A sequence of tools for multimodal analysis of verbal and nonverbal behaviors related to interactional rapport

Dan Loehr
The MITRE Corporation
7515 Colshire Drive
McLean VA 22102
703-983-6765
loehr@mitre.org

Sue Duncan
University of Chicago
5484 South University Avenue
Chicago IL 60637
773-702-8832
deng@uchicago.edu

Gina-Anne Levow
University of Manchester
131 Princess Street
Manchester UK M1 7DN
44-161-306-3094
gina-anne.levow@manchester.ac.uk

ABSTRACT
We describe a method for annotation and analysis of multimodal video data using a sequence of specialized tools that we have made interoperable. We present initial, test-sample results derived with this method. This report details the building blocks of cross-language, multi-modal analyses planned for a large corpus of audio-videotaped, dyadic, conversation data comprising elicitations from Gulf-region Arabic speakers, Mexican Spanish speakers, and American English speakers. We discuss how our approach meets the challenge of readying audio, video, and transcription text data from these three diverse languages for annotation and comparative analysis of multimodal language behaviors related to maintenance of interactional rapport.

Categories and Subject Descriptors
H.5.1 [**HCI**]: Multimedia Information Systems – *video, methodology, virtual reality.*

General Terms
Human Factors

Keywords
Multimodal annotation, multimodal analysis, interactional rapport.

1. INTRODUCTION
Within the last decade, dozens of tools have been created that support annotation and analysis of multimodal behavior captured on digitized audio and video. These tools are typically specialized for a single purpose. For instance, one tool may permit accurate speech phoneme annotation, another enable semi-automated extraction of video events matching certain parameters, still another may support morpho-syntactic labeling of sign language discourse, and another may allow rich annotation of discursive texts, but at the expense of fine-grained video frame observation. Each of these tools has its strengths, however all are largely un-interoperable, so that the researcher is unable to combine these strengths. Further, though most tools claim to support both annotation and analysis, in reality the primary focus of most is in annotation, with analysis limited to counting annotated tokens and answering questions based on these counts. Few tools provide more sophisticated analysis of temporal patterning in multimodal behaviors, which is arguably the desired goal for multimodal researchers.

Recently, a group of developers of diverse tools for annotation and analysis of multimodal language data have been working to promote interoperability among the tools (Schmidt, et al. 2008). This allows a processing sequence in the spirit of UNIX utilities; that is, an approach focused on solving larger problems with a sequence of smaller tools, each designed for a specific purpose. This approach also permits metadata accumulated with annotation tools to be input into temporal pattern-analysis tools. To date, the interoperability achieved by these developers has been proof-of-concept, limited to a small, relatively constrained data set. In this paper, we describe using a collection of tools for a real-world analysis of verbal and nonverbal behaviors related to maintenance of rapport in dyadic conversations in three language/cultural groups: Gulf (Iraqi and Emirati) Arabic, Mexican Spanish, and American English. We discuss issues found along the way, and present preliminary results on elicitation data from each group derived from this methodology, focusing on aspects of listeners' reactions to behaviors that speakers engage in during a story-telling activity.

2. BACKGROUND
There are a variety of tools for annotating and analyzing multimodal communication in digitized audio and video (see Bigbee et al. 2001, Knudsen et al. 2002, Rohlfing et al. 2006 *inter alia* for overviews). A transcription and speech-acoustics analysis tool popular in the linguistics research community is Praat (www.praat.org). For video, a number of tools exist that provide a video window with playback controls embedded in a "music score" interactive annotation interface, in which horizontal tracks or "tiers" (one to represent each behavior stream of interest) scroll right or left as the video is played forwards or backwards. Examples of music score annotation interfaces include Anvil (www.anvil-software.de), EXMARaLDA (www.exmaralda.org), and ELAN (www.lat-mpi.eu/tools/elan). Figure 1 depicts ELAN's user interface with an interval of annotations of three time-aligned

videos (different camcorder angles on one elicitation) from our corpus of dyadic discourse data (described below).

Figure 1. ELAN's user interface

The pattern-analysis software Theme (Magnusson et al. 2004), applied to such annotation data—labeled intervals—detects significant patterns in sequences of behaviors in time. Figure 2 shows Theme's interface, displaying recurring patterned sequences. In this example, the pattern consists of four events: (1) listener begins blink, (2) listener begins nod, (3) listener ends blink, (4) listener ends nod. In other words, the listener's blink and nod co-occur. This pattern repeats 15 times within 280 seconds.

