Incorporating Gesture and Gaze into Multimodal Models of Human-to-Human Communication

Lei Chen
Dept. of Electrical and Computer Engineering
Purdue University
West Lafayette, IN 47907
chenl@ecn.purdue.edu

Abstract

Structural information in language is important for obtaining a better understanding of a human communication (e.g., sentence segmentation, speaker turns, and topic segmentation). Human communication involves a variety of multimodal behaviors that signal both propositional content and structure, e.g., gesture, gaze, and body posture. These non-verbal signals have tight temporal and semantic links to spoken content. In my thesis, I am working on incorporating non-verbal cues into a multimodal model to better predict the structural events to further improve the understanding of human communication. Some research results are summarized in this document and my future research plan is described.

1 Introduction

In human communication, ideas tend to unfold in a structured way. For example, for an individual speaker, he/she organizes his/her utterances into *sentences*. When a speaker makes errors in the dynamic speech production process, he/she may correct these errors using a *speech repair* scheme. A group of speakers in a meeting organize their utterances by following a *floor control* scheme. All these structures are helpful for building better models of human communication but are not explicit in the spontaneous speech or the corresponding transcription word string. In order to utilize these structures, it is necessary to first detect them, and to do so as efficiently as possible. Utilization of various kinds of knowledge is important; For example, lexical and prosodic knowledge (Liu, 2004; Liu et al., 2005) have been used to detect structural events.

Human communication tends to utilize not only speech but also visual cues such as gesture, gaze, and so on. Some studies (McNeill, 1992; Cassell and Stone, 1999) suggest that gesture and speech stem from a single underlying mental process, and they are related both temporally and semantically. Gestures play an important role in human communication but use quite different expressive mechanisms than spoken language. Gaze has been found to be widely used in coordinating multi-party conversations (Argyle and Cook, 1976; Novick, 2005). Given the close relationship between non-verbal cues and speech and the special expressive capacity of non-verbal cues, we believe that these cues are likely to provide additional important information that can be exploited when modeling structural events. Hence, in my Ph.D thesis, I have been investigating the combination of lexical, prosodic, and non-verbal cues for detection of the following structural events: *sentence units*, *speech repairs*, and *meeting floor control*.

This paper is organized as follows: Section 1 has described the research goals of my thesis. Section 2 summarizes the efforts made related to these goals. Section 3 lays out the research work needed to complete my thesis.

2 Completed Works

Our previous research efforts related to multimodal analysis of human communication can be roughly grouped to three fields: (1) multimodal corpus col-

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 211–214,
New York, June 2006.

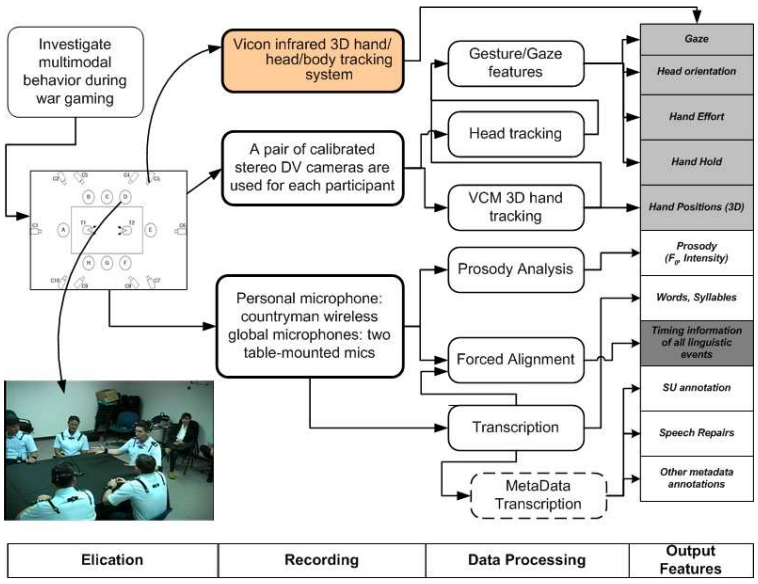

Figure 1: VACE meeting corpus production

lection, annotation, and data processing, (2) measurement studies to enrich knowledge of non-verbal cues to structural events, and (3) model construction using a data-driven approach. Utilizing non-verbal cues in human communication processing is quite new and there is no standard data or off-the-shelf evaluation method. Hence, the first part of my research has focused on corpus building. Through measurement investigations, we then obtain a better understanding of the non-verbal cues associated with structural events in order to model those structural events more effectively.

2.1 Multimodal Corpus Collection

Under NSF KDI award (Quek and et al.,), we collected a multimodal dialogue corpus. The corpus contains calibrated stereo video recordings, time-aligned word transcriptions, prosodic analyses, and hand positions tracked by a video tracking algorithm (Quek et al., 2002). To improve the speed of producing a corpus while maintaining its quality, we have investigated factors impacting the accuracy of the forced alignment of transcriptions to audio files (Chen et al., 2004a).

