Multimodal Integration – A Biological View

Michael H. Coen
MIT Artificial Intelligence Lab
545 Technology Square
Cambridge, MA 02139
mhcoen@ai.mit.edu

Abstract

We present a novel methodology for building highly integrated multimodal systems. Our approach is motivated by neurological and behavioral theories of sensory perception in humans and animals. We argue that perceptual integration in multimodal systems needs to happen at all levels of the individual perceptual processes. Rather than treating each modality as a separately processed, increasingly abstracted pipeline – in which integration over abstract sensory representations occurs as the final step – we claim that integration and the sharing of perceptual information must also occur at the earliest stages of sensory processing. This paper presents our methodology for constructing multimodal systems and examines its theoretic motivation. We have followed this approach in creating the most recent version of a highly interactive environment called the Intelligent Room and we argue that doing so has provided the Intelligent Room with unique perceptual capabilities and gives insight into building similar complex multimodal systems.

Introduction

This paper proposes a novel perceptual architecture motivated by surprising results about how the brain processes sensory information. These results, gathered by the cognitive science community over the past 50 years, have challenged century long held notions about how the brain works and how we experience the world we live in. We argue that current approaches to building multimodal systems that perceive and interact with the real, human world are flawed and based largely upon assumptions described by Piaget (1954) – although tracing back several hundred years – that are no longer believed to be particularly accurate or relevant.

Instead, we present a biologically motivated methodology for designing interactive systems that reflects more of how the brain actually appears to process and merge sensory inputs. This draws upon neuroanatomical, psychophysical, evolutionary, and phenomenological evidence, both to critique modern approaches and to suggest an alternative for building artificial perceptual systems. In particular, we argue against *post-perceptual* integration, which occurs in systems where the modalities are treated as separately processed, increasingly abstracted pipelines. The outputs of these pipelines are then merged in a final integrative step, as in Figure 1. The main difficulty with this approach is that integration happens after the individual perceptions are generated. Integration occurs *after* each perceptual subsystem has already "decided" what it has perceived, when it is too late for intersensory influence to affect the individual, concurrent unimodal perceptions. Multimodal integration is thus an *assembly* rather than a *perceptual* process in most modern interactive systems. In this, it is an abstraction mechanism whereby perceptual events are

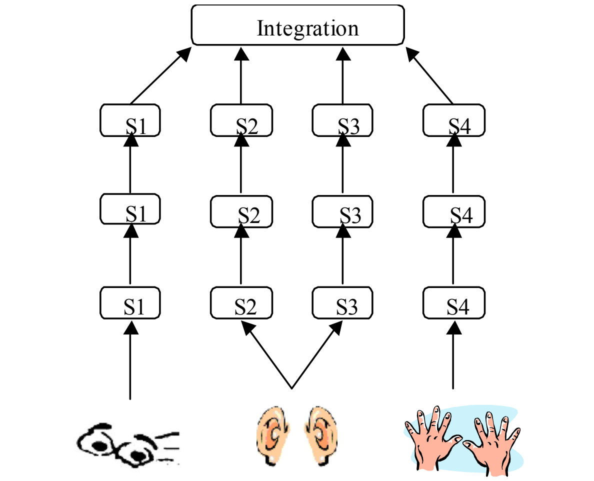

Figure 1 – Post-perceptual integration in a multimodal system. Here, visual, auditory, and haptic inputs pass through specialized unimodal processing pathways and are combined in the integration mechanism on top, which creates multimodal perceptions by extracting and reconciling data from the individual inputs. Notice that multiple pathways may process a single sensory input. For example, the auditory signal in this system might be fed into separate *speech recognition* and *spatial localization systems*, designated S2 and S3 respectively.

This material is based upon work supported by the Advanced Research Projects Agency of the Department of Defense under contract number DAAD17-99-C-0060, monitored through Rome Laboratory. Additional support was provided by the Microsoft Corporation.

separated from the specific sensory mechanisms that generate them and then integrated into higher-level representations.

This paper examines the idea that multimodal integration is a fundamental component of perception itself and can shape individual unimodal perceptions as much as does actual sensory input. This idea is well supported biologically, and we argue here for the benefits of building interactive systems that support *cross-modal influence* – systems in which sensory information is shared across all levels of perceptual processing and not just in a final integrative stage. Doing this, however, requires a specialized approach that differs in basic ways from how interactive systems are generally designed today.

This paper presents our methodology for constructing multimodal systems and examines its theoretic motivation. We have followed this approach in creating the most recent version of a highly interactive environment called the Intelligent Room, and we argue that doing so has provided the Intelligent Room with unique perceptual capabilities and provides insight into building similar complex multimodal systems. We specifically examine here how the Intelligent Room's vision and language systems share information to augment each other and why the traditional ways these systems are created made this sharing so difficult to implement. Our approach is similar in spirit to the work of (Atkeson et al. 2000, Brooks et al. 1998, Cheng and Kuniyoshi 2000, Ferrell 1996, Nakagawa 1999, and Sandini et al. 1997). Although they are primarily concerned with sensorimotor coordination, there is a common biological inspiration and long-term goal to use knowledge of human and animal neurophysiology to design more sophisticated artificial systems.

