

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

Lei Chen

Dept. of Electrical Engineering
Purdue University
West Lafayette, IN 47906
chenl@ecn.purdue.edu

Abstract

In human communication, utterances are expressed with some structural events, e.g., sentence, speech repairs, control of floor, and etc. These structural events bring important information and quite helpful for a better understanding of human communication. Meanwhile, the human communication is also full of multimodal behaviors, e.g., gesture, gaze, and etc. As non-verbal signals, gesture and gaze show closed temporal and semantic links to spoken content. In my thesis, I am working on incorporating non-verbal cues in a multimodal model to better predict the structural events to further improve the understanding of human communication. Some research efforts made are summarized and my future research plan is described.

1 Introduction

In human communication, the ideas are expressed via some structural formats. For example, for an individual speaker, he/she organizes his/her utterances using *topics* and *sentences*; for a group of speakers in a meeting, they organize their utterances by following *floor control* scheme. All these structural events are helpful for a potential better understanding of human communication and machine processing. Focusing on analysis and detection of these structural events, there have some works done using lexical and prosodic knowledges alone (Liu, 2004; Liu et al., 2005).

The human communication is indeed a multimodal process. People do not only use speech but also non-verbal signals, e.g., gesture, gaze, and so on. Some studies (McNeill, 1992; Cassell and Stone, 1999) have shown 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 and use quite different expressive mechanisms than spoken language. Gaze are found 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, non-verbal cues are likely to provide additional important information that can be exploited when modeling structural events. Hence, in my Ph.D thesis, I propose to combine lexical, prosodic, and non-verbal cues for the investigation and detection of following structural events: *sentence units*, *speech repairs*, and *meeting floor control*.

The paper is organized as follows: Section 1 has proposed the research goal of my thesis. Section 2 summaries previous efforts made towards the research goal presented above. Section 3 describes my planned future research work to finish my thesis.

2 Completed Works

The previous research efforts can be roughly grouped to three fields: (1) multimodal corpora collection, annotation and data processing, (2) measurement study to enrich knowledge of non-verbal cues to structural events, and (3) model building us-

ing data-driven approach. Utilizing non-verbal cues in human communication processing is quite new and there is no standard data and ready evaluation method. Hence, the first part of my research is focused on corpus building. Through measurement investigations, we enrich the knowledge about non-verbal cues to interested structural events in order to model those structural events for prediction.

2.1 Multimodal Dialog 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 transcription, prosodic analysis, and hand position tracked by video tracking algorithm (Quek et al., 2002). To improve the speed of producing a corpus while maintain its quality, I have survey factors impacting the accuracy of forced alignment of transcriptions to audio files (Chen et al., 2004a).

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 cases, speakers need refocus or revise what they are saying and therefore speech repairs appear in overt speech. A typical speech repair contains *reparandum*, *editing phrase* and *correction*. Based on relationship between reparandum and the correction, speech repairs can be classified into three types: *repetitions*, *content replacement*, and *false start*. Since utterance content has been modified in last two repair types, we denoted them as content modification (CM) repairs. The presence of speech repairs seriously degrades the accuracy of current speech recognition systems. We carried out a measurement study (Chen et al., 2002) to find patterns of gesture that co-occur with speech repairs that can be exploited by a multimodal processing system to more effectively proceed spontaneous speech. We observed that modification gesture (MG), that exhibits a change in gesture state during speech repair, has 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 provide evidence that gestures provide an impor-

tant additional cue for identifying speech repairs and their types.

2.3 Incorporating Gesture in SU detection

An 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 to fuse lexical, prosodic and gestural cues together to detect *sentence unit* boundaries on conversational dialog (Chen et al., 2004b). As can be seen

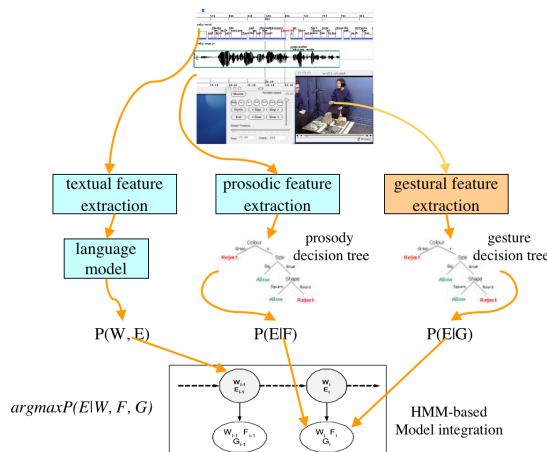


Figure 1: Data flow diagram of multimodal SU model using textual, prosodic and gestural cues

in Figure 1, our multimodal model combines lexical, prosodic and gesture 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 SRI prosodic feature extraction tool ¹by collaborators in ICSI and then were used to train a CART decision tree as prosodic model ($P(E|F)$). Similar to prosodic model, we computed gestural 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

