Can Modeling Redundancy In Multimodal, Multi-party Tasks Support Dynamic Learning?

Edward C. Kaiser

Oregon Health and Science University
OGI School of Science & Engineering
20000 NW Walker Road, Beaverton, OR., 97006, USA
kaiser@cse.ogi.edu

Abstract

In multi-party interactions humans use available communication modes in predictable ways. For example, the dialogue theories of Conversational Implicature (Grice 1975) and Giveness Theory (Gundel, Hedberg et al. 1993) have both been applied successfully in the analysis of multi-party, multimodal settings (Chai, Prasov et al. 2005). At times people use multiple modes of communication in complementary or mutually disambiguating ways (Oviatt and Olsen 1994), while at other times the information in multiple modes is redundant (Anderson, Hoyer et al. 2004) (Fig. 2). Our position is that gaining a better understanding of why, when and how people choose to communicate multimodal information redundantly is very important for emerging computational systems that aim to be intelligent assistants for humans. Our technique of Multimodal New Vocabulary Recognition (MNVR) learns the spelling, pronunciation and semantics of new, out-of-vocabulary (OOV) words from a single observation of redundant handwriting and speech in a naturally occurring exchange of information within a multi-party scheduling meeting (Fig. 1). Similar redundancies can occur across other modes (e.g. gazing at someone while speaking their name). We believe that empirical research into the nature of communicative redundancy could be a very informative guide to the development and integration of a generalized dynamic learning approach in evolving multimodal interfaces.

Introduction

Our goal is to create computer systems that learn as easily as humans. As machines move closer to being observant and intelligent assistants for humans it is not enough that they rely on off-line models for the support of recognition. They need to automatically adapt and acquire new models and new knowledge as they are running, particularly from single instance, natural demonstrations.

Current recognition systems need sophisticated models of both features and higher level sequential or combinatory patterns; for example, speech recognizers are trained at the feature level on large numbers of corpus-based examples of phonetic segments and then at higher levels are constrained by either rule-based symbolic or corpus-based statistical language models. But whether recognition is rule-based or statistical no system of such static models

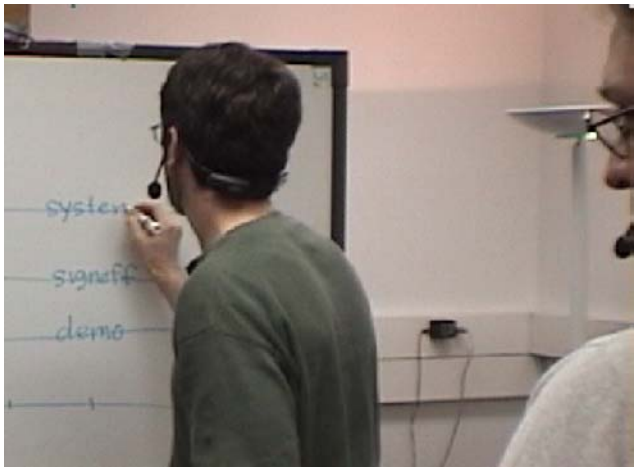

Figure 1: Using handwriting and speech to label task-lines on a Gantt chart in a multimodal, multi-person schedule meeting.

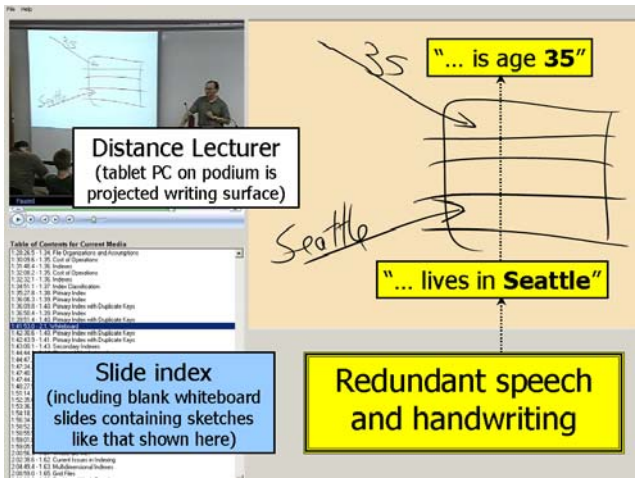

Figure 2: Redundant speech and handwriting during delivery of a distance learning lecture, using the Classroom Presenter system. Shown here on CXP Web Viewer (©UW 2002-4).*