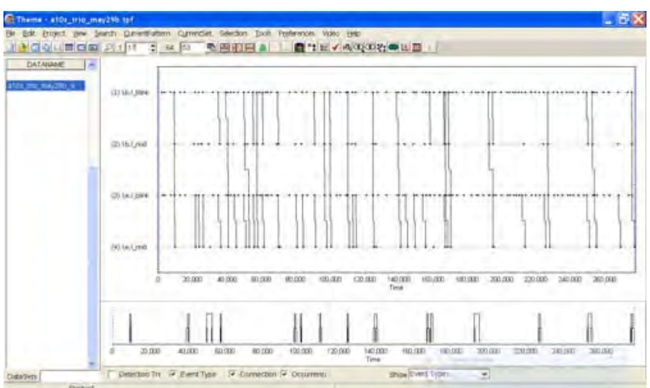

Figure 2. Theme's display of temporal patterns.

Figure 2's upper pane traces the linear chains of events making up the repeated pattern along a timeline (up to four events per chain in this example). The bottom pane shows the patterns (with internal sub-pattern organization) occurring along a timeline.

As mentioned, the many existing tools for the above and other purposes have been largely uninteroperable, until a recent initiative by Schmidt et al. to promote data sharing. As a starting point, Schmidt et al. took the Annotation Graph (AG) framework (Bird and Liberman 2001), which is designed for linguistic annotation along a timeline. Schmidt et al. extended AG to include a core set of annotations common to most music-score video annotation tools, including the ability to assign annotation labels to different tiers. Developers of a variety of tools (including all the tools mentioned above) then wrote converters to and from this common AG-based exchange format. As a proof of concept, an enduser of each tool then annotated verbal and nonverbal behaviors in one minute of a common video clip, which was a recording of a speaker telling a story. Annotations accumulated using each annotation tool were successfully exchanged with and displayed in all other tools.

3. CROSS-CULTURE COMPARISON OF INTERACTIONAL RAPPORT

The goal of our current research, and of our analysis 'pipeline', is to elucidate behavioral similarities and differences among three language/cultural groups—American English, Gulf Arabic, and Mexican Spanish—focusing specifically on how individuals within each culture establish and maintain rapport. A near-term goal of this research is to annotate, analyze, and describe multiple verbal (speech and speech-prosodic) and nonverbal (gesture, facial expression, posture, etc.) markers of rapport to feed modeling of behavioral repertoires for "virtual human" interlocutors. In human-computer interaction settings, these characters, functioning as "listening agents" interact with humans, who then evaluate the level of rapport using subject questionnaires, along the lines of Gratch 2007. The cross-culture comparative dimension of our study potentiates the development of agents that model culturally distinctive behaviors that are related to maintenance of interactional rapport. Ultimately this will make possible HCI elicitations in which the reactions of human participants to agents modeling culturally familiar *versus* culturally unfamiliar interactive behaviors may be compared. Briefly summarized, our research strategy is to *videotape*, *annotate*, *elucidate*, *generate*, and *evaluate* rapport-ful behavior in three cultures. That is, *videotape* human-human dyads (Arabs, Mexicans, and Americans), *annotate* the videos, analyze and *elucidate* rapport cross-culturally, *generate* culture-appropriate rapport behaviors in virtual humans, and *evaluate* the effect of these agent behaviors on human participants in HCI elicitations. This paper focuses on issues with the 2^{nd} and 3^{rd} steps, i.e. annotating videos for the purpose of elucidating markers of rapport. To this end, we describe our methodology.

4. METHODOLOGY

We audio-videotaped dyads from each language/culture, engaged in an unrehearsed story-telling activity. A participant in the role of *Speaker* who had seen the "Pear Film" (Chafe 1975) told the story of the film to a naïve *Listener*. We instructed listeners to be "active and engaged" in the story-telling task. They understood that, after hearing the story, they would be videotaped themselves, re-telling it to an investigator. Figure 3 presents sample stills from our data, showing a close-up on each participant (giving the resolution necessary for observation of facial expression) and a wider view (for observation of manual gesture and larger body movements).

After videotaping, our processing sequence consisted of the following steps and tools. We list them exhaustively to illustrate the complex reality (particularly, as concerns the challenges of dealing with the Arabic language in interfaces designed only for

Figure 3. 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.

left-to-right running languages) of what might be assumed to be a straightforward process.

