

Locate Salient Portions of Meeting Using Multimodal Cues

Lei Chen

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

1 Introduction

Meetings are gatherings of humans for the purpose of communication. Such communication may have various purposes: planning, conflict resolution, negotiation, collaboration, confrontation, etc. Facing the increasing size of meeting recording archives, it is becoming more important to efficiently browse, summarize, and retrieve important information in the meetings.

In this proposal for the AMI training program, I describe my planned research: locating salient portions of a meeting based on multimodal cues, including verbal cues and non-verbal cues, e.g., prosody, gaze and gesture. With the knowledge of the salient portions, the textual content of these regions can be used for summarization, and the audio/video content can be used to support audio/video skimming.

The proposal is organized as follows: Section 2 briefly reviews previous research. Section 3 describes my research plan. Section 4 describes my previous research that is related to this project and tools/data that can be used to support this research.

2 Previous Research

To investigate meetings, several corpora have recently been collected, including the **ISL** audio corpus [1] from Interactive Systems Laboratory (ISL) of CMU, the **ICSI** audio corpus [2], the **NIST** audio-visual corpus [3], and the **MM4** audio-visual corpus [4] from the Multimodal Meeting (MM4) project in Europe. There is also some previous research that has been carried out to find the salient parts of meetings, e.g., “hot spots”, attention focus, important part, and so on, in order to support more effective browsing and summarization of meeting.

Waibel et al. [5] have proposed a meeting summarization system for their meeting browser [6]. The recognized sentences of a meeting are ranked according to the relevance, and users can control the relevance threshold to obtain summaries of different sizes.

“Hot spots”, are regions in which participants are highly involved in the meeting. These regions are likely to contain important information that is useful for meeting summarization. Wrede and Shriberg [7] have used prosodic cues, e.g., pitch and energy, to find the “hot spots” in the ICSI meeting data.

In [8], Arons showed that emphasized segments in long monologues could be extracted by locating segments with heightened pitch. He also showed that these emphasized segments tend to contain information that would be helpful for gathering summaries from speech. Kennedy and Ellis [9] ported Arons’s work to meeting recordings. They showed that a pitch-based scheme can locate emphasized segments in a meeting and that the extracted segments provide important information for understanding the meeting.

In [10], Erol, Lee, and Hull used multimodal processing to summarize meeting recordings. In particular, they used a well-known text retrieval method, Term Frequency-Inverse Document Frequency (TF-IDF). Using TF-IDF together with pitch-based emphasis detection (similar to [8]) and video activity change detection (such as a person entering a room), they were able to generate a video summary (or skim) of a meeting. A trivial model combination method, that is, just giving equal weights to predictions from textual, prosodic and visual cues, is used to generate the final video skim output.

For two important non-verbal communication behaviors, gaze and gesture, there are some emerging applications for the summarization. Participants’ gaze patterns are important for tracking their attention, which can be useful for finding important portions of a meeting. Stiefelhagen [11] has shown that a participant’s attention focus can be reliably detected by head orientation and such attention focus can also be used in meeting summarization. In a videotaped technical presentation, gesture, especially pointing to a location on a slide, has been used to obtain a planned summary of the presentation [12].

3 Planned Research

In previous research, individual modalities (e.g., text, prosody, gaze and gesture) have been investigated for meeting summarization; however, there has been little work concerning the combination of multimodal cues to locate salient portions of meeting audio/video recordings.

During the AMI training program, I plan to conduct research on locating salient portions of meetings using multimodal cues, including text, prosody, gaze, and gesture. The research plan is divided into four aspects.

metadata Some structural events used to enrich transcription (human or ASR output) of a meeting, such as DARPA EARS and Rich Transcription (RT) metadata [13, 14] (e.g., location and type of sentence unit) and dialog act [15] annotations have proven to be useful for meeting summarization. For example, the question/answer chain in a meeting may act as a major thread. Commands at the end of a meeting tend to be related to next steps in a work plan and so should have more weight in the meeting summarization. I plan to investigate the use of metadata information to identify salient portions of meetings and determine their layout.