* http://www.cs.washington.edu/education/dl/confxp/webviewer.html

can achieve full coverage. Natural language is replete with new words, new word patterns, and new topics. Thus, symbolic rule-based systems are notoriously brittle (because they cannot handle new words or word patterns), while statistical models typically fail on input that has little relation to their training data (as for example when a dialogue shifts to a previously unseen topic area). Thus our position is that automatically acquiring new knowledge — for example, the semantics, orthography and pronunciation of out-of-vocabulary (OOV) terms — as the system is running, particularly by a single, natural demonstration is critical not only to significantly enhancing the usability of observant, intelligent systems across the virtuality continuum, but also to supporting the evolution of those system into dynamic learning machines capable eventually of language acquisition.

We believe that an important step towards systems that learn as easily as humans is creating cognitive systems that learn implicitly from observation. For example, in our multimodal application for tracking the creation of a whiteboard schedule chart in a multi-party meeting (Figure 1), observation of redundant speech and handwriting input can support the recognition and dynamic enrollment of new vocabulary (Kaiser, Demirdjian et al. 2004; Kaiser 2005). So, rather than asking our users to engage in explicit *socially guided learning* — as exemplified by MIT's Leonardo system (Breazeal, Brooks et al. 2004) — where an instructor specifically guides the cognitive machine to understand commands for performing new actions through a process of iterative interaction, we instead ask the computer to leverage aspects of implicit observations of naturally occurring human-human communication to learn dynamically from a single interaction.

Capturing events in which handwriting and speech co-occur and carry redundant information is integral to our MNVR technique. In our previous work we have shown that for the task of labeling a schedule chart task-line (Fig. 1) in which a user writes the task-line name while speaking it, it is more effective to combine the redundant speech and handwriting information than to rely solely on either mode alone (Kaiser 2004; Kaiser 2005). In the human-computer-interaction (HCI) literature on bi-modal, speech and pen wizard-of-oz systems for map-based and form-filling tasks speech and handwriting have been found to co-occur redundantly in this way for less than 1% of all interactions (Oviatt and Olsen 1994; Oviatt, DeAngeli et al. 1997). However, in the educational-technology literature on human-human, computer-mediated interactions like the presentation of distance-learning lectures as much as 15% of all pen interactions were found to be handwriting (Anderson, Anderson et al. 2004), and a follow-on study to that work noted that in a tablet-PC-based, distance-learning, lecture-presentation application 100% of the randomly sampled instances of handwritten text were accompanied by semantically redundant speech (Anderson, Hoyer et al. 2004). Thus, when humans believe they are directly addressing a computer the current evidence is that they use multiple modes for presenting their input in a complementary rather than redundant fashion, but in contexts where the computer is a mediator or observer of natural multi-party interactions then redundancy in human-human multimodal presentation does occur.