Background

Who would question that our senses are distinct? We see, we feel, we hear, we smell, and we taste, and these are qualitatively such different experiences that there is no room for confusion among them. Even those affected with the peculiar syndrome *synesthesia*, in which real perceptions in one sense are accompanied by illusory ones in another, never lose awareness of the distinctiveness of the senses involved. Consider the woman described in (Cytowic 1988), for whom a particular taste always induced the sensation of a particular geometric object in her left hand. A strange occurrence indeed, but nonetheless, the tasting and touching – however illusory – were never confused; they were never merged into a sensation the rest of us could not comprehend, as would be the case, for example, had the subject said something *tasted* octagonal. Even among those affected by synesthesia, the sensory channels remain extremely well defined.

Given that our senses appear so unitary, how does the brain coordinate and combine information from different sensory modalities? This has become known as the *binding problem,* and the traditional assumption has been to assume that the sensory streams are abstracted, merged, and integrated in the cortex, at the highest levels of brain functioning. This was the solution proposed by Piaget (1954) and in typical Piagetian fashion, assumed a cognitive developmental process in which children slowly developed high-level mappings between innately distinct modalities through their interactions with the world. This position directly traces back to Helmholtz (1856), and even earlier, to Berkeley (1709) and Locke (1690), who believed that neonatal senses are congenitally separate and interrelated only through experience. The interrelation *does not diminish the distinctiveness of the senses themselves*, it merely accounts for correspondences among them based on perceived cooccurrences.

The Piagetian assumption underlies nearly all modern interactive, multimodal systems – it is the primary architectural metaphor for multimodal integration. The unfortunate consequence of this has been making integration a post-perceptual process, which assembles and integrates sensory input after the fact, in a separate mechanism from perception itself. In these multimodal systems, each perceptual component provides a generally high-level description of what it sees, hears, senses, etc. (Some systems, such as sensor networks, provide low-level feature vectors, but the distinction is not relevant here.) This, for example, may consist of locations of people and how they are gesturing, what they are saying and how (e.g., prosody data), and biometric data such as heart rates. This information is conveyed in modal-specific representations that capture detail sufficient for higher-level manipulation of the perceptions, while omitting the actual signal data and any intermediate analytic representations. Typically, the perceptual subsystems are independently developed and trained on unimodal data; each system is designed to work in isolation. (See Figure 2.) They are then interconnected through some fusive mechanism (as in Figure 1) that combines temporally proximal, abstract unimodal inputs into an integrated event model. The integration itself may be effected via a neural network (e.g., Waibel et al. 1995), hidden Markov models (e.g., Stork and Hennecke 1996), unification logics (e.g., Cohen et al. 1997), or various ad hoc techniques (e.g., Wu et al. 1999). The output of the integration process is then fed into some higher-level interpretative mechanism – the architectural equivalent of a cortex.

This post-perceptual approach to integration denies the possibility of cross-modal influence, which is pervasive in biological perception. Our visual, auditory, proprioceptive, somatosensory, and vestibular systems influence one another in a complex process from which perceptions emerge as an integrated product of a surprising diversity of components. (For surveys, see Stein and Meredith 1993, Lewkowicz and Lickliter 1994, and to a lesser extent, Thelen and Smith 1994.) For example, consider the seminal work of McGurk and MacDonald (1976). In preparing an experiment to determine how infants reconcile conflicting information in different sensory modalities, they had a lab technician dub the audio syllable /ba/ onto a video of someone saying the syllable /ga/. Much to their surprise, upon viewing the dubbed video, they repeatedly and distinctly heard the syllable /da/ (alternatively, some hear /tha/), corresponding neither to the actual audio nor video sensory input. Initial assumptions that this was due to an error on the part of the technician were easily discounted simply by shutting their eyes while watching the video; immediately, the sound changed to a /ba/. This surprising fused perception, subsequently verified in numerous redesigned experiments and now known as the *McGurk effect*, is robust and persists even when subjects are aware of it.