1. Synchronize all three camcorder views into one composite video, as shown in Figure 1, using Apple Computer's Final-CutPro media editing software. Editing the three views together is to ensure that the videos remain synchronized, when played in the other interfaces involved in this processing sequence.
2. Generate a sound-only file from the "trio-ed" video file. Import this sound file into Praat.
3. For Spanish and English, transcribe the speech into a Praat tier, applying the RT-03 transcription guidelines (NIST 2008) for annotating, for instance, filled pauses, speech interruptions, non-speech sounds, and so on.
   a. For Arabic, first transcribe into fully-vowelled Arabic orthographic script using Basis Technology's Arabic Editor (basistech.com/arabic-desktop-suite), in two versions: (1) Modern Standard Arabic ("dialect-neutral") and (2) colloquial, dialect-sensitive Arabic. Then, using a custom-built transcoder/transliteration tool, transliterate each version into Latin characters according to two schemes: ARPABET, an ASCII-based phonetic alphabet (for step 4, below), and a standard Arabic-English dictionary format (Wehr 1993), for human readability. Finally, import the tiers of transliterated Arabic into Praat.
4. Input the sound file and the Praat speech transcriptions into the University of Colorado Sonic speech recognition system (Pellom, 2001) in forced-alignment mode. This automatically adds word and phoneme boundaries to the Praat transcriptions to match the actual boundaries in the speech stream. For Arabic and Spanish, Sonic's language porting capabilities, in conjunction with custom dictionaries, enable alignment. Re-import these time-aligned transcriptions into Praat.
5. For Spanish and Arabic, add Praat tiers with interlinear English glosses (phrase- and word-level), as an assist to English-speaking researchers who will annotate and interpret the verbal and nonverbal behavioral data.
6. Before continuing, visually inspect and correct the results of all previous processing steps—a process of manually adjusting approximately 5-10% of the onset/offset boundary markers on the phrase- and word-level Praat tiers, as well as deleting inserted spurious intervals of silence.
7. Open the composite video in an ELAN annotation file and import the Praat tiers with annotations, with one ELAN tier for each Praat tier imported.
8. Annotate verbal and nonverbal phenomena in ELAN on additional tiers as needed. The resulting tiers (including those from Praat) are all time-aligned with each other and with the video.
9. Import the ELAN tiers into Theme, as temporal "events", where each event consists of a timestamp, an annotation label, and an "actor" (e.g. speaker or listener).
10. Analyze with Theme to discover patterned temporal sequences of verbal and nonverbal behaviors in interaction.

To our knowledge, we are the first to undertake such a sophisticated verbal and nonverbal time-aligned annotation of these languages/cultures. One of our contributions, therefore, is the description of the above methodology for doing so. However, the goal is carry out the analysis afforded by this methodology, to unearth verbal and nonverbal indicators of rapport. We now turn to these findings.

5. INITIAL FINDINGS

At this time, our corpus of dyadic discourse data comprises elicitations from 45 Arab dyads (Iraqi and Emirati) videotaped in Amman, Jordan and Al-Ain, the United Arab Emirates, 20 Mexican dyads videotaped in Chicago, Illinois, and 30 American dyads, also videotaped in Chicago. All dyads were close friends or family members. In other words, had already-established rapport. For purposes of demonstrating use of the methods described in this report we selected one dyadic elicitation from each group for which we have completed all steps in our processing sequence at this time. The listener in each dyad is a male in his 20s or 30s. For our initial run of Theme analyses on these elicitations we focused on patterns involving a subset of annotated listener behaviors. First, the Arab and the American listener both manifested a frequently recurring pattern of association of back-channel behaviors. This was a pattern of pairing head nods with eye blinks (as shown in Figure 2). Theme discovered no such pattern of co-occurrence in the listener in the Mexican dyad. A recurring pattern discovered in the American dyad, but not in the Arab or Mexican dyads, was a multi-behavior sequence in which the listener blinked and nodded, then engaged in fidgety behaviors of some kind, after which the speaker participant suffered two identical intervals of speech hesitation and pausing (see Figure 4). This interaction between speaker and listener occurred three times within 80 seconds. Finally, a pattern discovered only in the Mexican dyad was one in which the speaker shifted gaze to the listener, then away, and then engaged in an interval of speech hesitation.

6. CONCLUSION AND FUTURE WORK

The initial patterns discovered by Theme analysis of our dyadic elicitation data provide just a suggestion of the sorts of patterns of behavior that the methods described here will enable us to discover. The goal will be to discern patterns of interaction in rapportful interactions that distinguish the three language/cultural groups that are the focus of our comparative study. Distinguishing such behavioral profiles will underpin future efforts at modeling culture-specific interaction behaviors in "virtual humans" and this work will lead to HCI research in which the effects on human participants of exposure to listening agents manifesting different cultural profiles may be studied.

7. ACKNOWLEDGMENTS

This matrial is based upon work supported by a grant from the National Science Foundation under grant #BCS-0729515.

8. REFERENCES

[1] Bigbee, A., Loehr, D., and Harper, L. 2001 Emerging Requirements for Multi-Modal Annotation and Analysis Tools, Proceedings, Eurospeech 2001.

[2] Bird, S. & Liberman, M. 2001. A formal framework for linguistic annotation. Speech Communication 33, 23-60.

[3] Chafe, W. 1975. The Pear Film. www.linguistics.ucsb.edu/faculty/chafe/pearfilm.htm

[4] Gratch, J., Wang, N., Gerten, J., Fast, E., and Duffy, R. 2007. Creating Rapport with Virtual Agents. International Conference on Intelligent Virtual Agents.

