Project Title:	Dyadic Rapport within and across Cultures: Multimodal Assessment in Human-Human and Human-Computer Interaction
Proposal #:	BCS-0729515
HSD Emphasis Area:	Dynamics of Human Behavior
Lead PI:	Prof. Gina-Anne Levow, University of Chicago
Lead PI:	Prof. Jonathan Gratch, University of Southern California
Co-PIs:	Susan Duncan, PhD, University of Chicago
International partners (if any):	Prof. Rima Aboudan, United Arab Emirates University, Al Ain, UAE

Project Goals:

- Phase 1 - Complete a cross-language/cultural comparison of dyadic story-telling interactions in three populations: American English, Gulf/Iraqi Arabic, and Mexican Spanish speakers. Analyses seek to discern verbal/vocal and nonverbal behaviors that are culture-specific and that function as signals enabling members of a given culture to establish and maintain a sense of rapport with one another. This work will take place in Chicago (English and Spanish corpora) and in the United Arab Emirates (Gulf/Iraqi Arabic corpus).
- Phase 2 - Program behavioral repertoires for 'virtual humans', or, 'Embodied Conversational Agents' (ECAs). ECAs are computer-generated, two-dimensional figures, human in appearance and capable of a range of interactive behaviors. This work will focus on interactive behavior patterns that our Phase 1 analyses show to be typical of *listeners* in our three language/cultural groups; for example, dimensions of posture, gaze, nodding, facial expression, and vocal feedback or 'back-channel' utterances. ECAs capable of real-time capture and analysis of certain dimensions video and audio data from a human partner in interaction will model the distinctive, rapport-inducing, behavioral repertoires of listeners in our three target cultures. This work will be undertaken at the University of Southern California, Institute for Creative Technologies.
- Phase 3 - Videotape human participants in interaction with the Phase 2 ECAs. We will be able to manipulate, in the ECAs, aspects of behavior identified in Phase 1 as related to the maintenance of rapport, so as to observe effects on our human participants. For example, if nonverbal cues necessary for maintenance of rapport among American English speakers are infrequent or absent in an ECA that models a member of Gulf Arab or Mexican culture, what effect may this have on the American's story-telling, interactional patterns, and his/her evaluation of the 'quality of interaction' achieved with the ECA?

Thematic Areas:

- Cross-cultural communication - Identification of aspects of behavior that are crucial for scaffolding successful intercultural interaction and communication among people of different cultures.
- Training - Further development of perceptive animated agents with potential roles in, for example, computer-aided language learning applications and training

procedures for U.S. personnel in foreign contexts involving negotiation or conflict.

- Basic science - Development of programmed interfaces (ECAs) for use in controlled scientific elicitations of human responses to subtle behavioral signals in interaction partners, which may be difficult to reliably instantiate, even with training, in human researcher confederates.

Methodologies:

- We are currently halfway through Year 1/Phase 1 of our three-phase project. Here we focus on methods employed and/or under development now that support Phase 1 work.

  Dyads with pre-existing rapport (e.g., friends, family members) participate in a story-telling interaction. Wallace Chafe's short 'Pear film' (<http://www.linguistics.ucsb.edu/faculty/chafe/pearfilm.htm>) serves as an eliciting stimulus. Figure 1 shows how we videotape with three camcorders as the participant who has seen the film (Speaker) tells the story of it to the participant who has not seen it (Listener). Listeners are actively engaged, knowing that they must later re-tell the story to an investigator.

Figure 1. Top-to-bottom: American English-, Iraqi Arabic-, and Mexican Spanish-speaking dyads engaged in the Pear Film elicitation. Listener close-ups are the leftmost stills.

The closeup camcorder views enable analysis of facial expression (<http://www.face-and-emotion.com/dataface/facs/description.jsp>). The view capturing both members of the dyad enables analyses of gesticulation and posture. Audio is collected from head-worn, fixed-distance microphones to enable precise acoustic analyses of vocal behavior. The elicitation data are digitized from videotape into audio and video files suitable for analysis using several software tools. For all the elicitation data, first the speech is transcribed by native speakers of the language of elicitation. For the English and Spanish speech data, initial transcription is accomplished with Praat (<http://www.fon.hum.uva.nl/praat/>) a tool for transcription, annotation, and analysis of speech data. For the Arabic language data, initial

transcription is created as fully vowellized Arabic orthography in text documents. These Arabic transcripts are then transliterated automatically into left-to-right Roman-character text, using a transcoder currently under development as part of this project at MITRE, Corp. The transliterated Arabic texts are then ported into Praat.

