

Utilizing Gestures to Better Understand Dynamic Structure of Human Communication

Lei Chen

School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN 47907-1285
chenl@ecn.purdue.edu

Categories and Subject Descriptors: H.5.1, H.5.5, I.2.7.

General Terms: Algorithms, Performance, Experimentation, Languages.

Keywords: multimodal fusion, gesture, prosody, language models, sentence boundary detection, dialog.

1. THESIS ABSTRACT

Motivation: Many researchers have highlighted the importance of gesture in natural human communication. McNeill [4] puts forward the hypothesis that gesture and speech stem from the same mental process and so tend to be both temporally and semantically related. However in contrast to speech, which surfaces as a linear progression of segments, sounds, and words, gestures appear to be nonlinear, holistic, and imagistic. Gesture adds an important dimension to language understanding due to this property of sharing a common origin with speech while using a very different mechanism for transferring information. Ignoring this information when constructing a model of human communication would limit its potential effectiveness.

Goal and Method: This thesis concerns the development of methods to effectively incorporate gestural information from a human communication into a computer model to more accurately interpret the content and structure of that communication. Levelt [5] suggests that structure in human communication stems from the dynamic conscious process of language production, during which a conversant organizes the concepts to be expressed, plans the discourse, and selects appropriate words, prosody, and gestures while also correcting errors that occur in this process. Clues related to this conscious processing emerge in both the final speech stream and gestures. This thesis will attempt to utilize these clues to determine the structural elements of human-to-human dialogs, including sentence boundaries, topic boundaries, and disfluency structure. For this purpose, the data driven approach is used, which requires three components: corpus generation, feature extraction, and model construction.

Previous Work: Some work related to each of these components has already been conducted. A data collection and processing protocol for constructing multimodal corpora has been created; details on the video and audio processing can be found in the *Data and Annotation* section of [3]. To improve the speed of producing a corpus while maintaining its quality, we have surveyed factors impacting the accuracy of forced alignments of transcriptions to audio files [2]. These alignments provide a crucial temporal synchronization between video events and spoken words (and their components) for this research effort. We have also conducted measurement studies in an attempt to understand how to model multimodal conversations. For example, we have investigated the types of gesture patterns that occur during speech repairs [1]. Recently, we constructed a preliminary model combining speech and gesture features for detecting sentence boundaries in videotaped dialogs. This model combines language and prosody models together with a simple gestural model to more effectively detect sentence boundaries [3].

Future Work: To date, our multimodal corpora involve human monologues and dialogues (see <http://vislab.cs.wright.edu/kdi>). We are participating in the collection and preparation of a corpus of multi-party meetings (see <http://vislab.cs.wright.edu/Projects/Meeting-Analysis>). To facilitate the multi-channel audio processing, we are constructing a tool to support accurate audio transcription and alignment. The data from this meeting corpus will enable the development of more sophisticated gesture models allowing us to expand the set of gesture features (e.g., spatial properties of the tracked gestures). Additionally, we will investigate more advanced machine learning methods in an attempt to improve the performance of our models. We also plan to expand our models to phenomena such as topic segmentation.

2. REFERENCES

- [1] L. Chen, M. Harper, and F. Quek. Gesture patterns during speech repairs. In *Proc. of the Fourth International Conference of Multimodal Interface (ICMI)*, Pittsburg, PA, Oct. 2002.
- [2] L. Chen, Y. Liu, M. Harper, E. Maia, and S. McRoy. Evaluating factors impacting the accuracy of forced alignments in a multimodal corpus. In *Proc. of Language Resource and Evaluation Conference (LREC)*, Lisbon, Portugal, June 2004.
- [3] L. Chen, M. Harper, Y. Liu and E. Shriberg. Multimodal model integration for sentence unit detection. In *Proc. of Sixth International Conference of Multimodal Interface (ICMI)*, College Park PA, Oct. 2004.
- [4] D. McNeill. *Hand and Mind: What Gestures Reveal about Thought*. Univ. Chicago Press, 1992.
- [5] W. Levelt. *Speaking: from intention to articulation*. MIT Press, Cambridge, MA, 1989.

Copyright is held by the author/owner.

ICMI'04 Oct. 13–15, 2004, State College, Pennsylvania, USA
ACM 1-58113-954-3/04/0010.