[5] Knudsen, M., Martin, J-C., Dybkjær, L., Ayuso, M., Bernsen, N., Carletta, J., Heid, Ul, Kita, S., Llisterri, J., Pelachaud, C., Poggi, I., Reithinger, N., van Elswijk, G., and Wittenburg, P. 2002 Survey of Multimodal Annotation Schemes and Best Practice, ISLE Deliverable D9.1, www.nis.sdu.dk/publications/year02.html.

[6] Magnusson, M.S., Burfield, I., Loijens, L., Grieco, F., Jonsson, G.K., and Andrew Spink. 2004. Theme; Powerful tool for detection and analysis of hidden patterns in behavior. Reference Manual. Version 5.0. 229 pages. PatternVision Ltd and Noldus Information Technology

[7] National Institute for Standards and Technology. 2008. Rich Transcription Evaluation Project, www.itl.nist.gov/iad/mig/tests/rt.

[8] Pellom, B. SONIC: The University of Colorado Continuous Speech Recognizer, University of Colorado, Tech Report #TR-CSLR-2001-01, Boulder, Colorado, March, 2001.

[9] Rohlfing, K., Loehr, D., Duncan, S., Brown, A., Franklin, A., Kimbara, I., Milde, J., Parrill, F., Rose, T., Schmidt, T., Sloetjes, H., Thies, A., Wellinghoff, S. 2006. Comparison of multimodal annotation tools: Workshop report. Gesprächsforschung 7.

[10] Schmidt, T., Ehmer, O., Hoyt, J., Kipp, M., Loehr, D., Rose, T., Sloetjes, H., Duncan, S., Magnusson, M. 2008. An exchange format for multimodal annotations. Proceedings, 6th international conference on Language Resources and Evaluation (LREC-08).

[11] Wehr, H. 1993. Arabic-English Dictionary: The Hans Wehr Dictionary of Modern Written Arabic (4th ed.). Spoken Language Services.

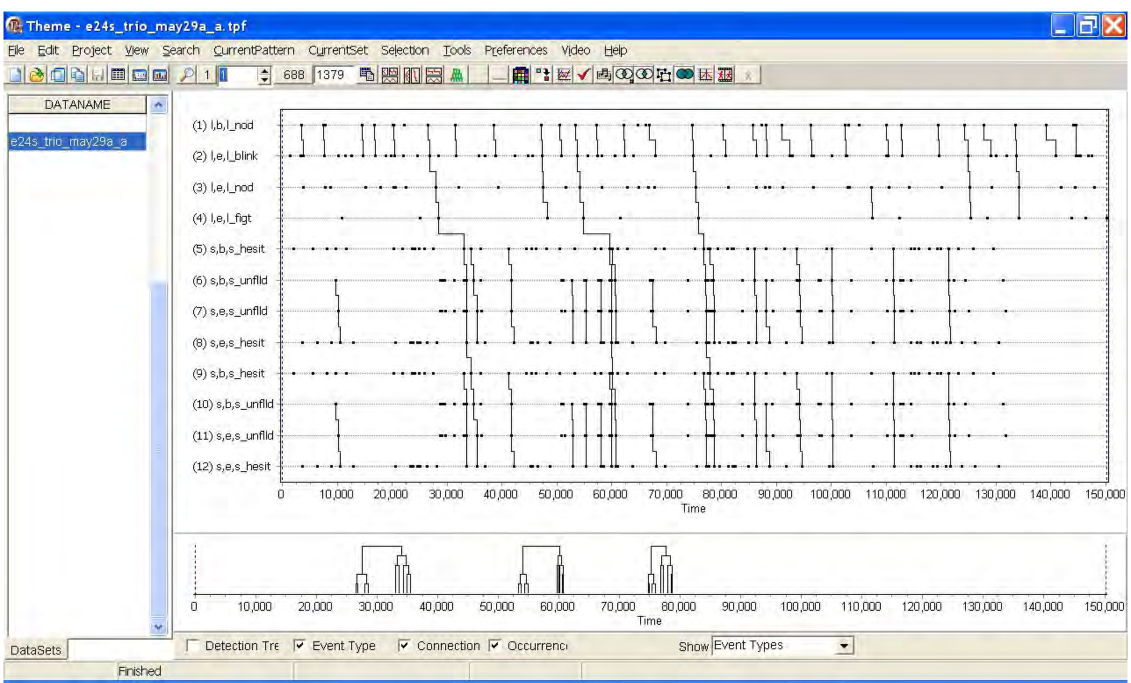

Figure 4. 12-step pattern: Listener blink/nod/fidget sequence followed by two speaker sequences of unfilled-pause/hesitation