¹A similar toolkit has been developed in our lab (Huang et al., 2006) on Praat.

gestural model ($P(E|G)$). Finally, a HMM based model combination scheme was used to fuse predictions from individual model to obtain overall SU prediction ($\text{argmax}(E|W, F, G)$). In our investigations, we found that gesture features complement the prosodic and lexical knowledge sources so by using all of the knowledge sources, the model is able to achieve the lowest overall detection error rate.

2.4 Multimodal Meeting Corpus Collection

Meeting, in which several participants communicate to each other, plays important role in our daily information exchange and brings more 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 large size of multimodal meeting corpus under the ARDA VACE award (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 collected multichannel time synchronized audio and video recordings. Using a series of audio and video processing techniques, we obtained the word transcription and prosodic features, as well as head, torso and hand 3D tracking traces from visual tracker and Vicon motion captures. Figure 2 depicts our meeting corpus collection process.

2.5 Floor Control Investigation on Meetings

An underlying, auto-regulatory mechanism known as “floor control”, enforces that meeting can be carried out in a coherent and smooth way among several participants. 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 de-

sign and automatic meeting analysis. we investigated floor control in multiparty 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 previous holder often gazes at the next floor holder and vice versa. 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 step research, I will focus on integrating previous efforts in a complete multimodal model for structural events study. In particular, I will improve current gestural feature extraction to utilize more temporal and spatial information of gesture, and expand non-verbal feature sets to other domain, i.e., eye gaze and body posture; I will investigate alternative integration architectures to the HMM shown in Figure 1; I will expand the quantity and quality of multimodal data so that I can use updated feature extraction, model training and combination on speech repairs and floor controls detection. Given the progress of structural events identification in human communication, I also plan to utilize detected structural events to further enhance meeting understanding. A particular task is to locating salient portions of meeting from multimodal cues (Chen, 2005)

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:

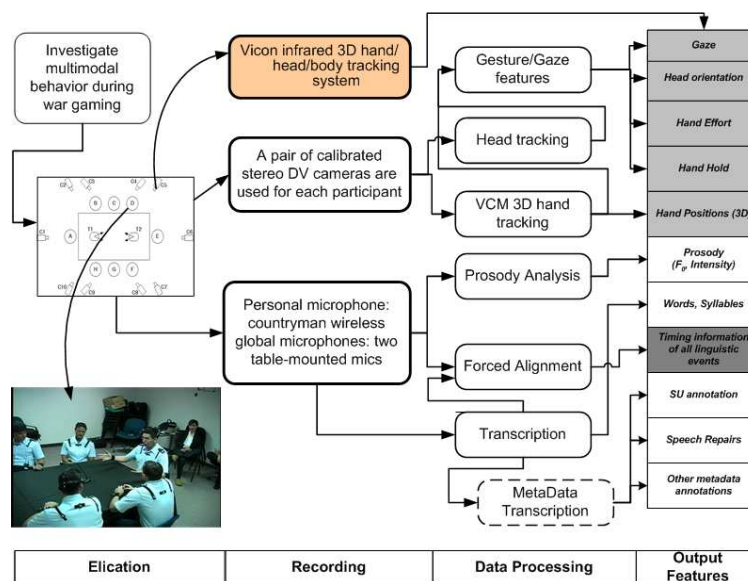


Figure 2: VACE meeting corpus production

- 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. Rose, F. Parrill, X. Han, J. Tu, Z. Huang, I. Kimbara, H. Welji, M. Harper, Q. Francis, 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, R. Travis Rose, I. Kimbara, Z. Q. Huang, and F. Quek. 2006. A multimodal analysis of floor control in meetings. In *Proc. of MLMI 06 (submitted)*, Washington, DC, USA, May.
- Lei 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 Revisited*, 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 algorithm for dynamic gesture tracking. In *ICCV Workshop on RATFG-RTS*, Gorfou, 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*.