Our later analyses of multimodal language and interaction behaviors depend on fine-grained time alignment of the words in the speech signal to the accompanying video achieved at this stage of our processing sequence. To reduce the burden on our transcriber-analysts, we have developed a semi-automatic process to support alignment of speech from our subjects in English, Spanish, and Arabic. The transcriber-analysts first create an initial coarse-grained transcription aligned at the level of the sentence or breath group. We then employ a version of the Sonic speech recognizer developed at the University of Colorado (Pellom et al., 2000. <http://cslr.colorado.edu/beginweb/speech_recognition/sonic.html>) to perform fine-grained word and phoneme alignment consistent with the coarse-grained transcription. Since Sonic has only an English pronunciation lexicon, letter-to-sound rules, and phone model set, we need to perform language porting to support alignment in Spanish and Arabic. In both cases, we create a mapping from a phone set suitable for the new language to the ARPABET representation employed by Sonic as well as a pronunciation lexicon covering the transcribed vocabulary using this new phone set. For the Arabic, additional steps currently under development at the University of Chicago, are required to convert from the Roman-character transliteration format to Sonic-based alignment. The above processing sequence, for all three languages, can reduce the time to achieve word-level alignment in the transcript for an elicitation, from several days to a few hours.

For annotating our multimodal (speech and nonverbal) language and interaction behavioral data we make use of the annotation and language archiving software package, ELAN (<http://www.lat-mpi.eu/tools/elan/>), developed by the Max-Planck Institute for Psycholinguistic Research in Nijmegen, The Netherlands. This 'interactive music score annotation interface' permits the human expert observer-analyst of nonverbal behaviors to isolate individual dimensions of behavior, such as posture, gaze, and gesticulation, on tiers that allow tagging and labeling of behavioral events according to a time line unifying all annotated behavioral streams—of the individual and at the level of the dyad—and permits assessment of their co-occurrence and sequential patterning through time.

The resulting combination of digitized audio-video data, time-aligned transcripts, and annotations reflecting human analysts' close observations of nonverbal behavior is then amenable to a variety of analyses of lexical and prosodic cues in connection with gesticulation and variations in posture, gaze, and other behaviors. To support automatic analysis and synthesis of behaviors associated with rapport in dyadic interactions, we will draw on analyses of regularities observed in the human-annotated behavior streams with the aim of developing and training automatic classifiers that can both identify the level of rapport and predict the timing and type of multi-modal signals of rapport.

Recent Research Findings:

A corpus of 24 English-speaking dyads, 13 Spanish-speaking dyads, and 2 Iraqi Arabic-speaking dyads is so far assembled. While full multi-modal annotation on several dozen ELAN tiers representing distinct behavioral streams is underway, we have begun with analyses of the speech data by focusing on lexical and prosodic cues which are supported by the semi-automatic transcriptions and alignment described above. We currently focus on developing a set of classifiers in an analysis-by-classification approach to explore which prosodic or lexical features best predict backchannels in dyadic communication. We employ a rich set of prosodic features including z-score channel-normalized pitch and intensity maximum, mean, and range across a speech span, duration, speaking rate, and stylized pitch contour. All measures are extracted using Praat for signal processing driven by Python scripts, except for speaking rate, computed using mrate (Morgan et al., 1997). Classification is performed with the Weka machine learning toolkit. Preliminary experiments found reduced intensity to be the best predictor of backchannel response in some English-speaking dyads.

Challenges and Opportunities:

To support the larger-scale signal processing required for audio and video analysis of our rapidly growing dataset as well as to support distributed research access to these materials, we have begun integration of these materials in the Social Informatics Data Grid ('SIDGrid') repository at the University of Chicago. (<https://sidgrid.ci.uchicago.edu/index.php?q=home>). We have uploaded exemplar dyadic interactions annotated in ELAN into a group-access restricted area of the SIDGrid repository. The SIDGrid architecture currently supports distributed parallel execution of several Praat-based prosodic extraction routines on multiple such interactions by dispatching these jobs to the TeraGrid, the largest distributed open science infrastructure. We plan to incorporate our new prosodic feature extraction routines into this framework to facilitate experimentation, as well as to support new multimedia analysis tools as they are created.