Beyond *just* the words

Upol Ehsan

PHIL 493: Honors Thesis

Advisor: Paul A. Gregory

March 29, 2013

Acknowledgements

This project is a manifestation of the synergy I see between philosophy (especially of mind and language) and the sciences (especially computer science, engineering, and mathematics). I have always felt these disciplines co-exist at the cutting edge of research and complement each other. Personally, engineering and mathematics equip me with the ability to indulge in empirical investigations while philosophy provides a privileged platform to rationally articulate and speculate on the next frontier of the issues. Although I entered as an engineering major at Washington and Lee University, my promiscuous brain fell in love with philosophy while maintaining a stable relationship with engineering and mathematics.

As embodied agents whose existence in the world is deeply tied to those of others, our successes are often products of our individual efforts and collective support of those around us. Thus, I am forever indebted to the support system at W&L that helped me realize a significant portion of my life's vision. First and foremost, I would like to thank Paul Gregory not only for the invaluable guidance he has provided for the thesis, but also for being an excellent mentor, advisor, and friend throughout my years here. Next, I would like to express my gratitude to Angela Smith, who has provided excellent feedback and support throughout the works-in-progress. I would also like to thank Emily Zankmann, Nathaniel Reisinger, Reilly Kidwell, and Wyn Boerkel for the insightful comments and critiques on the works-in-progress. Rachel Urban provided excellent editorial feedback, and I am very grateful for that. I would not have been able to fully comprehend the technical aspects of this project with the help of my mentor and friend Jonathan Erickson. His guidance, especially in engineering, has definitely influenced the conceptual framework of this project. I would also like to thank Alison Bell for facilitating stimulating conversations about autism and her critiques on many parts of the project.

This acknowledgement section will be incomplete if I do not mention Andy Clark, whose encouragement and support not only initiated the entire project, but also helped me navigate through much of it. Although the final state of the project looks very different its initial state, the journey, in itself, taught me more than I could have ever asked for. I would like to thank Rob Rupert, Jesse Prinz, Peter Godfrey-Smith, and Randall Beer for their insightful suggestions and their time to personally meet with me. Peter Robinson, Justine Cassell along with Jon Gratch shared valuable insights about their work in Affective Computing, which greatly aided my research on the empirical aspect of the project.

I am grateful for the Johnson Opportunity Grant and the R.E. Lee Research Scholarships that facilitated the research for this project. Last but not the least, I would to thank my parents from the bottom of my heart for always encouraging me to follow my passions and never influencing my academic decisions. Without all of your help, this work would not have reached its current potential…so thank you for everything.

— *Upol Ehsan*

Motivating scenario and introduction

Given mutual interest in a conversation (such as an interview), why is it that we prefer a phone conversation over an email conversation, a Skype conversation over a phone one, and an in-person, face-to-face conversation over all other forms of communication?

One often hears the response: "there's something extra added to the experience as we move from the email conversation to the in-person one, which facilitates a deeper understanding of the communication". While the response is a valid one, it leaves us with more questions than it answers. What *exactly* is the "something extra"? How and why does the addition of "something extra" increase the communication level in the conversation? If syntax and semantics were all there is to language, why bother with the "something extra"?

Human communication is embodied—facial expressions, gestures, and prosody (tone of voice and rhythm) add layers of important information to *just the words* used by a cognitive agent. In fact, *just* the use of words often fails in expressing the thoughts expressed easily by non-verbal modes of communication. Actions such as the inflection of the voice or the lop-sided smile convey our emotive and expressive thoughts more effectively than *just* words. Thus, the "something extra" mentioned in the conversational example is constituted by the layers of non-verbal communication. As the bandwidth of communication increases from email to in-person, each layer of non-verbal communication is added, enabling us to have a deeper understanding of the conversation. In this paper, I propose that non-verbal modes of communication are integral in our understanding of human language and thought. Moreover, I argue that a systematic understanding of non-verbal communication produces a novel way of looking at human communication. The incorporation and systematic understanding of non-verbal communication breaks the chain of the traditional syntax-

semantic centric viewpoint of language and thought, removing the blinders that have long narrowed the epistemological field of vision.

Before proceeding, I would like to state a few words of caution. It is extremely important to realize that I am *not* trying to undermine the role of verbal communication, for I appreciate the power of words. What I am suggesting is a shift in focus that incorporates non-verbal communication in models of human language and communication. This incorporation will present a novel way of looking at human communication, empowering us with a better understanding of language and thought. Ideally, we would want equal attention paid to both verbal and non-verbal communication when talking about language and thought.

Moreover, I am *not* using the phrase "systematic understanding" to entail a strictly rule-based formal syntactic approach. Systematic understanding might take the form of learning algorithms (in computers) which model the heuristics humans use in understanding and navigating social situations, especially on the basis of non-verbal cues. I will get into details about these models later. The important part to realize is that these models are not strictly rule-driven syntactic manipulation and are about processing non-verbal communication. We hope we can extract some interesting and generalizable patterns from the data provided by the models, enabling us to gain a deeper understanding of human-human interaction. I divide this paper into 5 sections. In section 1, I present viewpoints of prominent scholars like Rene Descartes, Daniel Dennett, Jerry Fodor, and Paul Churchland on the issue of language, representation, and thought. Using their works, I exhibit the lack of attention paid to the roles of non-verbal communication. Then, I consider the reasoning behind this negligence towards non-verbal communication and propose three plausible causes. Next, I begin section 2 by briefly addressing the importance of emotion in cognition and decision making. I proceed to present empirical work done in Affective Computing and Human-Computer

Interaction (HCI) and stop at "checkpoints" to consider their philosophical implications. Using the empirical work presented, I also consider both the practical and philosophical implication of the improvement of social signal processing abilities of individuals with Autism Spectrum Condition (ASC). Social signal processing can be roughly defined as the ability to infer one's mental states from one's emotional (or affective states). Next, I present further empirical research that pays equal attention to *both* verbal and non-verbal communication. In section 3, I extend the philosophical implications mentioned beforehand by focusing on the motivations behind using computer models to gain a systematic understanding of non-verbal communication. Moreover, I focus on the power of an embodied language on thought. I delineate how a systematic understanding of non-verbal communication augments our understanding of human-human interaction. In section 4, I consider objections to the current arguments and provide appropriate responses to them. Prior to the conclusion of the paper, section 5 deals with future implications of the current work both from a societal and philosophical point of view.

1: About non-verbal communication—*what has (not) been said so far and why*

Throughout human history, our ability to use (verbal) language is typically taken as evidence for our superior intellectual capacities compared to other higher mammals. Most philosophy of mind and language in the Western tradition has been heavily focused on a syntax-semantic viewpoint of language and a linguistic concept of communication. From the past to the present, from Descartes to Churchland, the debate on the role of language on human thought has centered on the power of words or verbal language. Little or no attention has been paid to the essential function of non-verbal aspects of language. I will draw on the works of Descartes, Fodor, Dennett and Churchland to illustrate the privileging of verbal language and the marginalization of non-verbal communication. The purpose of the following discussion is *not* to adjudicate the fine details of each

view; rather, it is to highlight the extreme focus on verbal language in the trajectory of the debate. Thus, I will paint a picture with broad strokes, crystallizing points in each author's views pertinent to the present discussion.

In Part Five of the *Discourse*(1985), Descartes uses the example of an automaton (or a machine) that is similar to a present-day conception of a humanoid. The automaton resembles the human body and carries out its actions in a realistic manner. According to Descartes, although the automaton is equipped with the power to produce words appropriate to certain situations, its non-human nature is revealed through its inability to use language meaningfully and reason like human souls (p.134). In other words, it is human beings' ability to use words and produce meaningful answers that makes us intelligent souls, which are different from the brutes and automatons. Interestingly, Descartes completely ignores the automaton's capacity to produce sufficient non-verbal expressions, making the existence of the non-verbal expressions insignificant in meaningful conversations. Instead, the ability to use verbal language is privileged as the litmus test of intelligence.

Jerry Fodor (1975) uses his Language of Thought Hypothesis (LOTH) to propose that we think in a mental language called *Mentalese*. Mentalese is different from our natural language, but it is syntactically similar to natural language. In essence, the way we think is through a syntactically structured language in our heads. On the other hand, Dennett (1991), instead of postulating the existence of a different language of thought, claims that language reprograms the brain, which helps in constituting ourselves as the rational creatures we are. According to him, (verbal) language "infects and inflects our thought at every level…The [syntactic and semantic] structure of grammar enforce[s] a discipline on our habits of thought, shaping the ways in which we probe our 'data-bases'" (1991, p.301). I am well aware of the powerful effects of verbal language in the formation of thought; however, it is interesting to note the complete lack of attention to the roles of non-verbal

communication in both Fodor's and Dennett's works. The focus on the syntax-semantic viewpoint of language is characteristic of the trajectory of the philoseophical debate on the role of language in human cognition.

Even Churchland, in his proposal of a non-linguistic mode of mental representation, does not pay attention to the roles of non-verbal communication in human language and thought. In "Outer Space and Inner Space: The New Epistemology" (2002), Churchland argues that the fundamental form of representation, one that is common to all organisms with a nervous system, is "the activation pattern across a proprietary population of neurons" (p.27). This neuro-computational model is different from the linguistically structured modes of representation of Fodor and Dennett. Broadly speaking, Churchland thinks that the activation patterns of neurons is the fundamental mode of information processes in the brain, and it can support a multitude of representational information processing strategies. Despite proposing such a novel way of looking at human cognition, Churchland does not elaborate on how these non-linguistic structures may support non-verbal modes of communication. Instead, he focuses on how the activation patterns of neurons can support *linguistically* structured language, exemplifying yet again the privileging of verbal language in the philosophical debate on the role of language in human cognition.