The McGurk effect is perhaps the most convincing demonstration of the intersensory nature of face-to-face spoken language and the undeniable ability of one modality to radically change perceptions in another. It has been one of many components leading to the reexamination of the Piagetian introspective approach to perception. Although it may seem reasonable to relegate intersensory processing to the cortex for the *reasoning* (as opposed to *perceptual*) processes that interested Piaget, such as in cross modal matching, it becomes far more implausible in cases where different senses impinge upon each other in ways that locally change the perceptions in the sensory apparatus themselves. One might object that the McGurk effect is pathological – it describes a perceptual phenomenon outside of ordinary experience. Only within controlled, laboratory conditions do we expect to have such grossly conflicting sensory inputs; obviously, were these signals to co-occur naturally in the real world, we would not call them conflicting. We can refute this objection both because the real world *is* filled with ambiguous, sometimes directly conflicting, perceptual events, and because the McGurk effect is by no means the only example of its kind. There is a large and growing body of evidence that the type of direct perceptual influence illustrated by the McGurk effect is commonplace in much of ordinary human and more generally animal perception, and it strongly makes the case that our perceptual streams are far more interwoven than conscious experience tends to make us aware. For example, the sight of someone's moving lips in an environment with significant background noise makes it easier to understand what the speaker is saying; *visual* cues – e.g., the sight of lips – can alter the signal-to-noise ratio of an *auditory* stimulus by 15-20 decibels (Sumby and Pollack 1954). Thus, a decrease in auditory acuity can be offset by increased reliance on visual input. Although the neural substrate behind this interaction is unknown, it has been determined that just the *sight* of moving lips – without any audio component – modifies activity in the *auditory* cortex (Sams et al 1991). In fact, psycholinguistic evidence has long lead to the belief that lip-read and heard speech share a degree of common processing, notwithstanding the obvious differences in their sensory channels (Dodd et al 1984).

Perhaps the most dramatic cross-modal interactions were demonstrated in the landmark studies of Meltzoff and Moore (1977), who showed that infants could imitate an investigator's facial expression within hours of birth. For example, the investigator sticking out his tongue lead the infant to do the same, although the infant had never seen its own face. Somehow, the visual cue is matched with the proprioceptive sensation of tongue protrusion. It is extraordinarily difficult to view this as a learned behavior; this will be quite relevant in considering below where the knowledge governing cross-modal influence should come from in computational systems.

We believe that the current approach to building multimodal interfaces is an artifact of how people like to build computational systems and not at all well-suited to dealing with the cross-modal interdependencies of perceptual understanding. Perception does not seem to be amenable to the clear-cut abstraction barriers that computer scientists find so valuable for solving other problems, and we claim this approach has lead to the fragility of so many multimodal systems. We also dispute the notion that perception generally corresponds well to a discrete and symbolic *event* model. Although such a model is well suited to many types of computational applications, it is frequently more useful to view perception as a dynamic process, rather than an event occurring at some point in time (Thelen and Smith 1994). This is all the more so during development (i.e. learning), where the space of events is fluid and subject to constant change, and during sensory fusion, where complex codependences and interactions among the different senses confound our simple event-biased predispositions. A similar dynamic approach has been taken by (Ullman 1996) for explaining cross-feature influence in visual perception.

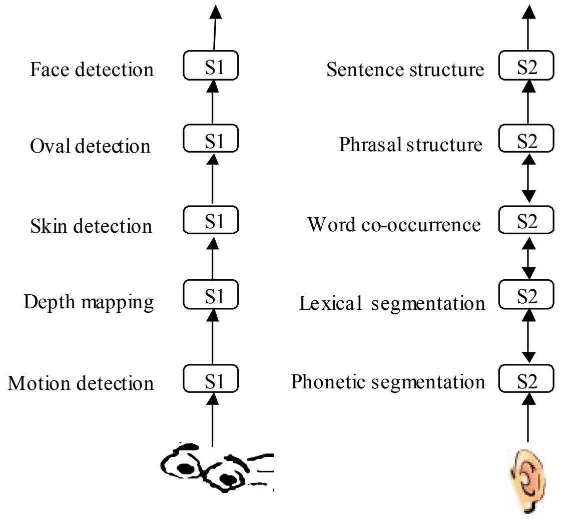

Figure 2 – Unimodal processing pathways. Individual modalities are processed in specialized pipelines. The visual pathway on the left uses 3-dimensional depth and color maps to find candidate regions for locating faces. (This follows Darrell et al. 1998.) The auditory pathway on the right performs speech recognition. Notice that higher-level syntactic constraints feed back into the lower-level morphological analysis.

A Multimodal System

The examples for our discussion of multimodal interactions will be drawn from the Intelligent Room project, as described in (Coen 1998, 1999). The Intelligent Room has multiple perceptual user interfaces, supporting both visual and verbal interactions, which connect with some of the ordinary human-level events going on within it. The goal of the room is to support people engaged in everyday, traditionally non-computational activity in both work and leisure contexts. Figure 3 contains a picture of the room, which contains nine video cameras, three of which the room can actively steer, several microphones, and a large number of computer-interfaced devices. Because the Intelligent Room is designed to support a range of activities that have never before explicitly involved computers, it in no way resembles a typical computer science laboratory. Its computational infrastructure is removed from view to enhance the room's appeal as a *naturally* interactive space with a decidedly understated "low-tech" atmosphere. The room's computational infrastructure (Coen et al. 1999), consisting of over 120 software agents running on a network of ten workstations, is housed in an adjacent laboratory.