Why should this be the case that in some contexts people use multiple modes redundantly while in others they do not? Grice's theory of Conversational Implicature (Grice 1975) proposes four maxims, two of which are the *Maxim of Quantity (MQ)* and the *Maxim of Manner (MM)*. These maxims, as described by (Chai, Prasov et al. 2005), posit that humans offer as much but no more information than is required for the purposes of conversational exchange *(MQ)*, and while avoiding obscurity and ambiguity are both brief and orderly in their expressions *(MM)*. Gundel's Giveness Hierarchy (Gundel, Hedberg et al. 1993), again as explained by Chai *et al*, is based on empirical matching of referring expressions (e.g. different determiners and pronominal forms — "it", "this one", "that house", "2020 Vision Street") to a hierarchy of giveness beginning with *Focus* (under conversation), moving to *Activation* (in short term memory), and eventually ending at *Indentifiable* (proper noun descriptor). Combining notions from these two theories Chai *et al* create a greedy algorithm for hierarchical reference resolution. Gesture, by virtue of the special effort it requires, is assigned hierarchical primacy. Given the user's referential expression and the status of the display objects with which they are interacting, Chai *et al's* algorithm performs comparably to NP graph matching algorithms while achieving sub-polynomial time execution. Thus it shows the benefit of using linguistic theory to inform approaches to hard problems in multimodal interface design.

We believe that by coming to a better understanding of why redundancy occurs in human-human communication we can better use it to support dynamic learning in observant cognitive systems. Redundancy in rich multimodal environments could provide the threshold ability that allows fully bootstrapped learning. For example redundant multimodal information may provide a basis for dynamic learning in the following perceptual environments:

- Redundant speech and 2D sketch could support dynamic enrollment of new sketch objects in a sketch recognizer.
- Redundant speech and 3D gesture or head/body-posture could support dynamic enrollment of new manipulative or iconic 3D gestures as well as new significations of assent/dissent or attention/inattention.
- Redundant gaze, speech and face recognition could support dynamic enrollment of new faces in a face recognition system.
- Redundant speech, gaze and visual activity recognition could support dynamic enrollment of new activity types.

We imagine a system that could learn the name of a new manipulative gesture through redundant demonstration, then using that known gesture transfer the semantics to a simultaneously uttered but previously unseen spoken utterance, or vice versa — using a known spoken reference transfer the newly acquired semantics of the reference to a simultaneously performed alternative gesture. This is what we mean by fully bootstrapped learning: multimodal redundancy serves as the basis for perceptual grounding, which in turn supports the transfer of semantics grounded in one mode to new, alternative symbols in another mode.

We envision that this kind of semantic bootstrapping could allow multimodal command languages in virtual and augmented reality environments to be much more adaptable and responsive to user preference. Adapting to users' command preferences is a need that we have identified in our previous work with MAVEN, our Multimodal Augmented and Virtual reality Environment for Natural interaction (Kaiser, Olwal et al. 2003) depicted in Fig. 3. In our studies with MAVEN users were constrained to learn a small set of manipulative hand/wrist gestures to be used in conjunction with speech to accomplish various object manipulations (like rotating a monitor as shown in Fig. 3). In general, user's found even such a simple four-gesture vocabulary awkward to use and difficult to remember. We believe that the ability to bootstrap-learn a customized gesture/speech vocabulary, while the system is running, through dynamic learning would be very desirable to users within MAVEN.

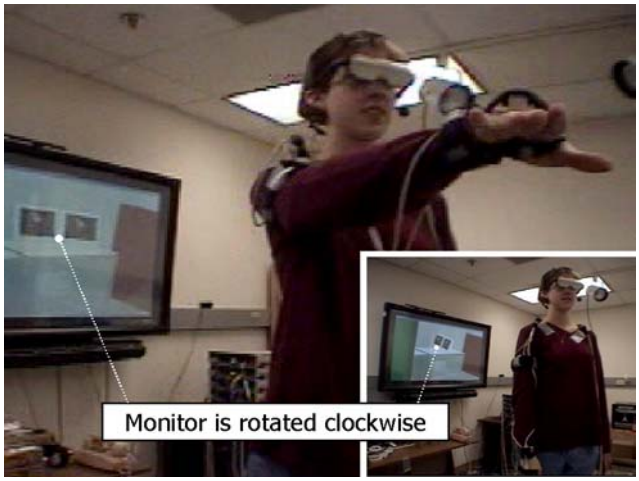

Figure 3: MAVEN virtual object manipulation: user says, "rotate the monitor clockwise," while making an awkward but required manipulative wrist gesture.