Having briefly covered the works of Descartes, Fodor, Dennett, and Churchland, it should be mentioned that none of them denies the essential roles of non-verbal communication in human thought and language. . In fact, a systematic understanding of non-verbal communication can be incorporated into Churchland's theoretical framework. However, none of them focus on it either. It should be clear that Philosophy of Mind and Philosophy of Language have paid very little or no attention to the roles of non-verbal communication, much less any attempt at speculating a possible systematization of it. But why? Why do we ignore such a primal and important form of

communication? I can think of three plausible reasons for overlooking the importance of non-verbal communication in the current debate.

The first reason for the omission of non-verbal communication is a subconscious form of anthropocentrism. I do not think the scholars are intentionally overlooking the non-verbal aspects of human communication. Rather, the obsession of the debate with syntax and semantics mainly arises from the culture of thinking that spoken (or natural) language sets us apart from our primate cousins. It is safe to say that we suffer from subliminal levels of chauvinism as a result of our improved cognitive abilities over other animals. Many a time, we feel superior to our primate cousins because of our ability to indulge in linguistically structured communication. Hence, whatever sets us apart from others often emerges as the focus of all our attention. There is no doubt that the development of natural language has changed the evolutionary game-space for humans, making the odds forever be in our favor. Thus, it is natural that we focus so much of our attention on the distinctive avenue. However, the primacy of non-verbal modes of communication is undeniable. Most babies use all flavors of non-verbal communication when they start to acquire spoken language, pointing to the intricate relationship between verbal and non-verbal aspects of communication. Then, why should we stop paying attention to such a rich avenue? It is most certainly not in our best interests to let the subconscious anthropocentrism hinder our endeavor to learn more about ourselves and how we communicate.

The second reason for the lack of attention paid to non-verbal communication is the privileged position of the model of language and thought as automated syntax. As briefly touched upon in the first reason, our ability to use and participate in linguistic (or verbal) communication often sets our intellectual faculties apart from other animals. This has led to the development of the model of physical symbol manipulation—automation of syntax—as the dominant paradigm for

systematic understanding of thought. This is the practical implication of Alan Turing's work that classical functionalism and classical Artificial Intelligence latch onto. Only with the recent emergence of fields such as Affective Computing are we able to use statistical tools such Bayesian analysis in order to develop a different paradigm for the models of thought and language. Since the only acceptable mainstream model for thought and language has been automated syntax, which has been derived from linguistic communication, the primary thrust in research has been on the verbal aspects of communication with very little emphasis on the non-verbal aspects.

The third reason for such a bias in focus against non-verbal communication comes from a pragmatic and current state of the research standpoint. Emotions and non-verbal modes of communication are intricately related, for the latter acts as medium of conveyance of the former. On the philosophical front, we have long considered emotion to be a contaminant of rational thinking, a line of thought that derives from the Platonic or Kantian epistemological tradition. The Platonic tripartite model of the soul employs *logos* (or reason) as the ruler of the *thymos* (emotion) and *eros* (appetite). Moreover, "emotions have a stigma in science" and "the role of emotions is marginalized at best" (Picard, 1995, p.1). The scientific endeavor, usually conceived as the epitome of rationality and logical hypotheses, has long shunned emotion as the unwanted contaminant in research. Much of the research in Psychology and other Mind Sciences has focused on explaining models of the mind and language by staying within the bounds of current schools of thought. As a result of the demotion of emotion, there has been a relatively weaker push for scientific research on computational models and studies of the systematization of non-verbal communication and inference of mental states from multimodal affective states. The usage of "emotional state" and "affective state" are interchangeable with "affect" defined as an observable expression of emotion. We use language to generate, transform, and convey our thoughts. Language, considered in the proper *holistic* sense, is comprised of both verbal and non-verbal aspects. Since non-verbal

expressions carry important information needed to gain a deeper understanding of communication, a tunnel-visioned focus on the verbal aspects only presents us with a distorted picture. If we are to paint a holistic picture of human thought and language, we need to have a proper balance of verbal and non-verbal aspects in the mixture.

2: At the cutting edge— *Empirical work in Affective Computing and philosophical implications thereof*

We have now arrived at a critical point in the paper—I have addressed theoretical frameworks, gaps in current research and debate pertaining to non-verbal modes of communication, and reasoning behind their omission. The rest of the paper will focus on addressing the systematization of non-verbal modes of communication. Computational models of facial and vocal affects will be considered along with models of communication paying equal attention to verbal and non-verbal communication. These considerations will indicate that the neglect of non-verbal communication hinders our progress in gaining a deeper understanding of human language and thought.

2.1: The emotional being

Let me begin by briefly addressing the importance of emotion in thinking and decision-making and the intricate connection it shares with non-verbal communication. Although they have had a bad reputation in the past, "emotions are making a comeback!" (Joshua Greene, personal correspondence, 06/21/12). The emerging evidence of the role of emotion in moral decision-making (Greene, 2009) as well as its necessary roles in thinking and cognition (Picard, 1995) has catalyzed the lagging research on emotion. Recently, the disciplines of Affective Computing and Human-Computer Interaction have served as catalysts of the treatment of human emotions as

essential ingredients of cognition and decision-making. Rosalind Picard claims that "the neurological evidence indicates emotions are not a luxury; they are essential for rational human performance…[Thus,] emotions are vital for us to function as rational decision-making human beings" (1995, p.2). Research in neuroeconomics has investigated how people make decisions during buying and selling stocks. What was initially assumed to be calculated "rational" decision making processes was later found to be largely emotional "gut reactions" of the customers. The label of "irrationality" for emotions is now fading away. Emotions are themselves rational, in the sense that they embody (often quick) rational assessments that are necessary for good decision-making. For instance, research in neuro-economics investigates how emotions drive rational decisions of a Wall Street specialist during her transactions in the stock market. Thus, emotions are fundamental to most of our rational thinking when making moral and cognitive decisions. Words are instrumental in expressing our emotions—writers and poets have long exhibited that. However, it is the actors of the play which make the words come *alive* with their acting, a combination of verbal and non-verbal aspects. To reemphasize a point made earlier, non-verbal modes of communication (facial expression, prosody, gestures, etc. contribute significantly to the emotional aspects of communication. Hence, it is incumbent on us to strive to understand it better. With emotions making a comeback, research in non-verbal modes of communication will equip us with a better understanding of language and thought.

2.2: Empirical work in Affective Computing and Human-Computer Interaction

Now I will present empirical work done in Affective Computing and Human-Computer Interaction (HCI), focusing on facial, vocal, and bodily affects. Rosalind Picard, one of the most influential figures in the founding of Affective Computing, roughly defined it as "computing that relates to, arises from, or influences emotions" (1995, p.1). Human-Computer Interaction , roughly

speaking, focuses on issues pertaining to the interaction of the human with a computing system, which may be as simple as a tablet computer or a sophisticated robot. Research in HCI and Affective computing intertwine to present us with computational models that detect mental states from vocal and facial affects. I will primarily focus on the works of researchers at the Rainbow Group (Cambridge University), Affective Computing Group (MIT Media Lab) and Justine Cassell (Carnegie Mellon, Northwestern). I will also briefly draw on the work by Jonathan Gratch (USC). While reading about the empirical research presented, it is important that we keep our primary question in mind: how does a systematic understanding of non-verbal communication, especially by building computational models of it, augment our understanding of human communication? I will take brief interludes to explicate on the question as we proceed through the different cases of systematization.

2.3: Facial and Vocal Affects

In collaboration with Peter Robinson, Rana El Kaliouby and Tal Shikler at the Rainbow Group have done tremendous work in the inference of affective states from facial and vocal non-verbal cues (that is, the facial and vocal affective states). I will present the work done by the Rainbow Group in two parts. The first part will contain El Kaliouby and Robinson's work on facial affect while the second part will include Shikler and Robinson's work on vocal affect. I will also consider the implications of the research by considering individuals with Autism Spectrum Condition (ASC) and their impairment in non-verbal social signal processing.

Before proceeding to the research, I will provide a brief overview of Simon Baron-Cohen's Mind Reading DVD, which will give us background information needed for the section on the Rainbow Group research. El Kaliouby and Shikler et al. train their computer models with the videos

from Simon Baron-Cohen's Mind Reading DVD (2007). The development of the DVD is primarily driven by Baron-Cohen's "Theory of mind" or mind-reading—the ability to attribute mental states to others by observing their behavior (Baron-Cohen et al., 1997). The Mind Reading DVD is an interactive computer-based guide to emotions, developed mainly to assist individuals with ASC in recognizing facial expressions of emotions.

Baron-Cohen and colleagues have developed an elaborate taxonomy of emotions, which consists of 412 distinct human emotions (excluding synonyms) (Baron-Cohen et al., 2002), which goes far and beyond the basic six Ekman faces/emotions—happy, sad, afraid, angry, surprised and disgusted (Ekman et al., 1976). Broadly speaking, Paul Ekman claimed the "basic" nature and universality of these set of emotions as "basic" in the sense that they have agreed meanings within and across cultures. Ekman showed cross-cultural agreement between the basic six emotions by correlating them with facial expressions, whose inferences were similar across and within cultures. Baron-Cohen and colleagues push the boundaries of Ekman faces by grouping the 412 emotions into 24 different groups, where 6 actors play each of 412 emotions using both audio and visual modalities. There are three main sections of the Mindreading DVD—*the emotions library, the learning center, and the games zone*. Refer to *Figure 1* to get a sense of how one may use the DVD to learn more about an affective state. Individuals, especially with Asperger's syndrome (mild ASC), can be trained to recognize certain emotional states using the learning center and the library, and they get to test out their skills in the game zone. The training, comprising of understanding non-verbal cues, has improved the social signal processing abilities of children with ASC, which markedly enriched their quality of life in social situations.