Meetings, in which several participants communicate with each other, play an important role in our daily life but increase the challenges to current information processing techniques. Understanding human multimodal communicative behavior, and how witting and unwitting visual displays (e.g., gesture, head orientation, gaze) relate to spoken content is critical to the analysis of meetings. These multimodal behaviors may reveal static and dynamic social structure of the meeting participants, the flow of topics being discussed, the control of floor of the meeting, and so on. For this purpose, we have been collecting a multimodal meeting corpus under the sponsorship of ARDA VACE II (Chen et al., 2005). In a room equipped with synchronized multichannel audio,video and motion-tracking recording devices, participants (from 5 to 8 civilian, military, or mixed) engage in planning exercises, such as managing rocket launch emergency, exploring a foreign weapon component, and collaborating to select awardees for fellowships. we have collected and continued to do multichannel time synchronized audio and video recordings. Using a series of audio and video processing techniques, we obtain the word transcriptions and prosodic features, as well as head, torso and hand $3D$ tracking traces from visual trackers and Vicon motion capture device. Figure 1 depicts our meeting corpus collection process.

2.2 Gesture Patterns during Speech Repairs

In the dynamic speech production process, speakers may make errors or totally change the content of what is being expressed. In either of these cases, speakers need refocus or revise what they are saying

and therefore speech repairs appear in overt speech. A typical speech repair contains a *reparandum*, an optional *editing phrase*, and a *correction*. Based on the relationship between the reparandum and the correction, speech repairs can be classified into three types: *repetitions*, *content replacements*, and *false starts*. Since utterance content has been modified in last two repair types, we call them content modification (**CM**) repairs. We carried out a measurement study (Chen et al., 2002) to identify patterns of gestures that co-occur with speech repairs that can be exploited by a multimodal processing system to more effectively process spontaneous speech. We observed that modification gestures (**MGs**), which exhibit a change in gesture state during speech repair, have a high correlation with content modification (**CM**) speech repairs, but rarely occur with content repetitions. This study does not only provide evidence that gesture and speech are tightly linked in production, but also provides evidence that gestures provide an important additional cue for identifying speech repairs and their types.

2.3 Incorporating Gesture in SU Detection

A sentence unit (SU) is defined as the complete expression of a speaker's thought or idea. It can be either a complete sentence or a semantically complete smaller unit. We have conducted an experiment that integrates lexical, prosodic and gestural cues in order to more effectively detect *sentence unit* boundaries in conversational dialog (Chen et al., 2004b).

As can be seen in Figure 2, our multimodal model combines lexical, prosodic, and gestural knowledge sources, with each knowledge source implemented as a separate model. A hidden event language model (LM) was trained to serve as lexical model ($P(W, E)$). Using a direct modeling approach (Shriberg and Stolcke, 2004), prosodic features were extracted using the SRI prosodic feature extraction tool[1] by collaborators at ICSI and then were used to train a CART decision tree as the prosodic model ($P(E|F)$). Similarly to the prosodic model, we computed gesture features directly from visual tracking measurements (Quek et al., 1999; Bryll et al., 2001): $3D$ hand position, Hold (a state when there is no hand motion beyond some adaptive threshold results), and Effort (analogous to the kinetic energy of hand movement). Using gestural features, we trained a CART tree to serve as the gestural model ($P(E|G)$). Finally, an HMM based model combination scheme was used to integrate predictions from individual models to obtain an overall SU prediction ($argmax(E|W, F, G)$). In our investigations, we found that gesture features complement the prosodic and lexical knowledge sources; by using all of the knowledge sources, the model is able to achieve the lowest overall detection error rate.

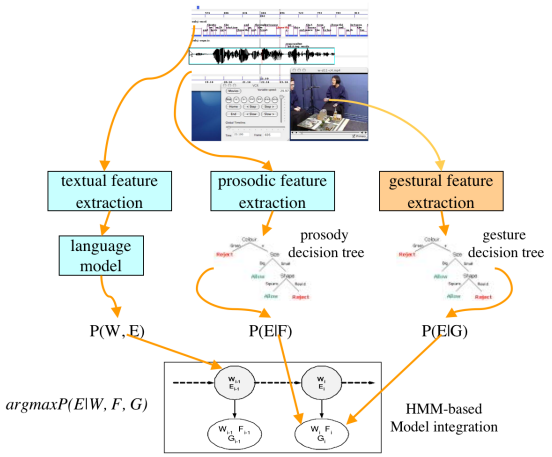

Figure 2: Data flow diagram of multimodal SU model using lexical, prosodic and gestural cues

2.4 Floor Control Investigation on Meetings

An underlying, auto-regulatory mechanism known as "floor control", allows participants communicate with each other coherently and smoothly. A person controlling the floor bears the burden of moving the discourse along. By increasing our understanding of floor control in meetings, there is a potential to impact two active research areas: human-like conversational agent design and automatic meeting analysis. We have recently investigated floor control in multi-party meetings (Chen et al., 2006). In particular, we analyzed patterns of speech (e.g., the use of *discourse markers*) and visual cues (e.g., eye gaze exchange, pointing gesture for next speaker) that are often involved in floor control changes. From this analysis, we identified some multimodal cues that will be helpful for predicting floor control events. Discourse markers are found to occur frequently at the beginning of a floor. During floor transitions, the

[1] A similar prosody feature extraction tool has been developed in our lab (Huang et al., 2006) using Praat.

previous holder often gazes at the next floor holder and vice verse. The well-known mutual gaze break pattern in dyadic conversations is also found in some meetings. A special participant, an active meeting manager, is found to play a role in floor transitions. Gesture cues are also found to play a role, especially with respect to floor capturing gestures.