Before exploring scenarios for new types of multimodal interactions, we first examine a traditionally explicit one in the resolution of a deictic reference, i.e., use of a word such as *this* in referring to a member of some class of objects. Suppose, for example, someone in the room says, "Dim this lamp." The room uses its ability to track its occupants, in conjunction with a map of its own layout, to dim the lamp *closest* to the speaker when the verbal command was issued. This kind of interaction can be implemented with a simple post-perceptual integration mechanism that reconciles location information obtained from the person tracker with the output of a speech recognition system. Here, multimodal integration of positional and speech information allows straightforward disambiguation of the deictic lamp reference.

Motivations

Given the simplicity of the above example, it seems far from obvious that a more complex integration mechanism is called for. To motivate a more involved treatment, we start by examining some of the problems with current approaches.

Despite many recent and significant advances, computer vision and speech understanding, along with many other perceptual interface research areas (Picard 1997, Massie and Salisbury 1994), are still infant sciences. The non-trivial perceptual components of multimodal systems are therefore never "perfect" and are subject to a wide variety of failure modes. For example, the room may "lose" people while visually tracking them due to occlusion, coincidental color matches between fore and background objects, unfavorable lighting conditions, etc. Although the particular failure modes of the modalities varies with them individually, it is a safe assumption that under a wide variety of conditions any one of them may temporarily stop working as desired. How these systems manifest this undesired operation is itself highly idiosyncratic. Some may simply provide no information, for example, a speech recognition system confused by a foreign accent. Far more troublesome are those that continue to operate as if nothing were amiss but simply provide incorrect data, such as a vision-based tracking system that mistakes a floor lamp for a person and reports that he is standing remarkably still.

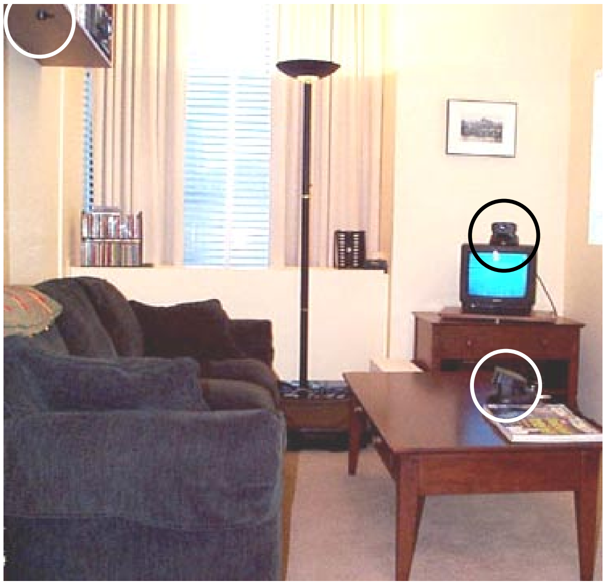

Figure 3 – A view of the Intelligent Room with three of its nine cameras visible and circled. The two lower of these in the picture can be panned and tilted under room control.

That perceptual systems have a variety of failure modes is not confined to their artificial instantiations. Biological systems also display a wide range of pathological conditions, many of which are so engrained that they are difficult to notice. These include limitations in innate sensory capability, as with visual blind spots on the human retina, and limited resources while processing sensory input, as with our linguistic difficultly understanding nested embedded clauses (Miller and Chomsky 1963). Stein and Meredith (1994) argue for the evolutionary advantages of overlapping and reinforcing sensory abilities; they reduce dependence on specific environmental conditions and thereby provide clear survival advantages. A striking example of this is seen in a phenomenon known as the "facial vision" of the blind. In locating objects, blind people often have the impression of a slight touch on their forehead, cheeks, and sometimes chest, as though being touched by a fine veil or cobweb (James 1890, p204). The explanation for this extraordinary perceptory capability had long been a subject of fanciful debate. James demonstrated, by stopping up the ears of blind subjects with putty, that audition was behind this sense, which is now known to be caused by intensity, direction, and frequency shifts of reflected sounds (Arias 1996). The auditory input is so successfully represented haptically in the case of facial vision that the perceiver himself cannot identify the source of his perceptions.

Research on interactive systems has focused almost entirely on unimodal perception: the isolated analysis of auditory, linguistic, visual, haptic, or to a lesser degree biometric data. It seems to put the proverbial cart before the horse to ponder how information from different modalities can be merged while the perceptory mechanisms in the sensory channels are themselves largely unknown. Is it not paradoxical to suggest we should or even could study integration without thoroughly understanding the individual systems to be integrated? Nonetheless, that is the course taken here. We argue that while trying to understand the processing performed within individual sensory channels, *we must simultaneously ask how their intermediary results and final products are merged into an integrated perceptual system*. We believe that because perceptual systems within a given species coevolved to interoperate, *compatibility pressures existed on their choices of internal representations and processing mechanisms*. In order to explain the types of intersensory influence that have been discovered experimentally, *disparate perceptual mechanisms must have some degree of overall representational and algorithmic compatibility that makes this influence possible*. The approach taken here is entirely gestalt, not only from a Gibsonian (1986) perspective, but because we have no example of a complex unimodal sensory system evolving in isolation. Even the relatively simple perceptual mechanisms in paramecium (Stein and Meredith 1994, Chapter 2) and sponges (Mackie and Singla 1983) have substantial cross-sensory influences. It seems that perceptual interoperation is a prerequisite for the development of complex perceptual systems. Thus, rather than study any single perceptual system in depth – the traditional approach – we prefer to study them *in breadth*, by elucidating and analyzing interactions between different sensory systems.