Figure 1: Screenshots from the Mindreading DVD. The emotion under consideration is *hysterical,* which falls under the *excited group (one of the 24 groups)*. *Top left*: the 6 facial expressions of actions of different age groups. *Top right*: 6 vocal recordings of actors displaying the affective state. *Bottom Left*: Drop-down menu showing the emotions listed under the *excited group*. Retrieved from *http://www.jkp.com/mindreading/demo/content/dswmedia/MRF_Load.html*

Coming back to the research on Affective Computing, Peter Robinson and Rana el Kaliouby (2004, 2005) focus on facial expression (facial affect) by using a multi-level dynamic Bayesian Network (DBN) classifier that represents high-level cognitive mental states given facial expression and head displays. A Bayesian approach "to learning starts with some *a priori* knowledge about the model structure and model parameters" (Ghahramani, 1998, p.176). The model is trained on some known parameters (in our case, the Mind Reading DVD) and its performance is tested on new data (in our case, a mental state based on facial features). For our purposes (simplistically speaking),

DBNs are probabilistic learning models (a breed of statistical model) that represent a set of variables associated with mental states and their conditional dependencies. For instance, after training, one may ask the question: what is the probability that one is frustrated given a set of 4 facial features, say, pursing of the lips, tilt, knitting of eyebrows, and eye movement? The DBN can then calculate the conditional probability of frustration (the desired event) and give us a probabilistic output.

In their 2004 article "Mind Reading Machines: Automated Inference of Cognitive Mental States from Video", El Kaliouby et al. focus on 6 mental states groups: agreement, concentrating, disagreement, interested, thinking and unsure. Combining machine vision and supervised statistical machine learning, they "model hidden mental states of a person based upon the observable facial and head displays of that person" (2004, p.1). The facial actions are identified "from component-based facial features (e.g. mouth) comprised of motion, shape and color descriptors" (El Kaliouby et al, 2004, p.2) [refer to *Figure 2*]. They discuss how the head and facial expressions are combined with learning algorithms that are robust in generalizing and detecting the facial affects. Experimental results show "an average recognition rate of 87.4% for 6 mental states groups" (El Kaliouby et al., 2004, p.1).

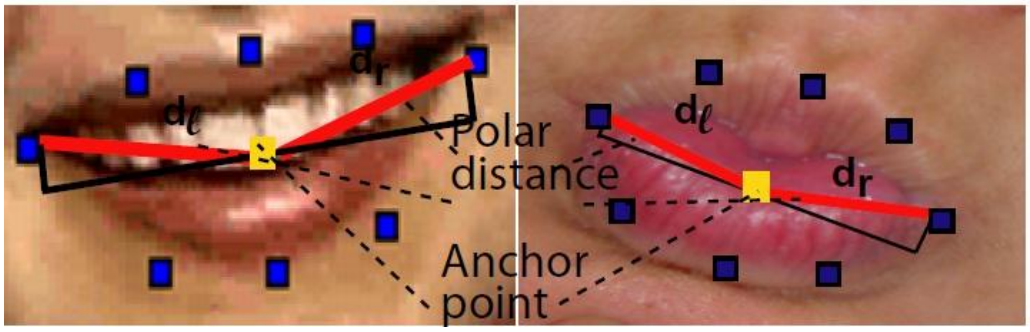

Figure 2: Facial feature extraction by component method. Note how the features points are used in each case to get the polar distance and anchor points which are later used to classify the features of the mouth. Adapted from "Mind Reading Machines: Automated Inference of Cognitive Mental States from Video" by R. El Kaliouby and P. Robinson, 2004 in *Proceedings of The IEEE International Conference on Systems, Man and Cybernetics*, pp. 1-7. Copyright 2004 by the Rainbow Group, Cambridge University, UK

In "Generalization of a Computational Model of Mind-Reading" (2005), El Kaliouby et al. extend their 2004 work by providing a vision-based computational model of mind reading "that infers complex mental states from head and facial expressions in ***real-time***" (p.1, emphasis added). The impressive aspect of the system lies in its inference of complex mental states that goes above and beyond the basic six Ekman emotions, enabling it to handle a greater variety of finely distinguished affective states. Instead of the basic six emotions, the system is able to analyze 412 distinct emotions sorted into 24 groups. Equipped with training from the Mind Reading DVD, "the results show that the system's accuracy is comparable to that of humans on the same corpus" (El Kaliouby et al. 2005, p.582) [refer to *Figures 3,4,5*]

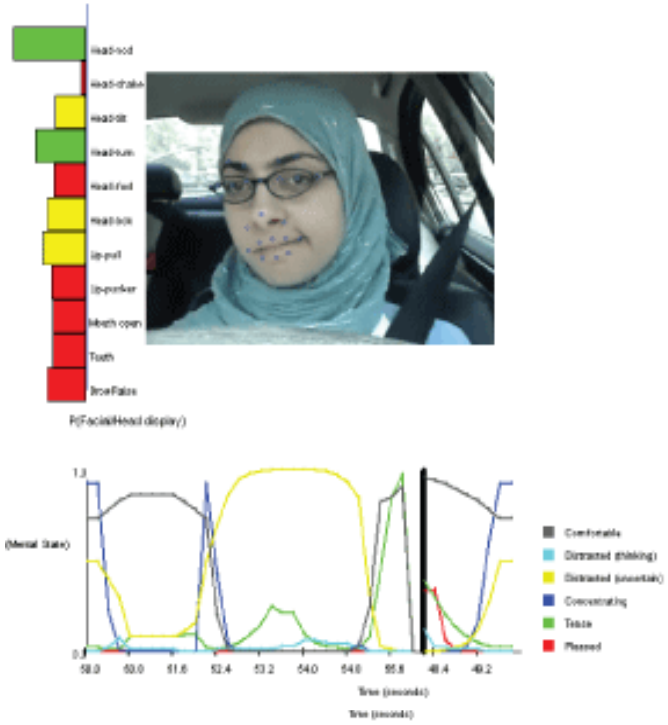

Figure 3: Monitoring a driver's affective states real time. A camera is placed in front of the driver (possibly using a head mount [not visible in the picture]) and the features from the face are used to give a probability distribution of the possible affective states of the driver. Adapted from "Mind Reading Machines" section of the Rainbow group's website http://www.cl.cam.ac.uk/research/rainbow/emotions/mrm.html Copyright 2005 by the Rainbow Group, Cambridge University, UK

Note: The figure is provided to give an illustration of the user-interface of the system, hence, inability to read the details on the axes is not problematic for all intents and purposes

El Kaliouby and colleagues tested the generalizability of the models during a CVPR (Computer Vision and Pattern Recognition) conference, yielding impressive success rates compared to human performance on the same data set. 16 conference attendees were asked to enact six mental states: agreeing, concentrating, disagreeing, interested, thinking and unsure (*Figures 4, 5*). No

instructions were given on *how* to express the mental states; rather, for labeling the videos, volunteers were asked to simply name the mental state immediately before they would act out the affective state. After establishing a proper human baseline, the mind-reading system was tested on 88 videos of the CVPR corpus. The overall accuracy of the system is 63.5%. Compared to human performance on the same set of videos, "the automated mind-reading system scores among the top 5.6% of humans…[and] generalizes well... to new examples of mental state enactments, which are posed (and labeled) by lay people in an uncontrolled setup" (El Kaliouby et al, 2005, p.588). In the process of training and inferring mental states, the system is evolving the robustness and generalizability of its inference from facial affects. It is safe to state that, through our life experiences, we as humans, gather correlations of various facial and vocal features and their respective affective inferences. However, we often do not discursively express exactly *how* we infer them. By modeling the detection and inference process in these systems and given its success, we can then extract interesting regularities (or irregularities) in the patterns of the feature points on the face (see *Figure 2*) and the corresponding inferences. Based on the research presented, I believe that a systematic understanding of non-verbal modes of communication, especially facial expression, is well on its way with substantial success already under its belt!