We know that establishing a common, working vocabulary is fundamental to human dialogue (Clark and Wilkes-Gibbs 1986). We know that humans expend all and only the necessary conversational energy to accomplish not only this grounding but also communication in general (Grice 1975; Chai, Prasov et al. 2005). If redundancy requires more energy, then what are the communicative purposes driving this? Recent work in multimodal information presentation (Zhou, Wen et al. 2005) outlines the factors involved in effective multimodal presentations. We know that **recallability** (which measures how well the presented data can be remembered) is affected by *transience*[1] and *overhead*[2]. We know that **affordance** (which measures how well a presentation captures a users attentional focus) is affected by aspects of *detectability* like *ordering*, *dependency* and *consistency*. These metrics from multimedia, human factors and attentional studies are related to metrics like **attentional focus,** which underlies the perception of *embodied intention* (Yu and Ballard 2003; Yu and Ballard 2003), a term from the growing literature on computational approaches to perceptual grounding in support of language learning. Perhaps redundancy affects **recallability**, **affordance**, and **attentional focus**. If so, then it may be that people consciously choose redundancy as a conversational strategy to bolster their communicative effectiveness. But it may also be true that in some instances redundancy is a reflex, an unconscious habit. Further empirical analysis needs to be done to gain insight into these questions.

In the literature on early childhood learning the 'Intersensory Redundancy Hypothesis' (**IRH**) (Bahrick, Lickliter et al. 2004) theorizes that sensory redundancy "facilitates attention to critical aspects of sensory stimulation." For example, the fact that infants are attentionally sensitive to *amodal* stimuli (e.g., synchrony, rythym and intensity — like the multisensory experience of seeing a ball bounce in which visual and aural stimuli are synchronous in both rythym and intensity) predicts that (1) *amodal* qualities like tempo will be learned more easily from multisensory input than unimodal input, and that conversely (2) unimodal qualities like direction will be learned more readily when presented unimodally. Such differences in sensory affordance underlie aspects of the common language parents use in addressing young infants, termed "multimodal motherese" (Gogate, Walker-Andrews et al. 2001), like moving objects while naming them, pointing at or touching objects while referring to them, highlighting specific words through exaggerated prosody while using shorter sentences or placing words in sentence final position.

The authors of the Intersensory Redundancy Hypothesis conclude that the "organizing influence of intersensory redundancy in guiding early attention, perception, and cognition likely constitutes a general developmental principle" (Bahrick, Lickliter et al. 2004). Could it be that aspects of this developmental principle are applicable to

[1] *Transience* is a measure of persistence (e.g. text is less transient than animation).

[2] *Overhead* is a measure of cognitive demand (e.g. a page of text has a higher overhead than a single word of text).

even adult interactions involving learning to establish common ground in multi-party human-human communication? Could this help to explain why we see redundant presentations in some contexts, particularly learning contexts like the delivery of lectures?

Given the fact that multimodal redundancy occurs in some multi-party human-human interactions (like distance lectures or whiteboard meetings) how can a better empirical understanding of the above factors be gained, and then used in predicting or recognizing instances of multimodal redundancy? We know that this redundant expenditure of energy can be perceived and used by cooperative, observant machines to acquire new knowledge in unobtrusive ways (e.g., MNVR). Thus the underlying developmental principle needs to be better understood, both theoretically from the computer science perspective and empirically from the computer interface perspective. We believe this is an important issue in the development of observant cognitive systems the can learn dynamically. We are actively designing studies that use pilot applications incorporating our Multimodal New Vocabulary Recognition technique that can begin to fill in our knowledge in this regard.

Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. NBCHD030010. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DARPA or the Department of Interior-National Business Center (DOI-NBC).