Cross-Modal Influences

How then might cross-modal influences be used in a system like the Intelligent Room? Answering this question is a two-step process. Because the Intelligent Room is an engineered as opposed to evolved system, we first need to explicitly find potential synergies between its modalities that can be exploited. Once determined, these synergies must then somehow be engineered into the overall system, and this emerges as the primary obstacle to incorporating cross-modal influences into the Intelligent Room and more generally, to other types of interactive systems.

We begin with the following two empirical and complementary observations:

1) People tend to talk about objects they are near. (Figure 4a)

2) People tend to be near objects they talk about. (Figure 4b)

These heuristics reflect a relationship between a person's location and what that person is referring to when he speaks; knowing something about one of them provides some degree of information about the other. For example, someone walking up to a video display of a map is potentially likely to speak about the map; here, person location data can inform a speech model. Conversely, someone who speaks about a displayed map is likely to be in a position to see it; here, speech data can inform a location model. Of course, it is easy to imagine situations where these heuristics would be wrong. Nonetheless, as observations they are frequently valid and it would be reasonable to somehow incorporate influences based on them into a system like the Intelligent Room. Mechanistically, we might imagine the person tracking system exchanging information with the speech recognition system. For example, the tracking system might send data, which we will call a *hint,* to the speech recognition system to preferentially expect utterances involving objects the person is near, such as a map. Conversely, we can also imagine that the speech recognition

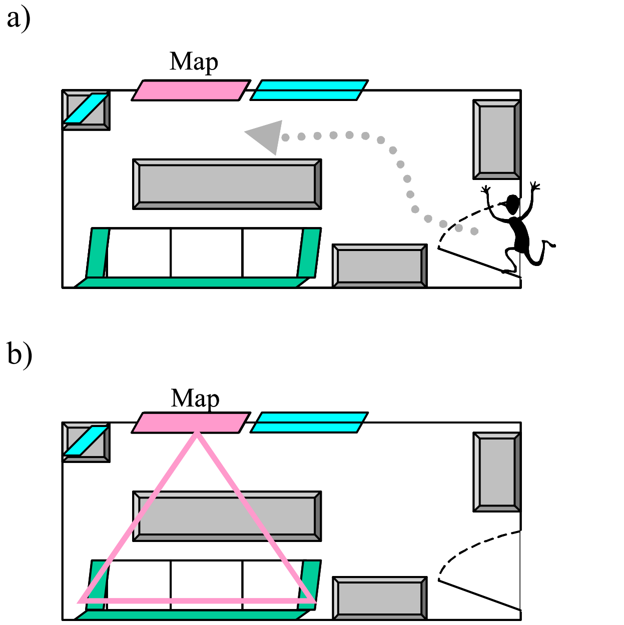

Figure 4a – People talk about objects they are near. Someone approaching a projected display showing, for example, a map, is more likely to make a geographical utterance. Here, location information can augment speech recognition.

Figure 4b – People are near objects they talk about. Someone speaking about the contents of a video display is more likely to be located somewhere (delineated by the triangle) from which the display is viewable to him. Here, speech information can augment person tracking.

system would send hints to the person tracking system to be especially observant when looking for someone in indicated sections of the room, based on what that person is referring to in his speech.

This seems reasonable until we try to build a system that actually incorporates these influences. There are both representational and algorithmic stumbling blocks that make this conceptually straightforward cross-modal information sharing difficult to implement. These are due not only to post-perceptual architectural integration, but also to how the perceptual subsystems (such as those in Figure 2) are themselves typically created. We first examine issues of representational compatibility, namely what interlingua is used to represent shared information, and then address how the systems could incorporate *hints* they receive in this interlingua into their algorithmic models.

Consider a person tracking system that provides the coordinates of people within a room in real-time, relative to some real-world origin – the system outputs the actual locations of the room's occupants. We will refer to the tracking system in the Intelligent Room as a representative example of other such systems (e.g., Wren et al. 1997, Gross et al. 2000). Its only input is a stereo video camera and its sole output are sets of *(x,y,z)* tuples representing occupants' centroid head coordinates, which are generated at 20Hz. Contrast this with the Intelligent Room's speech recognition system, which is based upon the Java Speech API (Sun 2001), built upon IBM's ViaVoice platform, and is typical of similar spoken language dialog systems (e.g., Zue et al. 2000). Its inputs are audio voice signals and a formal linguistic model of expected utterances, which are represented as probabilistically weighted context free grammars.