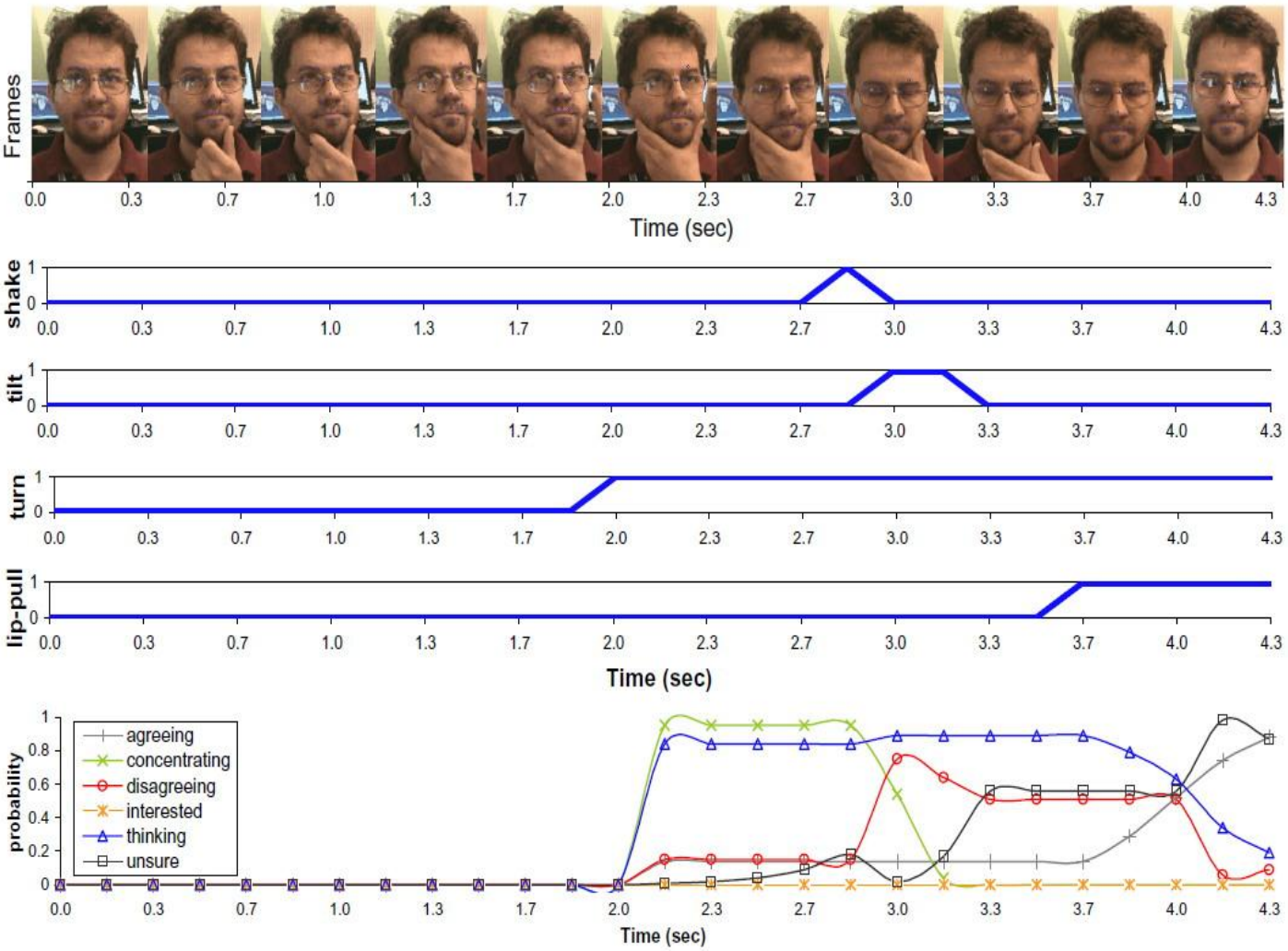

Figure 4: Mental state inference: (top) frames from a video labeling as *thinking* (CVPR corpus); (middle) head and facial displays; (bottom) mental state inferences. Adapted from "Generalization of a Vision-Based Computational Model of Mind-Reading" by R. El Kaliouby and P. Robinson, 2005 in *Affective Computing and Intelligent Interaction*, pp. 582-589. Copyright 2005 by the Rainbow Group, Cambridge University, UK

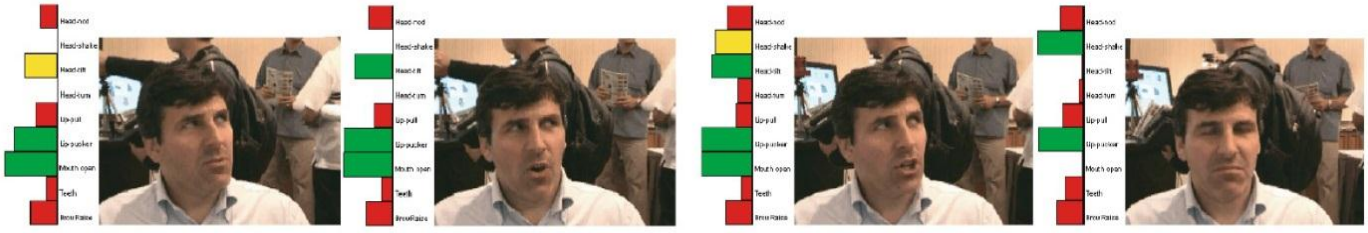

Figure 5: Real time display recognition (frames sampled every 0.7s). The bars represent the output probabilities of the HMM classifiers (top to bottom): head nod, shake, tilt, turn, lip corner pull, lip pucker, mouth open, teeth and eye-brow raise. Adapted from "Generalization of a Vision-Based Computational Model of Mind-Reading" by R. El Kaliouby and P. Robinson, 2005 in *Affective Computing and Intelligent Interaction*, pp. 582-589. Copyright 2005 by the Rainbow Group, Cambridge University, UK

Moving on to the inference of mental states from vocal affects, let us take a deeper look into the work of Shikler et al. (2010). Instead of using the commercially available Mind Reading DVD, the researchers used an experimental version, which had 700 affective states arranged into 24 groups. Shikler et al. used a classification algorithm called Support Vector Machines (SVMs) to conduct independent pair-wise comparisons between nine affective-state groups—joyful, thinking, concentrating, stressed, excited, opposed or disagree, interested, confident or sure, and unsure. (Perhaps over-generalizing,) SVMs are supervised learning algorithms similar to DBNs. They are efficient at classification problems on a given data set (see *Figure 6* for a simple illustration). Just like DBNs, once the SVMs are trained, they can be used to classify new data sets. The input set consisted of a large set of vocal features and, after each utterance, metrics were extracted for comparison.

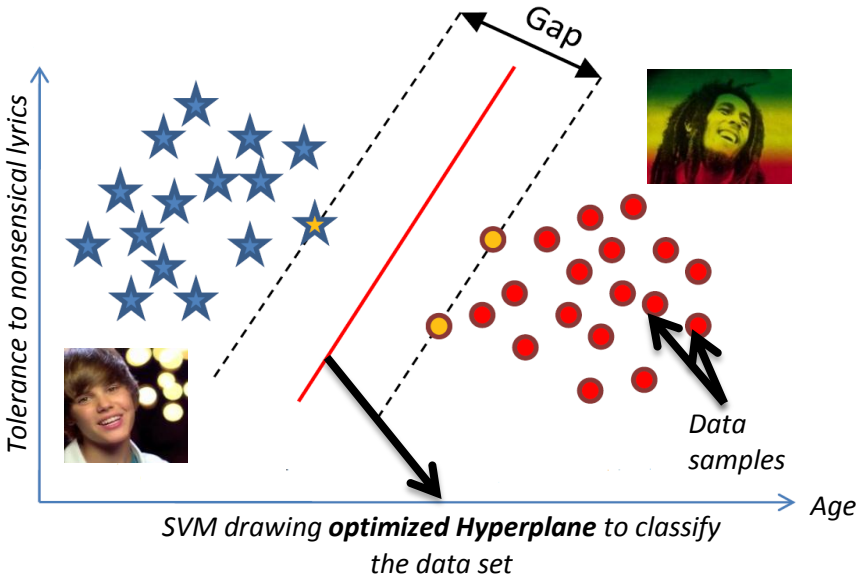

Figure 6: Illustrating principles of SVM operation on a hypothetical set of data– the blue data points correspond to Bieber fans while the red ones refer to Marley fans

For Shikler's experiment, the output of the system consisted of a consolidated list of single-ranked list of the nine-affective state groups. Shikler et al. claim that "average classification accuracy of the 36 pairwise machines was 75%, using tenfold cross-validation" (2010, p.1). The results of this research are impressive, for the distinguishing capabilities of the system are comparable to human

performance. Limitations of previous research include the inability to detect affective states that are co-occurring. Examples of co-occurring affective states may include happiness and interest coinciding during the time-period, making the detection a challenging problem. Apart from extracting mental states from vocal affects, the system performs credibly when dealing with co-occurring affective states (Shikler, 2001, p.2). The engineering and computational leap of the algorithms indicate bright future prospects for the inference of vocal affects.

Another important aspect of the empirical work done is to realize that the generalization and inference of the affective states from facial and vocal affects are *not* look-up table in nature. That is, the system is not looking up the answer in a pre-established catalogue; rather, it is relying on the statistical patterns it has extracted from past experience (in the form of training data). It relies on "learned" background "knowledge", where the model's sense of "learning" and "knowing" are hypothesized to be similar to (but perhaps not as rich as) human's sense of learning and knowing. Thus, it is *not* the case that there is a huge stored database and the algorithm goes in and searches for the entry that matches the new input. Rather, it *is* the case that, once the model (DBN or SVM) is trained using the Mind Reading DVD and a new data point is given to classify (the affective state) using feature points (of the face or the voice), the algorithm uses the previous learning to classify the new data point based on its conditional probabilities. Therefore, in a very naïve way, one can think of the algorithms as evolutionary learning algorithms, the more pertinent data points one feeds the algorithm, the better it *usually* is in terms of making a classification.

2.4: Checkpoint—*Why do we need computational models?*

We have now arrived at a check point where we can relate the empirical research to the big-picture philosophical question. One may wonder: Why do we need to build computational models

of non-verbal communication anyway? The researchers are trying to build better human-computer interfaces, so how does that help us to gain a deeper understanding of human-human communication? Excellent questions! Read on…

Let's say you are asked to give a list of all the things you would want in a human-human communication. In essence, you are asked, what *is it* that makes the communication human? An exhaustive list may be hard to formulate given the numerous factors in a human-human interaction. Hence, let us think of another way at getting at the issue. Now, instead of being asked to make a list, imagine that you were asked to interact with a robot, which has the ability to synthesize speech and generate bodily actions. Given sufficient experience with human-human interaction, it will be easy for you to notice the discrepancy that exists between HCI and HHI. Thus, I believe that it will be far easier for you to pick out the things that you *wished* the humanoid did such that the communication would be more like human-human interaction. That is, the interaction with the robot evoked the *humanness* in you during the conversation. It is similar to figuring out what you *wish* your current phone could do in order to build the perfect phone. This "method of difference" outlook on the HHI and human-humanoid interaction distills the missing ingredients needed to make the communication *human*. In order to upgrade and improve the interaction, the company models the required attributes the humanoid previously lacked and adds them to its personality matrix. The enhanced humanoid now converses with you, utilizing both verbal and non-verbal actions, in a much more *humanized* way.

In the process of upgrading, something interesting has happened here—it is precisely through the modeling of the previously lacking attributes that we gain a deeper understanding of *them*; otherwise, the modeling would fail. Moreover, the required attributes, by definition, are the missing pieces that you think will make the interaction more humanlike. Thus, the modeling of them

not only made the interaction more human, but they also empower us with a deeper understanding of those attributes. Overall, the entire process enriches our notion of what *it is* that makes the communication *human.* Although, on the surface, it looks as if researchers like El Kaliouby and Robinson are *just* building better human-computer interfaces by building computational models of non-verbal communication, the implications of their research reach the deeper depths of human-human communication, augmenting our understanding of it. That is why it is integral that we understand and develop these models in order to succeed in our quest of attaining an enriched understanding of human communication.