prosody Pitch has been widely used in emphasis detection [8, 9, 10] and “hot spots” detection [7]; I plan to investigate additional prosodic features to detect the salient portions of meetings.

gaze/gesture In [10], some visual activities have been tracked and used to find the salient portions of meetings. However, gaze and gesture play a more important role in human communication. In a meeting, the participants may not make many gross movements; they may not frequently stand up or go to other locations, but they still have very active gaze and gesture behaviors. I will focus on these two aspects of the visual signals to investigate the relation between salient portions of a meeting and the gaze and gesture patterns of meeting participants.

model combination Given predictions from metadata/prosody/gaze and gesture cues, it is essential to effectively combine their predictions to make a final decision. I plan to investigate different schemes for multimodal combination.

4 Supporting Tools/Data

Human communication involves multimodal cues; non-verbal cues (e.g., gesture and gaze) play an important role in understanding of communication. The topic of my Ph.D thesis is utilizing gestural cues to detect structural events in human communication. Focusing on this topic, I have conducted research on gesture patterns during speech repairs [16], on sentence boundary detection using textual/prosodic/gestural cues [17], and on speaker floor change using multimodal cues. In this research, I have become expert with techniques for language processing (e.g., hidden-event language modeling), prosody processing [18], gesture processing, and multimodal fusion.

Currently, I am supported by the VACE (Video Analysis and Content Exploration) program supported by ARDA. Working with Vislab at Virginia Tech, McNeill’s lab at University of Chicago, Thomas Huang’s lab at UIUC, and AFIT, we have been collecting a multimodal meeting corpus and investigating multimodal behaviors in meetings. Details about the data collection can be found in [19]. The overall data collection process is summarized in Figure 1. Attributes of the data collected are listed in Table 1. The multimodal meeting data collected in the VACE program can be used in my research for the AMI training program.

References

- [1] Burger, S., MacLaren, V., Yu, H.: The ISL meeting corpus: The impact of meeting type on speech type. In: Proc. of Int. Conf. on Spoken Language Processing (ICSLP). (2002)
- [2] Morgan, N., et al.: Meetings about meetings: Research at ICSI on speech in multiparty conversations. In: Proc. of ICASSP. Volume 4., Hong Kong, Hong Kong (2003) 740–743
- [3] Garofolo, J., Laprum, C., Michel, M., Stanford, V., Tabassi, E.: The NIST Meeting Room Pilot Corpus. In: Proc. of Language Resource and Evaluation Conference. (2004)

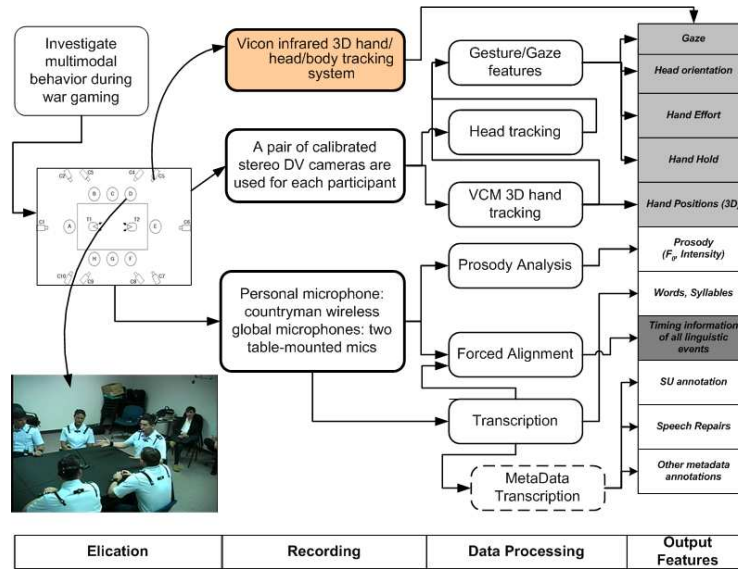


Fig. 1. VACE meeting corpus production

Video	MPEG4 Video from 10 cameras
Audio	AIFF Audio from all microphones
Vicon	3D positions of Head, Torso, Shoulders and Hands
Visual Tracking	Head pose, Torso configuration, Hand positions
Audio Processing	Speech segments, Transcripts, Alignments
Prosody	Pitch, Word & Phone duration, Energy, etc.
Gaze	Gaze target estimation
Gesture	Gesture phase/phrase, Semiotic gesture coding, e.g., deictics, iconics
Metadata	Language metadata, e.g., sentence boundaries, speech repairs, floor control change