How then should these two systems exchange information? It does not seem plausible from an engineering perspective, whether in natural or artificial systems, to provide each modality with access to the internal representations of the others. Thus, we do not expect that the tracking system should know anything about linguistic models nor we do expect the language system should be skilled in spatial reasoning and representation. Even if we were to suppose the speech recognition system *could* somehow represent spatial coordinates, e.g. as *(x,y,z)* tuples, that it could communicate to the person tracking system, the example in Figure 4b above involves *regions* of space, not isolated point coordinates. From an external point of view, it is not obvious how the tracking system internally represents regions, presuming it even has that capability in the first place. The complementary example of how the tracking system might refer to classes of linguistic utterances, as in Figure 4a above, is similarly convoluted.

Unfortunately, even if this interlingua problem were easily solvable and the subsystems had a common language for representing information, the way most perceptual subsystems are implemented would make incorporation of cross-modal data difficult or impossible. For example, in the case of a person tracking system, the real-world body coordinates are generated via three-dimensional spatial reconstruction based on correspondences between sets of image coordinates. The various techniques for computing the reconstructed coordinates, such as neural networks or fit polynomials, are in a sense closed – once the appropriate coordinate transform has been learned, there is generally no way to bias the transformation in favor of particular points or spatial regions. Thus, there is no way to incorporate the influence, even if the systems had a common way of encoding it. Here again, the complementary situation with influencing speech recognition from a tracking system can be similarly intractable. For example, not all linguistic recognition models (e.g., bigram-based) support dynamic preferential weighting for classes of commonly themed utterances. So, even if the tracking system could somehow communicate what the speech recognition system should expect to hear, the speech recognition system might not be able to do anything useful with this information.

We see that not only are the individual modal representations incompatible, the perceptual algorithms (i.e., the contents of the sensory pipelines) are incompatible as well. This comes as no surprise given that these systems were engineered primarily for unimodal applications. Unlike natural perceptual systems within an individual species, artificial perceptual systems do not co-evolve, and therefore, have had no evolutionary pressure to force representational and algorithmic compatibility. These engineered systems are intended to be data sources feeding into other systems, such as the ones performing multimodal integration, that are intended to be data sinks. There is no reason to expect that these perceptual subsystems would or even could directly interoperate.

Designing for Interaction

Our solution was to redesign the Intelligent Room's perceptual systems with the explicit intension that they should interact with each other. Doing so required fundamental representational and algorithmic changes but has made possible subtle types of cross-modal interactions that were previously unworkable. We first detail two different categories of intersensory function and then explain how the cross-modal influences described above were implemented in the Intelligent Room.

Consider, for example, the effect of touching someone and having his head and eyes turn to determine the source of the stimulus. This is clearly an example of cross-modal influence – the position of the touch determines the foveation of the eyes – but it is fundamentally different than the interaction described in the McGurk effect above, where the influence is evidenced solely in perceptual channels. The touch scenario leads to behavioral effects that center the stimulus with respect to the body and peripheral sensory organs. The McGurk effect is an example of sensory influence within perceptual channels and has no behavioral component.

Motor influences – i.e., ones that cause attentive and orientation behaviors – are by far the more understood of the

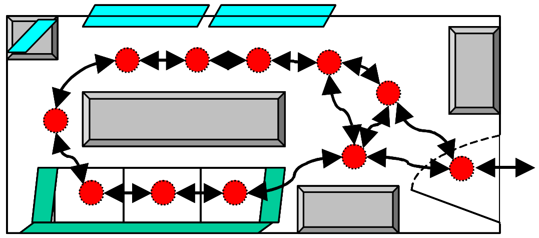

Figure 5 – A learned topographic map overlaid on top of the room's floor plan. The dots represent locations in the room that were used as reference points to interactively train a multi-camera vision system. The arrows represent observed transitions between these reference locations. Once trained, the room can then locate people at and between these points.

two. The primary neurological substrate behind them is the *superior colliculus,* a small region of the brain that produces signals that orient peripheral sensory organs based on sensory stimuli. The superior colliculus contains layered, topographic sensory and motor maps that are in register; that is, co-located positions in the real world – in the sensory case representing derived locations of perceptual inputs and in the motor case representing peripheral sensory organ motor coordinates that focus on those regions – are all essentially vertically overlapping. The actual mechanisms that use these maps to effect intersensory influence are currently unknown – variants on spreading vertical activation are suspected – but there is little doubt the maps' organization is a fundamental component of that mechanism.

Far less is known neurologically about purely semantic influences – i.e., ones that have effects confined to perceptual channels. The superior colliculus itself has been directly approachable from a research perspective because the brain has dedicated inner space, namely, the tissue of the topographic maps, to representing the outer space of the real-world; the representation is both isomorphic and perspicacious, and it has made the superior colliculus uniquely amenable to study. The perceptual as opposed to spatial representations of the senses are far more elusive and are specialized to the individual modalities and the organs that perceive them.

We have used the notion of layered topographic maps to represent both motor *and* semantic information in the Intelligent Room. Even though the superior colliculus has not yet been identified as a substrate in *non-behavioral* cross-modal influences, its extensive intra-map connections to higher cortical areas – particularly the visual cortex – may indicate its role in other types of intersensory function that are confined to perceptual channels and have no behavioral component.