2.5: Autism and a systematic understanding of non-verbal communication

Apart from the aforementioned philosophical implication of the research in itself, it has applications in the assistance of social signal processing abilities of individuals with ASC (or its milder cousin, Asperger's Syndrome). The systematization of non-verbal communication can augment our social signal processing abilities, especially for those who have difficulties in such processing. As El Kaliouby et al. claim, "while subtle and somewhat elusive, the ability to [process social signals] is essential to the social functions we take for granted" (2004, p.1). With impairment in their abilities to interact in complex social environments, autistic individuals "need to be taught explicitly how to read other people's mind from nonverbal communication channels" (El Kaliouby et al, 2005, p.1). If we are to improve the quality of lives of individuals lacking social signal processing skills, the present lack of assistive tools for autistic individuals makes it incumbent on us to develop wearable computing devices that can assist them. Such advancements will significantly reduce the otherization that autistic individuals experience in society, enabling us to create an atmosphere where empathy and understanding of the other is the norm instead of the exception.

El Kaliouby and Peter Robinson respond to this need in their 2005 article "The Emotional Hearing Aid: An Assistive Tool for Children with Asperger's Syndrome" wherein they propose the development of "a portable assistive computer intended to help children diagnosed with Asperger syndrome read understand and react to facial expressions in a socially appropriate ways" (p.21). The device draws inspiration from El Kaliouby et al.'s work on facial affect inference (2004, 2005) as well as the "emotional indexing" method, an approach for teaching children with autism how to read and respond to emotions. The usefulness of such devices is contingent on their real-time processing abilities of affective states. The hearing aid aims to provide "real time assistance with reading facial expressions of other people and advice on reacting to it in a child's natural social environment" (El Kaliouby et al, 2005, p.6). It consists of a personal digital assistant (PDA), an earpiece speaker and a wearable camcorder (*Figure 6*). There is a reaction advisor which suggests an appropriate reaction to the other person's affective state (*Figure 8*). The accuracy of the automated mind-reading device was tested using the Mind Reading DVD with "the overall accuracy of the system compares favorably to that of human performance on a similar recognition task" (El Kaliouby, 2005, p.20). However, the researchers correctly claim that a 77.4% success rate is not sufficient for such real-time applications.

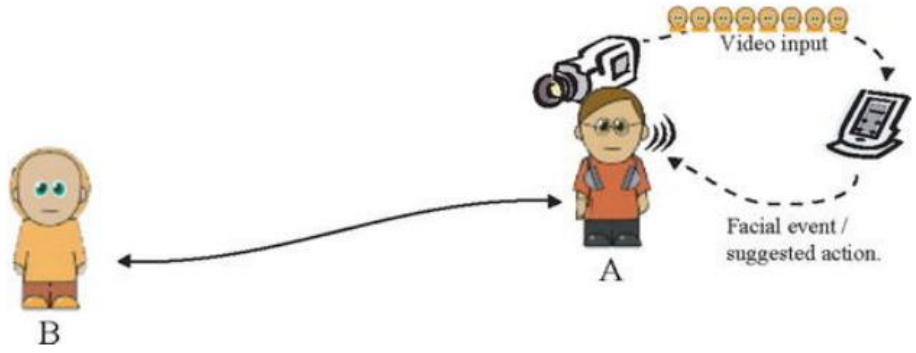

Figure 7: A model of operation of the Emotional Hearing Aid. Child A (diagnosed with Asperger's Syndrome) is using the emotional hearing aid in an interaction with person B. Video sequences of B and situational context cues are sent to the PDA for analysis, and suggested reactions. Depending on the mode of interaction, the output can be visual or audio, and can vary in the degree of detail presented. Adapted from "The Emotional Hearing Aid: An Assistive tool for Children with Asperger's Syndrome" by R. El Kaliouby and P. Robinson, 2005 in *Cambridge Online Archives. Retrieved December 26, 2012, from* http://www.cl.cam.ac.uk/~pr10/publications/uais05.pdf . Copyright 2005 by the Rainbow Group, Cambridge University, UK

Figure 8: A sequence of screenshots from the reaction advisor in the Emotional Hearing Aid. Adapted from "The Emotional Hearing Aid: An Assistive tool for Children with Asperger's Syndrome" by R. El Kaliouby and P. Robinson, 2005 in *Cambridge Online Archives. Retrieved December 26, 2012, from* http://www.cl.cam.ac.uk/~pr10/publications/uais05.pdf . Copyright 2005 by the Rainbow Group, Cambridge University, UK

Although there are many avenues of improvement for the Emotional Hearing Aid, its development has important implications for the research on a systematic understanding of non-verbal communication. Given the current rate of progress, it is not unreasonable to envision a future in which the facial, vocal, and bodily modalities are combined to infer mental states. Effects of enculturation could be incorporated in the learning models with contextual cues taken from surroundings as real-time data processing occurs over cloud computing. Although there are several engineering and technical challenges in the integration of modalities (cf. Shikler et al. 2008), there has been significant progress in data-processing over the internet through the development of Affdex. Affdex is a part of Affectiva (www.affectiva.com/affdex), a startup from the MIT Media Lab formed by El Kaliouby and Picard. It reads emotional states using any webcam and uses state-of-the-art machine learning and computer vision techniques to infer mental states from the facial affect. Anyone with access to the internet and a webcam can use it. The commercialization and success of such technologies is a good indicator of its progress and credibility in the long run.

Advancements in the systematization of non-verbal communication can reach a level where the deficit experienced by the child with ASC is completely eliminated as his/her abilities to process social signals is no less than that of a neurotypical (individual with normal brain functioning). I can easily imagine a future in which autistic individuals are wearing Google Glasses that process social signals, facilitating their communication with others. "Google Glasses" is another name for the wearable computing device recently developed by Google under Project Glass, which can be used as an augmented reality device as well as your phone on the go. (click here to see a demo)

2.6: Checkpoint—*Bridging the Neurotypical-Autistic Divide*

We have reached another checkpoint where we can pause for a moment to reflect on the importance of the empirical research in non-verbal modes of communication and its applications in

the assistance of individuals with ASC. Consider Mahir, a neurotypical who has average abilities in processing facial and vocal affects and responding in a socially appropriate manner. He can navigate through social gatherings, utilizing his naturally acquired heuristic understanding of non-verbal modes of communication. Now think of Ahnaf, an individual with ASC, one whose abilities to process social signals are impaired. Enhancing Ahnaf's social signal processing with Google glasses, equipped with the appropriate affective processing abilities, can reset the deficiency in processing social signals. Enhancement of the social signal processing abilities also augments his thinking and decision making abilities. Hence, the non-verbal modes of communication are instrumental in sustaining meaningful human-human interaction, playing a vital role in the conveyance of thoughts.

What the Ahnaf-Mahir example delineates is the stark contrast that exists between neurotypicals and individuals with ASC in terms of their social signal processing skills. The difference in their abilities determines whether their navigations through social interactions are smooth or rough. More importantly, their abilities to communicate properly depend on their abilities to process the social signals. For individuals with ASC, the absence of the ability to process mainly non-verbal communication denies the autistic individuals the pleasure of multimodal, enriching human-human interactions. Therefore, this "method of difference" way of looking at the problem makes the vital role of non-verbal communication, which has long been ignored by typical Philosophy of Language, evident in human communication. However, the void left by the cognitive impairment can be filled by the technology derived from a systematic understanding of non-verbal communication. Thus, it is important that we pursue the development of assistive technologies not only to benefit the lives of non-neurotypicals, but also because it empowers us with a deeper understanding of human-human communication.

2.7: "Grounding" a conversation and establishing rapport—*Embodied Conversational Agents*

Apart from applications in the assistance of individuals with ASC, proper implementation of *both* verbal and systematized non-verbal communication on artificial systems augments our understanding of human-human interaction. Justine Cassell has done laudable work in the development of Embodied Conversational Agents (ECAs). ECAs are "cartoon-like, often life-size, depictions of virtual humans that are projected on a screen" (Cassell, 2007, p.5). ECAs can be used as virtual tour guides and virtual peers to assist classroom learning amongst neurotypicals and children with ASC. The development of NUMACK (Northwestern University Multimodal Autonomous Conversational Kiosk) (*Figure 9*) supports the claim that an optimal conglomeration of verbal and non-verbal communication is instrumental in the human communication enterprise. NUMACK is an ECA who interacts with humans by providing directions around the campus, generating novel language and gestures in coordination with a computation model of language and gestures (Striegnitz et al. 2009, NUMACK (n.d)). One of the most impressive features of NUMACK lies in the generation of its verbal, non-verbal and multimodal behaviors—they are generated using a kinematic body model and automatically synthesized speech (Striegnitz et al. 2009, NUMACK (n.d)).

Figure 9: NUMACK, our ECA, producing (a) a route perspective gesture, (b) a non-locating gesture, (c) a survey perspective gesture. Reprinted from *Spatial language and dialogue* (p.14) by K. Striegnitz, P. Tepper, A. Lovett, & J. Cassell, 2009, Oxford: Oxford University Press. Copyright 2009 by authors.

Prior to the development on NUMACK, Cassell and Thorisson (1999) looked at human interactions with a less developed ECA, named Gandalf, consisting of a head with one disembodied hand. The studies done with Gandalf involved three versions: the first version consisted of "content only" where the ECA only spoke with no non-verbal communication; the second version called "content + envelope" involved Gandalf speaking and using eye gaze and brow movements with occasional hand waving; the third version, "content + emotion" incorporated activities of Gandalf like speaking, smiling, frowning, and looking puzzled (Cassell, 2007, p. 14-15). The interacting humans were asked to fill out questionnaires assessing the "lifelikeness" of the three versions of Gandalf. Cassell et al. found that "participants tended to mimic the virtual human: if he stood rigid, so did they; if he was animated, so were they" (2007, p.16). People interacting with the first version were "most animated in their expressions of frustration" (2007, p.16) whereas they interacted more positively with the second and third versions. This finding substantiates the claim that a proper balance of verbal and non-verbal communication is needed to establish a meaningful, *human* conversation. Hence, the lack of attention towards non-verbal communication has only hindered our progress towards understanding human communication in depth.

