Spontal – first glimpses of a Swedish database of spontaneous dialogue

Jens Edlund
KTH
Stockholm, Sweden
edlund@speech.kth.se

Abstract

This demonstration provides a first glimpse of a large multimodal database of Swedish spontaneous dialogue that is currently being collected within the ongoing project *Spontal: Multimodal database of spontaneous speech in dialog*. This accompanying paper briefly gives background and motivation for the project.

1 Introduction

This demonstration provides a first glimpse of *Spontal: Multimodal database of spontaneous speech in dialog*. The demonstration will touch upon annotation, audio, video and motion capture data, the recording studio, and some initial analyses.

2 Background

Spontal is an ongoing data collection project aimed at gathering multimodal data on spontaneous spoken dialogues. The project, which began in 2007 and will be concluded in 2010 and is funded by the Swedish Research Council, KFI - Grant for large databases (VR 2006-7482), It takes as its point of departure the fact that both vocal signals and gesture involving the face and body are key components in everyday face-to-face interaction – arguably the context in which speech was borne – and focuses in particular on spontaneous conversation.

There is a lack of data with which we can make more precise measurements of many aspects of spoken dialogue. We have for example an increasing understanding of the vocal and visual aspects of conversation, but there is little data with which we can measure with precision multimodal aspects such as the timing relationships between vocal signals and facial and body gestures. Furthermore, we need data to gauge acoustic properties that are specific to conversation, as opposed to read speech or monologue, such as those involved in floor negotiation, feedback and grounding, and resolution of misunderstandings. As a final example, there is a current surge in research on the related topics of incremental processing in dialogue on the on hand, and synchronous and converging behavior of interlocutors on the other – studies that are also hampered by a lack of data.

3 Scope

120 half-hour dialogues, resulting in a total in excess of 60 hours, will be recorded in the project. Sessions consist of three consecutive 10 minute blocks. All subjects are native speakers of Swedish and balanced (1) for gender, (2) as to whether the interlocutors are of opposing gender and (3) as to whether they know each other or not. The balancing results in 15 dialogues of each configuration: 15x2x2x2 for a total of 120 dialogues. Currently (May, 2009), about 45% of the database has been recorded. The remainder is scheduled for recording during 2009. Subjects permit, in writing, that (1) the recordings are used for scientific analysis, that (2) the analyses are published in scientific writings and that (3) the recordings can be replayed in front of audiences at scientific conferences and suchlike.

The recordings are comprised of high-quality audio and high-definition video. In addition, a motion capture system is used on virtually all recordings to capture body and head gestures, although the treatment and annotation of this data are outside the scope of the project and for this, resources have yet to be allocated.

4 Instruction and scenarios

Subjects are told that they are allowed to talk about absolutely anything they want at any point

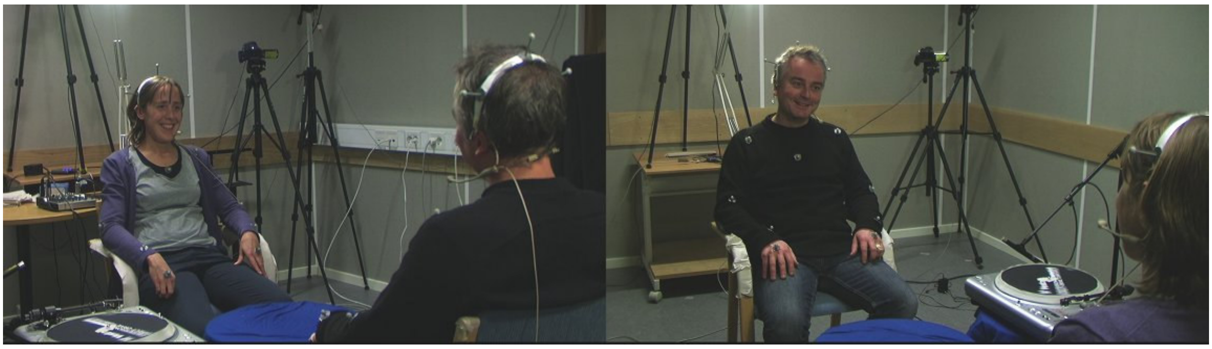

Figure 1. Example showing one frame from the two video cameras taken from the Spontal database.

in the session, including meta-comments on the recording environment and suchlike, with the intention to relieve subjects from feeling forced to behave in any particular manner. The recordings are formally divided into three 10 minute blocks, although the conversation is allowed to continue seamlessly over the blocks, with the exception that subjects are informed, briefly, about the time after each 10 minute block. After 20 minutes, they are also asked to open a wooden box which has been placed on the floor beneath them prior to the recording. The box contains objects whose identity or function is not immediately obvious.

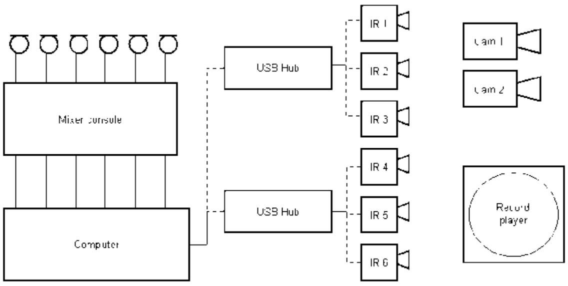

Figure 2. Schematic representation of the recording setup.

5 Technical specifications

The recording setup is illustrated in Figure 2. The audio is recorded on four channels using a matched pair of omni-directional microphones for high audio quality, and two headset microphones to facilitate subject separation for transcription and dialogue analysis. Two high definition video cameras are placed to obtain a good view of each subject from a height that is approximately the same as the heads of the subjects. The cameras work at 1920x1080 resolution at a bitrate of 26.6 Mbps. Audio, video and motion-capture are synchronized during post-processing with the help of a turntable placed between the subjects and a bit to the side, in full view of the motion capture cameras. A motion capture marker is placed near the edge on the turntable which rotates with a constant speed (33 rpm), enabling high-accuracy synchronization.

Figure 1 shows a frame from each of the two video cameras next to each other, so that both dialogue partners are visible. The opposing video camera can be seen centrally in the images, and a number of tripods with motion capture cameras are visible. Figure 3, finally, shows a 3D representation of motion-capture data. Each of the dots correspond to a reflective marker placed on the interlocutors' hands, arms, shoulders, trunks and heads, as can be seen in Figure 1.

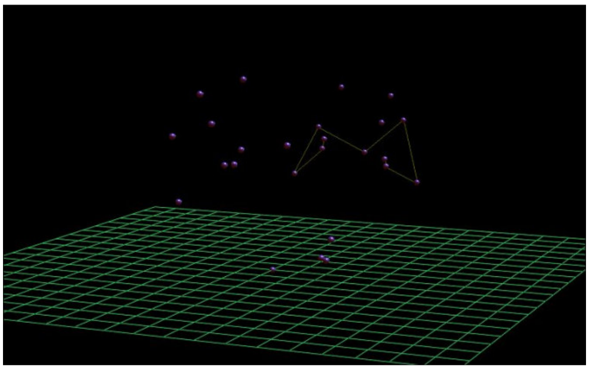

Figure 3. A single frame of motion-capture data from a Spontal dialogue.

The Spontal database will be made available to the research community after project completion. When recorded in its entirety, the Spontal database will be the largest of its kind in the world, and one of the richest dialogue data resources in Sweden.

Acknowledgments
Thanks are due to the Swedish Research Council for funding and support, the Spontal team at KTH Speech Music and Hearing for everything, and last but not least to all the volunteering subjects for making the data collection possible in the first place.