Using a topographic organization, we created a new model for visually tracking people within a room (Coen and Wilson 1999). The model takes advantage of the observation that much of the information needed in human-computer interaction is qualitative in nature. For example, it may be necessary to distinguish between a person sitting on a chair, a person standing in front of a bookcase, and a person standing in a doorway, but obtaining the actual real-world coordinates of these people is generally unimportant. In our system, locations in a room that are likely to contain people are used as reference points, as in Figure 5, to interactively train a multi-camera vision system, whose current implementation has three steerable and six fixed cameras. The system learns to combine event predictions from the multiple video streams in order to locate people at these reference points in the future. The system can also dynamically track people with the steerable cameras as they move between these locations. Because most rooms have natural attractors for human activity, such as doorways, furniture, and displays, the selection of training points is usually readily apparent from the layout of the room.

Once this topographic map is created, the tracking system activates locations on the map to correspond to its observations of people in the room. As we will see in a moment, other systems in the room can also *weakly* activate locations on the map, *which causes the room to turn a steerable camera to view the corresponding real world location.* If a person is found there as a result of orienting the camera, the system then completely activates that location on the map.

The ability to interact with topographic representations of the room is not confined to the tracking system. Once the map is learned, the room builds corresponding maps to spatially categorize events in its other sensory systems, even if they have no explicit spatial component. For example, speech recognition events are topographically organized on a map dedicated just for that purpose. As utterances are heard for which the room has spatial information (either learned or explicitly provided by its programmers), it activates locations in the speech system's topographic map, which in turn activates locations in other modalities' topographic maps via vertical spreading activation, as shown in Figure 6. Conversely, other systems can *weakly* activate locations on the speech system's topographic map, which causes the speech system to increase the expectation probabilities of utterances associated with that location.

Thus, activations in the map are *bi-directional*: perceptual events in a given modality can directly activate its map locations. Maps locations can also be activated via spreading activation from corresponding positions in other system's topographic maps, *which causes a corresponding change in that modality's perceptual state* — here, these spreading activations cause the secondary system to either look or listen for something. It is this bi-directional activation – through which the systems can react to intersensory stimuli – that has made possible the cross-modal influences that were presented in Figure 4.

Conclusion

This paper has described our approach to incorporating cross-modal influences into the perceptual processing of the

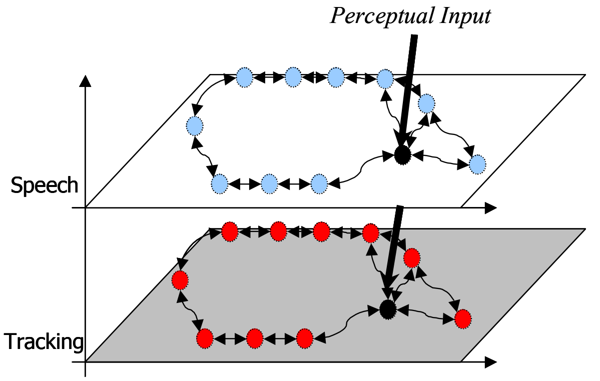

Figure 6 – Layered topographic maps containing location-dependent speech and tracking information. Positions on the maps can be activated directly by perceptual inputs or activated by spreading vertical activation from adjacent maps.

Intelligent Room. We simultaneously argued against conventional post-perceptual integration and have motivated this position with biological evidence that unimodal perceptions are themselves integrated products of multimodal sources. Our position has allowed us to explore representational and algorithmic issues in *unimodal* perception that can only be approached from an integrated, *multimodal* perspective. It has also allowed us to investigate creating more sophisticated interactive systems by incorporating more subtle intersensory cues into the Intelligent Room. Future work is both exciting and promising.