In addition to eliciting a positive reaction from the interacting human, non-verbal behaviors play an important role in the phenomenon dubbed "grounding". Grounding refers to the ways in which interlocutors ensure mutual understanding of each other, updating the common ground. Common ground is often referred to as "the sum of mutual knowledge, mutual beliefs and mutual suppositions necessary for a particular stage of a conversation (Clark 1992)" (Cassell, 2007, p.16). Non-verbal actions such as nodding to indicate one is following the conversation enable grounding to happen. Cassell et al. tested the reactions of people to two versions of an ECA, one with grounding turned off and the other with grounding turned on. With grounding turned off, the human acted as if they were simply in front of a kiosk and not another human; however, "when the

ECA did engage in grounding behaviors, the human acted strikingly…human, looking back and forth between the map and the ECA" (Cassell, 2007, p.16). This result underscores the integral roles of non-verbal communication by showing the increased meaningfulness in the communicative abilities once grounding is turned on.

Non-verbal modes of communication, especially facial expressions and prosody, play an integral part in establishing rapport amongst humans. The Oxford English dictionary defines rapport as "a close and harmonious relationship in which the people or groups concerned understand each other's feelings or ideas and communicate well" (Oxford Dictionaries online entry). Rapport is essential for success in developing friendships, negotiations, classroom performance and other human-human interactions (Gratch, (n.d.)). Jonathan Gratch and colleagues at USC have done noteworthy work in developing computational models of emotion (Marsella et al. 2010) and incorporating systematized verbal and non-verbal communication in Virtual Humans (similar to Embodied Conversational Agents like NUMACK). In "Can Virtual Humans Build Rapport and Promote Learning?" (2009), Wang and Gratch investigate "the effectiveness of nonverbal immediacy using a virtual human" (p.1). Analyzing facial expression and gestures, the virtual human uses machine vision and prosody analysis to establish rapport with the learner (in a classroom environment). Results indicate that the virtual humans successfully established rapport with the learners and also suggest that "creating rapport is related to higher self-efficacy, and self-efficacy is related to better learning results" (Wang and Gratch, 2009). In trying to build computational models of emotion, Gratch and colleagues have distilled the need to have a systematic understanding of non-verbal communication. In succeeding to establish rapport, mainly by utilizing non-verbal actions, Virtual Humans delineate the importance of non-verbal aspects in learning environments where effective communication is crucial. Not only do the findings help us create more effective learning environments by training our teachers to establish rapport with their students, but the

modeling also helps us unearth one of the key ingredients that make human-human communication enriching. In turn, the research into a systematic understanding of non-verbal communication gives us a deeper understanding of *ourselves*, especially in the way we communicate.

3: The question about ourselves and the power of embodied language

Till now, we have looked at empirical research showcasing the importance of a systematic understanding of non-verbal communication as well as the equal importance of verbal and non-verbal communication in human-human interaction. We have looked at how the development of computational models of affective states assists both neurotypicals and non-neurotypicals in attaining a rich, multimodal human-human interaction. We have also briefly looked at the philosophical implications of the empirical research on a systematic understanding of non-verbal communication. At this point, one may legitimately wonder: why bother creating all these computational models when we can just observe humans? Why should we invest our energies in instantiating the non-verbal and verbal aspects in artificial systems? What are we trying to do by building these artificial systems? Earlier, I had briefly touched upon the philosophical implications of the empirical research in connection to the question of "what it is that makes communication *human*?" Now, I will respond to the questions posed in a more general sense using a hybrid approach comprising of my views merged with some important points made by Justine Cassell.

I believe that building computational models of non-verbal and verbal communication and their instantiation in artificial systems enhances our understanding of human communication. If we are to instantiate a phenomenon, say X, into an artificial system, it is incumbent that we know what it *is* that we are trying to instantiate. For instance, consider the simple cases of algorithms that perform mathematical operations in software like MATLAB, Maple, etc. Only after achieving a

critical level of understanding can we write the algorithms for the operations. Moreover, upon successful implementation of the algorithm, we have the luxury of carrying out further complex operations using the computation power of the machines. Some of the operations, like simulation of models, could not have been carried out had the algorithm of the model not been written. The ability to tinker around with the phenomenon with greater control over parameters augments our understanding of the concepts that are modeled.

Similarly, the case of a systematic understanding of non-verbal communications and its instantiation in artificial systems is no different. The success in the research on Affective Computing and HCI has produced novel findings about the way humans communicate, focusing on the principal question: what is it that makes communication *humanlike*? For instance, when Cassell and colleagues realized that "the phenomena of hand gesture, intonation and facial expression [is]…derived from one common set of communicative goals,… [the] result fundamentally [not only] changed the way [they] build embodied conversational agents, but it [also] was an advance in understanding human communication" (2007, p.12). Moreover, the development of computational models and artificial systems often enables greater scope and control over experimental analysis. It is convenient to tinker the simulation of the model and observe their predictive abilities, enabling us to observe the problem space under multiple scenarios in a short amount of time.

Apart from the successful implementation of the models, *we can also learn a lot from their deficiencies*. I share Cassell's enthusiasm and viewpoint that we "learn [so much] about human behavior when [we] try to recreate it – in particular when, and because, [our] imitations are partial and imperfect" (2007, p.3). We have observed that if the artificial system can enact many of the "humanlike" actions, such as maintain eye contact, follow gaze, generate gestures and facial expressions, then the interactions becomes more human. Hence, by modeling it in artificial systems

and watching how a human interacts with it, we are not only recreating the actions in the artificial system, but gaining a deeper understanding of the actions while modeling it.

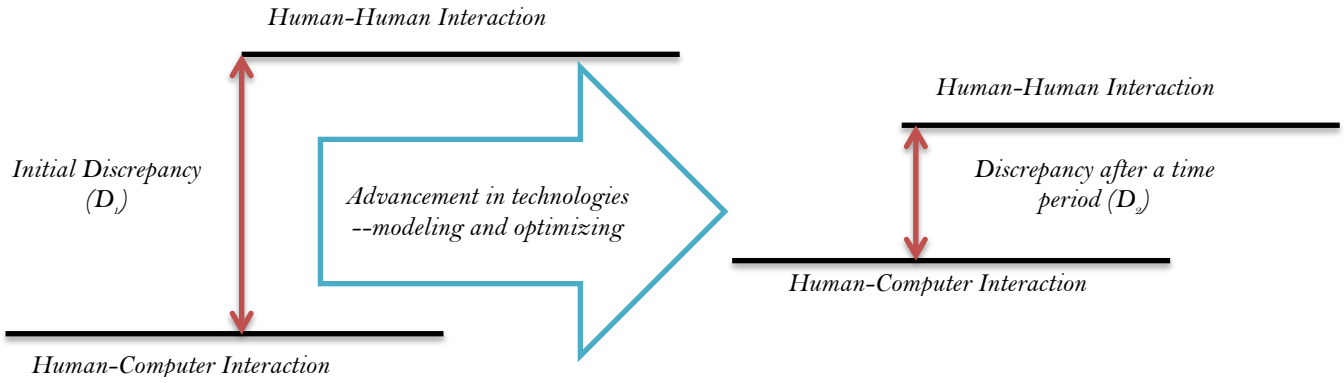

Figure 10: The decrease in the discrepancy between HCI and HHI as advancement in technology continues over time. In the process of decreasing discrepancy and modeling, we gain deeper understanding of the aspects, whose addition to HCI, lessens the gap.

The question, "what *is it* that makes the communication human?", is of interest and importance in Philosophy of Language and Mind. I believe that, by minimizing the deficiencies (see *Figure 10*) between Human-Computer Interaction (HCI) and Human-Human Interaction (HHI), we are getting closer to answering the question. Success in modeling and filling in the discrepancy as a result of technological advancements lessens that gap between the initial discrepancy (D_1) and the discrepancy after a time period (D_2). Although we may not succeed in perfectly matching the HCI and HHI levels and the HHI level may remain forever as an asymptote, in the process of trying to minimize the gap, we are gaining essential knowledge about ingredients that make the conversation more *human*. In turn, we attain a deeper traction on the question posed above, enriching our philosophical understanding of it. Furthermore, instead of thinking whether we can build an artificial human with the development in the current research, I tend to think down the same lines as Cassell. The question should *not* be whether we can build an artificial mind; rather, the question should be:

"what can we learn about humans when we make a machine that evokes humanness in us – a machine that acts human enough that we respond to it as we respond to another human?" (Cassell, 2007, p.4). Successful instantiations of artificial systems "evoke distinctly human characteristics in our interaction with them" (Cassell, 2007, p.14). Thus, the empirical work in the sciences is instrumental in answering the question that is essential in understanding our own nature.

Language, viewed as a beautiful and intricate synergy of verbal and non-verbal aspects, has transformative powers in the formation, development, and conveyance of thought. Andy Clark (1998) rightly claims that language transforms and simplifies the problem space of cognitive function for an agent, playing an integral role in the mechanisms of thought. At the start of the paper, I claimed the embodied nature of communication, explaining how the facial expressions, prosody, and gestures add deep contextual layers that are intricately related to human thought. Throughout this paper, I have demonstrated the crucial roles of non-verbal communication in human thought. I have also commented on how a proper balance is needed between verbal and non-verbal aspect to get a deeper understanding of a communication. Communication is a fluid and dynamic process in which thoughts are not merely conveyed; rather, they are formed, developed and conveyed throughout the interaction. The richness and understanding of a proper communication arises from this dynamic interplay between thoughts of the conversational agents.

Moreover, the human being as the embodied agent actively couples (a deep dynamic relationship) with its environment in the way it processes and conveys information. Non-verbal expressions, as discussed before, play an instrumental role in the coupling process. In turn, along with the verbal aspect, non-verbal aspects form an integral part of the mechanisms of thought in the human mind. Think about the many times when incorporation of bodily gestures and movements not only helped explain a Physics problem, but it also helped the receiver to understand it properly.