3 Research Directions

In the next stage of my research, I will focus on integrating previous efforts into a complete multimodal model for structural event detection. In particular, I will improve current gesture feature extraction, and expand the non-verbal features to include both eye gaze and body posture. I will also investigate alternative integration architectures to the HMM shown in Figure 2. In my thesis, I hope to better understand the role that the non-verbal cues play in assisting structural event detection. My research is expected to support adding multimodal perception capabilities to current human communication systems that rely mostly on speech. I am also interested in investigating mutual impacts among the structural events. For example, we will study SUs and their relationship to floor control structure. Given progress in structural event detection in human communication, I also plan to utilize the detected structural events to further enhance meeting understanding. A particularly interesting task is to locate salient portions of a meeting from multimodal cues (Chen, 2005) to summarize it.

References

M. Argyle and M. Cook. 1976. *Gaze and Mutual Gaze*. Cambridge Univ. Press.

R. Bryll, F. Quek, and A. Esposito. 2001. Automatic hand hold detection in natural conversation. In *IEEE Workshop on Cues in Communication*, Kauai,Hawaii, Dec.

J. Cassell and M. Stone. 1999. Living Hand to Mouth: Psychological Theories about Speech and Gesture in Interactive Dialogue Systems. In *AAAI*.

L. Chen, M. Harper, and F. Quek. 2002. Gesture patterns during speech repairs. In *Proc. of Int. Conf. on Multimodal Interface (ICMI)*, Pittsburg, PA, Oct.

L. Chen, Y. Liu, M. Harper, E. Maia, and S. McRoy. 2004a. Evaluating factors impacting the accuracy of forced alignments in a multimodal corpus. In *Proc. of Language Resource and Evaluation Conference*, Lisbon, Portugal, June.

L. Chen, Y. Liu, M. Harper, and E. Shriberg. 2004b. Multimodal model integration for sentence unit detection. In *Proc. of Int. Conf. on Multimodal Interface (ICMI)*, University Park, PA, Oct.

L. Chen, T.R. Rose, F. Parrill, X. Han, J. Tu, Z.Q. Huang, I. Kimbara, H. Welji, M. Harper, F. Quek, D. McNeill, S. Duncan, R. Tuttle, and T. Huang. 2005. VACE multimodal meeting corpus. In *Proceeding of MLMI 2005 Workshop*.

L. Chen, M. Harper, A. Franklin, T. R. Rose, I. Kimbara, Z. Q. Huang, and F. Quek. 2006. A multimodal analysis of floor control in meetings. In *Proc. of MLMI 06*, Washington, DC, USA, May.

L. Chen. 2005. Locating salient portions of meeting using multimodal cues. Research proposal submitted to AMI training program, Dec.

Z. Q. Huang, L. Chen, and M. Harper. 2006. An open source prosodic feature extraction tool. In *Proc. of Language Resource and Evaluation Conference*, May 2006.

Y. Liu, E. Shriberg, A. Stolcke, B. Peskin, J. Ang, Hillard D., M. Ostendorf, M. Tomalin, P. Woodland, and M. Harper. 2005. Structural Metadata Research in the EARS Program. In *Proc. of ICASSP*.

Y. Liu. 2004. *Structural Event Detection for Rich Transcription of Speech*. Ph.D. thesis, Purdue University.

D. McNeill. 1992. *Hand and Mind: What Gestures Reveal about Thought*. Univ. Chicago Press.

D. G. Novick. 2005. Models of gaze in multi-party discourse. In *Proc. of CHI 2005 Workshop on the Virtuality Continuum Revisted*, Portland OR, April 3.

F. Quek and et al. KDI: Cross-model Analysis Signal and Sense- Data and Computational Resources for Gesture, Speech and Gaze Research, http://vislab.cs.vt.edu/kdi.

F. Quek, R. Bryll, and X. F. Ma. 1999. A parallel algorighm for dynamic gesture tracking. In *ICCV Workshop on RATFG-RTS*, Gorfu,Greece.

F. Quek, D. McNeill, R. Bryll, S. Duncan, X. Ma, C. Kirbas, K. E. McCullough, and R. Ansari. 2002. Multimodal human discourse: gesture and speech. *ACM Trans. Comput.-Hum. Interact.*, 9(3):171–193.

E. Shriberg and A. Stolcke. 2004. Direct modeling of prosody: An overview of applications in automatic speech processing. In *International Conference on Speech Prosody*.