References

1. Atkeson CG, Hale J, Pollick F, Riley M, Kotosaka S, Schaal S, Shibata T, Tevatia G, Vijayakumar S, Ude A, Kawato M: Using humanoid robots to study human behavior. *IEEE Intelligent Systems***,** Special Issue on Humanoid Robotics, 46-56. 2000.
2. Brooks, R.A., C. Breazeal (Ferrell), R. Irie, C. Kemp, M. Marjanovic, B. Scassellati and M. Williamson, Alternate Essences of Intelligence. *In Proceedings of The Fifteenth National Conference on Artificial Intelligence.* (AAAI98). Madison, Wisconsin. 1998.
3. Butterworth, G. The origins of auditory-visual perception and visual proprioception in human development. In *Intersensory Perception and Sensory Integration*, R.D. Walk and L.H. Pick, Jr. (Eds.) New York. Plenum. 1981.
4. Cheng, G., and Kuniyoshi**,** Y. Complex Continuous Meaningful Humanoid Interaction: A Multi Sensory-Cue Based Approach *Proc. of IEEE International Conference on Robotics and Automation* (ICRA 2000), pp.2235-2242, San Francisco, USA, April 24-28, 2000.
5. Coen, M. Design Principles for Intelligent Environments. *In Proceedings of The Fifteenth National Conference on Artificial Intelligence.* (AAAI98). Madison, Wisconsin. 1998.
6. Coen, M. The Future Of Human-Computer Interaction or How I learned to stop worrying and love My Intelligent Room. IEEE Intelligent Systems. March/April. 1999.
7. Coen, M., and Wilson, K. Learning Spatial Event Models from Multiple-Camera Perspectives in an Intelligent Room. In *Proceedings of MANSE'99*. Dublin, Ireland. 1999.
8. Coen, M., Phillips, B., Warshawsky, N., Weisman, L., Peters, S., Gajos, K., and Finin, P. Meeting the computational needs of intelligent environments: The Metaglue System. In *Proceedings of MANSE'99*. Dublin, Ireland. 1999.
9. Cohen, P. R., Johnston, M., McGee, D., Smith, I. Oviatt, S., Pittman, J., Chen, L., and Clow, J. QuickSet: Multimodal interaction for simulation set-up and control. 1997, *Proceedings of the Applied Natural Language Conference,* Association for Computational Linguistics. 1997.
10. Cytowic, R. E., Synesthesia: A Union of Senses. New York. Springer-Verlag. 1989.
11. Darrell, T., Gordon, G., Harville, M., and Woodfill, J., Integrated person tracking using stereo, color, and pattern detection, *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR '98),* pp. 601-609, Santa Barbara, June 1998.
12. Ferrell, C. Orientation behavior using registered topographic maps. In Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior (SAB-96). Society of Adaptive Behavior. 1996.
13. Gross, R., Yang, J., and Waibel, A. Face Recognition in a meeting room. Fourth IEEE International Conference on Automatic Face and Gesture Recognition, Grenoble, France, March 2000
14. Held, R. Shifts in binaural localization after prolonged exposures to atypical combinations of stimuli. Am. J. Psychol. 68L526-266. 1955.
15. Helmholtz, H. v. *Handbook of Physiological Optics.* 1856. as reprinted. in James P.C. Southall. (Ed.) 2000.
16. James, H. 1890. Principles of Psychology. Vol. 2. Dover. 1955.
17. Kohler, I. The formation and transformation of the perceptual world. Psychological Issues 3(4):1-173. 1964.
18. McGurk, H., and MacDonald, J. Hearing lips and seeing voices. Nature. 264:746-748. 1976.
19. Meltzoff, A.N. and Moore, M.K. Imitation of facial and manual gestures by human neonates. Science 198:75-78. 1977.
20. Miller, George and Noam Chomsky (1963). Finitary models of language users. In Luce, R.; Bush, R. and Galanter, E. (eds.) *Handbook of Mathematical Psychology, Vol 2.* New York: Wiley. 419-93.
21. Piaget, J. *Construction of reality in the child,* London: Routledge & Kegan Paul, 1954.
22. Sams, M., Aulanko, R., Hamalainen, M., Hari, R., Lounasmaa, O., Lu, S., and Simola, J. Seeing speech: Visual information from lip movements modified activity in the human auditory cortex. Neurosci. Lett. 127:141-145. 1991.
23. Sandini G., Metta G. and Konczak J. "Human Sensori-motor Development and Artificial Systems". In: AIR&IHAS '97, Japan. 1997.
24. Stein, B., and Meredith, M. A. The Merging of the Senses. Cambridge, MA. MIT Press. 1994.
25. Stork, D.G., and Hennecke, M. Speechreading: An overview of image processing, feature extraction, sensory integration and pattern recognition techniques", *Proc. of the Second Int. Conf. on Auto. Face and Gesture Recog.* Killington, VT pp. xvi--xxvi 1996.
26. Sumby, W.H., and Pollack, I. Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am. 26:212-215. 1954.
27. Sun. http://www.javasoft.com/products/java-media/speech/. 2001.
28. Thelen, E., and Smith, L. A Dynamic Systems Approach to the Development of Cognition and Action. Cambridge, MIT Press. 1998.
29. Ullman, Shimon. High-level vision: object recognition and visual cognition. Cambridge. MIT Press. 1996.
30. Waibel, A., Vo, M.T., Duchnowski, P., and Manke, S. Multimodal Interfaces. Artificial Intelligence Review. 10:3-4. p299-319. 1996.
31. Wren, C., Azarbayejani, A., Darrell, T., and Pentland, P., "Pfinder: Real-Time Tracking of the Human Body ", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, July 1997.
32. Wu, Lizhong, Oviatt, Sharon L., Cohen, Philip R., Multimodal Integration -- A Statistical View, *IEEE Transactions on Multimedia*, Vol. 1, No. 4, December 1999, pp. 334-341.
33. Zue, V., Seneff, S., Glass, J., Polifroni, J., Pao, C., Hazen, T., and Hetherington, L. "JUPITER: A Telephone-Based Conversational Interface for Weather Information," IEEE Transactions on Speech and Audio Processing, Vol. 8, No. 1, January 2000.