A theory about human thought that does not incorporate a *holistic* view of language is essentially ignoring a rich pool of information about human thought, painting a blind-sighted version of the picture. Thought is not mere syntax or semantics; rather, it is embodied, enactive, rich in emotion and context. Engaging in a rich conversation with someone may, at first, seem as if the connection is only from the outside. However, by seemingly connecting well from the outside, we gain access to a deeper understanding of the activity of each other's minds inside. Thus, if philosophy of language and mind incorporates a holistic viewpoint of language, the epistemic blinders that long hindered its progress will be removed. More importantly, we will have a much richer and intricate theoretical framework to work on. Coupled with computer models that can provide insight on the inner workings of the brain, an enlightened philosophy of mind and language unleashes the human potential in terms of discovering new and more interesting things about us.

4: Objections and Responses

We have now reached a point in the paper where it is appropriate to address concerns and objections to the present argument.

Objection 1: Current research shows that inference of mental states from facial expressions is extremely limited since our emotions are highly context-dependent, in which our facial expressions, prosody, gestures are integrated. For instance, pictures of facial expressions are often devoid of context, preventing us from making a correct interpretation of the mental state. Hence, unless we learn to integrate the modalities, we benefit from a systematic understanding of non-verbal communication.

The need for the integration of the modalities for improved inference of affective states is an astute observation. Indeed, the embodiment of our communication is intertwined in the modalities

of non-verbal expressions—facial, vocal, and bodily expressions. The concern raised, however, is a challenge more to the engineering aspect of the developments than the philosophical thrust behind it. Currently, the aforementioned research cannot integrate the modalities in real-time. Shikler et al. (2004) mention the engineering challenges behind the integration of the modalities and propose a few models to overcome the issue. Therefore, at present, the problem is an open empirical research question. The importance of non-verbal communication in achieving a deeper understanding of human-human interaction has already been discussed. Granted that current research cannot integrate the modalities yet, it does not, by any means, entail that we cannot benefit from a systematic understanding of non-verbal communication. This is especially true when we consider the current success of the research in the development of assistive technology for people with ASC, which has markedly improved their social signal processing skills. Thus, it is in our best interests to refrain from broad, sweeping claims about the benefits (or lack thereof) of a systematic understanding of non-verbal communication. Premature, knee-jerk negative reactions can impede the progress of research into an avenue that equips us with a novel of philosophical approach to human communication.

Objection 2: The current computational models are oblivious to the effect of enculturation on the production and inference of non-verbal communication. Hence, the algorithms are not generalizable and the scope of the model is limited to subcultures of the Western part of the globe.

This is an excellent point. It is certainly true that cultural construction has important influences on the ways we communicate. The enculturation of an individual influences the production of affective states and subsequent inference of the states. For instance, certain cultures encourage the inhibition of outward anger in their non-verbal expressions. If the cultural variability is not taken into account, the present ability of the computational models to infer the affective states

is put to question, especially when applied to non-Western parts of the world. However, just because it has not been done yet does not entail one cannot do it. Advancements in data mining technologies combined with an increased awareness of cultural variability are increasing the feasibility to conduct research projects from different parts of the world. Therefore, the models can be made region specific and trained with appropriate data set from that region. Although it is not a trivial problem, it is conceivable to think of models that can indulge in cross-cultural inferences after a certain period of experiential training just as humans do if they have experiences in different cultures. In fact, in the near future, I plan to pursue the incorporation of cultural variability and enculturation in these models by working with Jon Gratch and Louis-Philippe Morency at the Institute of Creative Technologies, USC.

Objection 3: the inference from the affective states can never be perfect since all of us express ourselves in different ways. Hence, the attempts to infer the mental states using these computational models are futile.

There is no doubt that not all of us express ourselves the same way. Given this realization, it is true that the inference of the algorithms will never be perfect. However, the question is: do we, as humans, have perfect inference of these states? No. It is precisely the diversity in expressing ourselves that prevents us from having a normalized inference of affective states. However, the diversity in our expressions requires us to possess a heuristic and systematic idea of inferring the mental states from non-verbal communication in order to facilitate social interactions. Surely, we can go wrong, but as our own neural networks in the brain are trained with experience in interactions, our social signal processing skills improve. Hence, it is unfair to expect a computational model to achieve perfect performance, especially when humans are unable to do so. To proceed from the imperfect nature of inference to claim that it is futile to even attempt the development is, at the least,

a premature proposition. The point of the development is not to achieve perfect inference; rather, it is to gain a heuristic and probabilistic understanding of non-verbal communication, which is akin to how humans understand it. The training of the models with appropriate data (in real life, these would be experiences) increases their predictive abilities. Thus, the development of these models is quite opposite to being futile. It is essential to our understanding of human thought and communication, for it empowers us with a holistic viewpoint of human-human interaction instead of tunnel-visioning ourselves with a syntax-semantic viewpoint.

Objection 4: The models do not *really* understand the affective states, they are *merely* displaying top-level mimicking of the inference ; therefore, there is no need to build these models since they will never *truly* understand the states.

This objection falls under a broader category of objections that have been postulated against any sort of artificial intelligence and has its root deeply cemented in a Cartesian framework. One potential way to address this concern is to bite the bullet and ask, "How do we, humans, know that we aren't *merely* mimicking the inference? What confirmation, other than self-reported confirmations as a species, do we have of the fact that we *truly* comprehend the affective states?" I do not think a satisfactory response to these questions will be something like, "Because we are humans, and we *know* that we *truly* understand it". This type of answer depends heavily on the questionably arbitrary metric of intelligence that we put on ourselves. Just because we are humans does not entail that the way we understand things is the *only* way those things could be understood in a genuine sense of understanding. Moreover, to expect a foreign entity to carry out "x" the same way we perform "x" is a form of naïve anthropocentrism.

This objection would have had a lot more impact had the aforementioned computational models deployed a mere look-up table method to infer the mental states. If the models did utilize

such a simplistic mechanism, it is plausible to surmise that, in fact, the models are displaying mere top-level mimicry. However, our models are anything but that. As mentioned earlier, using pertinent training data, the models use sophisticated statistical techniques to infer the affective states. Roughly speaking, they gather predictive power by deploying a probabilistic model trained on appropriate data (read *experiential knowledge* for humans) to infer the mental states. It is as if the models were using their past "experiences" as a guide to make future inferences, which is a good model of how we think humans work in general. Thus, they have a fairly complex underlying structure that governs their performance. There is also a level of biological plausibility in the way these models infer affective states compared to the way we think humans do it. Thus, a fairly complex underlying structure of the algorithms should allay some of the worries about top-level mimicking.

Moreover, mere top-level is very unlikely to yield the high success rates (87%) of the models compared to human performance (El-Kaliouby et. al, 2005). Such a high level of performance can only be expected when the underlying structures of the computational models are behaving as initially predicted. In achieving reliable performance, we are presented with models that have a complex underlying structure. Moreover, they should also capture a fair amount of the cognitive functions that underlie human performance in similar tasks of inferring non-verbal cues. The degree to which the modeling captures the underlying human performance is an open empirical question, one that can only be answered by conducting more research. However, given the biological basis of the modeling and the current empirical success, it is reasonable to expect positive results in the future.

Now that the mere top-level mimicking objection has been addressed, I would like to draw our attention to an interesting avenue created by the present discussion. I have just pointed to the open empirical nature of the question: are the underlying functional structures— the heuristics, the

strategies, etc. — of the models similar to the way the brain carries out the functions in humans? If further empirical research shows that the answer is *yes*, then we have gained a deeper understanding of not only the crucial roles of non-verbal communication, but have also gained deeper knowledge about human cognition and understanding. If the answer to our question is *no*, then that's a fantastic result too! If, despite their differences, the two modes of affective inference exhibit comparable performance levels, then we have uncovered another way of solving the same problem— namely a different path towards emotional inference! Contrasting the two modes of inference can yield as rich an understanding of our own brain functions as it would have had the answer been a *yes*. Hence, in light of the previous discussion, the pursuit of a systematic understanding of non-verbal communication is very likely to be a successful one.

Objection 5: Granted there is a need for assistive technology for autistic individuals, why should we even want Google glasses equipped with the social signal processing technologies for neurotypical-neurotypical interactions (NNIs)? Why would I want to *expose* myself?

There are many instances of NNI when the usage of such Google glasses can provide a deeper understanding of communication. For instance, there are forms of theater and dance that require the inference of the intricate non-verbal expressions to comprehend the story being told. Examples of such art forms include South Asian dance forms of *Kathakali* and *Manipuri*. Viewers, especially those from a different cultural background, have difficulty enjoying the art form and often need an interpretive commentary from a friend to enjoy its richness. On one hand, while the commentary can be helpful to the uninitiated viewer, on the other hand, it can also be disruptive to the overall flow of the experience. Here, if we have Google Glasses whose firmware is quipped to infer the non-verbal aspects of the dance form, then it will be able to provide a far smoother and richer experience for the uninitiated viewer.

More importantly, there are many instances in our daily lives when an important conversation falls apart due to lack of understanding of non-verbal communication. For instance, think of a business negotiation between two business partners Sarah and Rachel. Sarah is oblivious to Rachel's facial expressions and body language of frustration, and as a result, the negotiation is going downhill. During a break, Sarah's friend Hasan, who picked up on the non-verbal cues, informs Sarah about Rachel's frustrated state of mind. Using this new piece of information, Sarah modulates her behavior in a way that saves the deal. Often, highlighting previously undetected non-verbal expressions can enhance one's ability to engage in meaningful interactions with others. Thus, it is reasonable to think that properly functioning Google Glasses can serve the role of Hasan in the aforementioned example, namely by highlighting an inference of a previously undetected non-verbal expression. In fact, if there is sufficient fluidity in human-computer interaction between the agent and the Google Glasses, then we can even expect conversations to be smoother than before. However, it is unlikely that the these Google Glasses will *radically* change the way we interact, for neurotypicals already modulate their behavior depending on the feedback they receive during a conversation. Thus, under these circumstances, a cognitive enhancement in terms of increased social signal processing capabilities can empower the neurotypicals with a deeper understanding human-human interaction. Therefore, instead of looking at the scenario as *exposing* oneself, I think we should look at it as *expressing* oneself in a manner that is more holistic than what *just* words can provide.