References

Anderson, R., C. Hoyer, C. Prince, J. Su, F. Videon and S. Wolfman (2004). *Speech, Ink and Slides: The Interaction of Content Channels*. ACM Multimedia.

Anderson, R. J., R. Anderson, C. Hoyer and S. A. Wolfman (2004). *A Study of Digital Ink in Lecture Presentation*. CHI 2004: The 2004 Conference on Human Factors in Computing Systems, Vienna, Austria.

Bahrick, L. E., R. Lickliter and R. Flom (2004). "Intersensory redundancy guides infants' selective attention, perceptual and cognitive development." *Current Directions in Psychological Science* **13**: 99-102.

Breazeal, C., A. Brooks, J. Gray, G. Hoffman, C. Kidd, H. Lee, J. Lieberman, A. Lockerd and D. Mulanda (2004). "Humanoid Robots as Cooperative Partners for People." *International Journal of Humanoid Robots (Forthcoming)* **1**(2).

Chai, J. Y., Z. Prasov, J. Blaim and R. Jin (2005). *Linguistic Theories in Efficient Multimodal Reference Resolution: An Empirical Investigation*. International Conference on Intelligent User Interfaces, San Diego, CA, ACM Press.

Clark, H. H. and D. Wilkes-Gibbs (1986). "Referring as a collaborative process." *Cognition* **22**: 1-39.

Gogate, L. J., A. S. Walker-Andrews and L. E. Bahrick (2001). "The Intersensory Origins of Word Comprehension: an Ecological-Dynamic Systems View." *Development Science* **4**(1): 1-37.

Grice, H. P. (1975). Logic and Conversation. *Speech Acts*. P. Cole and J. Morgan. New York, Academic Press**:** 41-58.

Gundel, J. K., N. Hedberg and R. Zacharski (1993). "Cognitive Status and the Form of Referring Expressions in Discourse." *Language* **69**(2): 274-307.

Kaiser, E., D. Demirdjian, A. Gruenstein, X. Li, J. Niekrasz, M. Wesson and S. Kumar (2004). *Demo: A Multimodal Learning Interface for Sketch, Speak and Point Creation of a Schedule Chart*. International Conference on Multimodal Interfaces (ICMI '04), State College, PA.

Kaiser, E., A. Olwal, D. McGee, H. Benko, A. Corradini, X. Li, P. Cohen and S. Feiner (2003). *Mutual Disambiguation of 3D Multimodal Interaction in Augmented and Virtual Reality*. International Conference on Mutimodal Interfaces (ICMI '03).

Kaiser, E. C. (2004). *Dynamic New Vocabulary Enrollment through Handwriting and Speech in a Multimodal Scheduling Application*. Making Pen-Based Interaction Intelligent and Natural, Papers from the 2004 AAAI Symposium, Technical Report FS-04-06, Arlington, VA., USA.

Kaiser, E. C. (2005). *Multimodal New Vocabulary Recognition through Speech and Handwriting in a Whiteboard Scheduling Application*. Proceedings of the International Conference on Intelligent User Interfaces, San Diego, CA.

Oviatt, S. and E. Olsen (1994). *Integration Themes in Multimodal Human-Computer Interaction*. International ConferenceonSpoken Language Processing (ICSLP '94).

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

Yu, C. and D. H. Ballard (2003). A Computational Model of Embodied Language Learning. Rochester, New York, Computer Science Deptartment, University of Rochester.

Yu, C. and D. H. Ballard (2003). *A Multimodal Learning Interface for Grounding Spoken Language in Sensory Perceptions*. International Conference on Multimodal Interfaces (ICMI '03), Vancouver, B.C., Canada, ACM Press.

Zhou, M. X., Z. Wen and V. Aggarwal (2005). *A Graph-Matching Approach to Dynamic Media Allocation in Intelligent Multimedia Interfaces*. International Conference on Intelligent User Interfaces, San Diego, CA, ACM Press.