Table 1. Composition of the VACE meeting corpus

- [4] McCowan, I., Gatica-Perez, D., Bengio, S., Lathoud, G., Barnard, M., Zhang, D.: Automatic analysis of multimodal group actions in meetings. *IEEE Trans. on Pattern Analysis and Machine Intelligence* **27** (2005) 305–317
- [5] Waibel, A. Bett, M., et al.: Advances in automatic meeting record creation and access. In: *Proc. of ICASSP*. (2001)
- [6] Waibel, A. Bett, M., Finke, M.: Meeting browser: Tracking and summarizing meetings. In: *Proc. of DARPA Broadcast News Workshop*. (1998)
- [7] Wrede, B., Shriberg, E.: The relation between dialogue acts and hot spots in meetings. In: *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, St. Thomas US VI (2003)
- [8] Arons, B.: Pitch-based emphasis detection for segmenting speech recordings. In: *Proc. of Int. Conf. on Spoken Language Processing (ICSLP)*. Volume 4., Yokohama, Japan (1994) 1931–1934
- [9] Kennedy, L.S., Ellis, D.: Pitch-based emphasis detection for characterization of meeting records. In: *Proc. of ICASSP*, Hongkong, China (2003)
- [10] Erol, B., Lee, D., Hull, J.: Multimodal summarization of meeting recordings. In: *IEEE Int. Conf. on Multimedia and Expo (ICME)*, Baltimore, MD (2003)
- [11] Stiefelhagen, R.: Tracking focus of attention in meetings. In: *Proc. of Int. Conf. on Multimodal Interface (ICMI)*, Pittsburg, PA (2002)
- [12] Ju, S.X., Black, M.J., Minneman, S., Kimber, D.: Summarization of videotaped presentations: Automatic analysis of motion and gesture. *IEEE Trans. on Circuits and Systems for Video Technology* **8** (1998)
- [13] Liu, Y., Shriberg, E., Stolcke, A., Peskin, B., Harper, M.: The ICSI/SRI/UW RT-04 Structural Metadata Extraction System. In: *Proceedings of EARS RT-04 Workshop*. (2004)
- [14] Liu, Y., Shriberg, E., Stolcke, A., Peskin, B., Ang, J., D., H., Ostendorf, M., Tomalin, M., Woodland, P., Harper, M.: Structural Metadata Research in the EARS Program. In: *Proc. of ICASSP*. (2005)
- [15] Ang, J., Liu, Y., Shriberg, E.: Automatic dialog act segmentation and classification in multiparty meetings. In: *Proc. of ICASSP*. (2005)
- [16] Chen, L., Harper, M., Quek, F.: Gesture patterns during speech repairs. In: *Proc. of Int. Conf. on Multimodal Interface (ICMI)*, Pittsburg, PA (2002)
- [17] Chen, L., Liu, Y., Harper, M., Shriberg, E.: Multimodal model integration for sentence unit detection. In: *Proc. of Int. Conf. on Multimodal Interface (ICMI)*, University Park, PA (2004)
- [18] Chen, L., Huang, Z.Q., Harper, M.: Praat based prosodic feature extraction toolkit. Technical report, Purdue University (2005)
- [19] Chen, L., Travis, R., Parrill, F., Han, X., Tu, J., Huang, Z., Kimbara, I., Welji, H., Harper, M., Francis, Q., McNeill, D., Duncan, S., Tuttle, R., Huang, T.: Vace multimodal meeting corpus. In: *Proceeding of MLMI 2005 Workshop*. (2005)