5: What the future holds

With a few objections addressed, it is time to focus on the future implications of the work presented both from a philosophical and societal point of view. From an epistemic point of view, traditional analytic philosophy of language has to abandon its tunnel-visioned syntax-sematic

viewpoint of language and communication. Instead of viewing non-verbal communication as just the icing on the cake of verbal communication, philosophical accounts of human-human interaction should acknowledge that non-verbal communication lies at the crux of language and thought along with the verbal aspect. Thus, the incorporation of a holistic mode of communication with equal importance placed on verbal and non-verbal aspects may benefit the way we account for human intelligence and thought in terms of communication. From an educational point of view, incorporation of a systematic understanding of non-verbal communication may enable us to design modules that train students on non-verbal communication along with verbal communication. Individuals can finally tap into the rich modes of non-verbal communication and obtain a systematic understanding of it, which will lead to deeper and more enriching communicative experiences. From a purely technological point of view, we can envision a future where our computers respond emotionally to our actions—imagine a GPS responding to your emotional states, like your frustration (the Emotional Computer— https://www.youtube.com/watch? v=DWu38dgk4s0).

From a sociological point of view, the development of assistive technologies that improve the social signal processing skills of autistic individuals will most likely serve two main purposes. Firstly, it will reduce the otherization that autistic individuals experience, greatly augmenting their quality of life. If the fluidity in their social interactions is comparable to neurotypicals then there is a lower probability of calling them "special" and/or "slow" and excluding them from mainstream society. Moreover, a more effective neurotypical-autistic interaction will likely generate empathy from both sides, leading to a more understanding environment. Most importantly, the more autistics and neurotypicals can communicate effectively, the higher the probability of increased levels of collective learning. For instance, (high functioning) autistic individuals often have gifted abilities for pattern recognition, which is integral in tackling mathematical or algorithmic problems. Neurotypicals, on the other hand, are equipped with their own set of unique skills. An increased

level of effective communication between the two groups is likely to yield a higher epistemic productivity of the group as a whole compared to their separated states. Thus, we can increase the collective learning potential of a population by facilitating an effective autistic-neurotypical interaction.

Concluding remarks

In this paper, I began with the claim that there is more to *just* the words we use in our communication and that non-verbal communication plays an instrumental role in human communication. Moreover, I have argued that the systematic understanding of non-verbal modes of communication empowers us with a deeper understanding of language and thought. The empirical research presented exhibits success in the systematization of non-verbal communication in facial and vocal affects. Moreover, research, especially on Virtual Humans and ECA, indicates that the proper integration of verbal and non-verbal modes of communication constitutes meaningful human-human interaction. Applications of the research can not only be used to augment the social signal processing skills of individuals with ASC, but it can also be used to improve learning technologies and interaction with neurotypicals. Most importantly, research in Affective Computing and HCI is pointing Philosophy of Language and Mind to expand its epistemological boundaries, especially when it comes to talking about human language and thought. Instead of putting on blinders that only enable us to see language as syntax and semantics, the research is ushering a change in the epistemological tide. The blinders need to go, for language and communication is embodied. Non-verbal modes of communication are instrumental in a deeper understanding of human thought. A systematic understanding of non-verbal communication can fundamentally change the way we look at things. Thus, it is in our best interests to pay equal attention to both verbal and non-verbal modes

of communication if we are to succeed in our endeavors to achieve a deeper understanding of human communication.

Coming back to the questions that started the essay: as the bandwidth of communication increases from email to in-person, the "extra thing" that makes the conversation more enjoyable is precisely the scaffolding layers of non-verbal communication. The addition of each layer—prosody with the phone and facial and bodily affects with the face-to-face interaction— enriches the communication by supplying rich, contextual information that cannot be expressed simply by using words. Justine Cassell shares my viewpoint as she claims that "in e-mail, we are obliged to compress all of our communication goals into textual form (plus the occasional emoticon). In face-to-face conversation,… humans have many more modalities of expression at their disposal" (2007, p.10).

Imagine if you could open this file and experience a holographic presentation of the paper. Even with the use of the exact words in this paper, the embodied state of the communication will be enhanced with rich multimodal non-verbal modes of communication, greatly enhancing the comprehension of the communication.

Maybe, that is why I am required to present and defend my thesis in-person, facilitating a rich and multimodal human-human interaction with the audience…

Bibliography

Baron-Cohen, S., Hill, J., Golan, O., & Wheelwright, S. (2002). Mindreading Made Easy. *Cambridge Medicine* , *17*, 28-29.

Baron-Cohen, S. (1997). *Mindblindness: an essay on autism and theory of mind.* (Reprint. ed.). Cambridge, Mass.: MIT Press.

Baron-Cohen, S. (2007). *Mind reading the interactive guide to emotions*. London: Jessica Kingsley.

Cassell, J., & Thorisson, K. (1999). The Power of a Nod and a Glance: Envelope vs. Emotional Feedback in Animated Conversational Agents. *Applied Artificial Intelligence*, *13*(4), 519-538.

Cassell, J. (2007). Body Language: Lessons from the Near-Human. *Genesis redux: essays in the history and philosophy of artificial life* (pp. 1-25). Chicago: University of Chicago Press.

Churchland, P. M. (2002). Outer space and inner space: The new epistemology. *Proceedings and Addresses of the American Philosophical Association*, *76*(2), 25-48.

Churchland, P. M. (1997). To Transform the Phenomena: Feyerabend, Proliferation, and Recurrent Neural Networks. *Philosophy of Science*, *64*, S408-S420.

Clark, A. (1998). Magic Words: How Language Augments Human Computation. *Language and thought: interdisciplinary themes* (pp. 162-183). Cambridge, UK: Cambridge University Press.

Clark, H. H. (1992). *Arenas of language use*. Chicago: University of Chicago Press ;.

Damasio, A. R. (1994). *Descartes' error: emotion, reason, and the human brain*. New York: Putnam.

Definition of rapport in Oxford Dictionaries (British & World English). (n.d.). *Oxford Dictionaries Online*. Retrieved March 29, 2013, from http://oxforddictionaries.com/definition/english/rapport

Dennett, D. C. (1991). *Consciousness explained*. Boston: Little, Brown and Co..

Descartes, R., & Cottingham, J. (1985). *The philosophical writings of Descartes*. Cambridge: Cambridge University Press.

Dretske, F. I. (1981). *Knowledge & the flow of information*. Cambridge, Mass.: MIT Press.

Ekman, P., & Friesen, W. V. (1976). *Pictures of facial affect*. Palo Alto, Calif.: Consulting Psychologists Press.

Fodor, J. A. (1975). *The language of thought*. New York: Crowell.

Ghahramani, Z. (1998). Learning Dynamic Bayesian Networks. *Lecture Notes in Computer Science*, *1387*, 168-197.

Gratch, J. (n.d.). Rapport. *Emotion and Virtual Human Research*. Retrieved December 27, 2012, from people.ict.usc.edu/~gratch/

Greene, J. D. (2009). Dual-process morality and the personal/impersonal distinction: A reply to McGuire, Langdon, Coltheart, and Mackenzie. *Journal of Experimental Social Psychology*, *45*(3), 581-584.

Kaliouby, R., & Robinson, P. (n.d.). Computer Laboratory: Mind-reading machines. *The Computer Laboratory*. Retrieved December 26, 2012, from http://www.cl.cam.ac.uk/research/rainbow/emotions/mrm.html

Kaliouby, R., & Robinson, P. (2004). Mind Reading Machines: Automated Inference of Cognitive Mental States from Video. *Proceedings of The IEEE International Conference on Systems, Man and Cybernetics*, 1-7.

Kaliouby, R., & Robinson, P. (2005). The Emotional Hearing Aid: An Assistive tool for Children with Aspergerâ€™s Syndrome. *Cambridge Online Archives*, Retrieved December 26, 2012, from http://www.cl.cam.ac.uk/~pr10/publications/uais05.pdf

Kaliouby, R., & Robinson, P. (2005). Generalization of a Vision-Based Computational Model of Mind-Reading. *Affective Computing and Intelligent Interaction*, 582-589.

Lewis, D. K. (1969). *Convention: a philosophical study,*. Cambridge: Harvard University Press.

Marsella, S., Gratch, J., & Petta, P. (2010). Computational Models of Emotion. *Blueprint for affective computing: a sourcebook*. Oxford: Oxford University Press.

NUMACK. (n.d.). *ArticuLab :: ArticuLab Home*. Retrieved December 27, 2012, from http://www.articulab.justinecassell.com/projects/numack/index.html

Picard, R. (1995). Affective Computing. *M.I.T Media Laboratory Perceptual Computing Section Technical Report*, *321*, 1-16.

Shikler, T., Robinson, P., & Kaliouby, R. (2004). Design Challenges in Multi Modal Inference Systems for Human Computer Interaction. *Proceedings of International Workshop on Universal Access and Assistive Technology*, Retrieved March 8, 2013, from http://www-edc.eng.cam.ac.uk/cwuaat/04/15-pat-cmc-cwuaatrobinson.pdf

Skyrms, B. (2010). *Signals evolution, learning, & information*. Oxford: Oxford University Press.

Sobol-Shikler, T., Kaliouby, R., & Robinson, P. (2008). Design Challenges in Multi Modal Inference Systems for Human Computer Interaction. *Cambridge Online Archives*, , 55-58.

Sobol-Shikler, T., & Robinson, P. (2010). Classification of Complex Information: Inference of Co-Occurring Affective States from Their Expressions in Speech. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(7), 1284-1297.

Striegnitz, K., Tepper, P., Lovett, A., & Cassell, J. (2009). Knowledge representation for generating locating gestures in route directions. *Spatial language and dialogue*. Oxford: Oxford University Press.

Wang, N., & Gratch, J. (2009). Can Virtual Human Build Rapport and Promote Learning?. *14th International Conference on Artificial Intelligence in Education*.

APA formatting by BibMe.org.