

Project presentation: Spontal – multimodal database of spontaneous speech in dialog

*Jonas Beskow, Jens Edlund, Kjell Elenius, Kahl Hellmer, David House & Sofia Strömbergsson
KTH Speech Music & Hearing, Stockholm, Sweden*

Abstract

We describe the ongoing Swedish speech database project Spontal: Multimodal database of spontaneous speech in dialog (VR 2006-7482). The project takes as its point of departure the fact that both vocal signals and gesture involving the face and body are important in everyday, face-to-face communicative interaction, and that there is a great need for data with which we more precisely measure these.

Introduction

Spontal: Multimodal database of spontaneous speech in dialog is an ongoing Swedish speech database project which began in 2007 and will be concluded in 2010. It is funded by the Swedish Research Council, KFI - Grant for large databases (VR 2006-7482). The project 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.

Although we have a growing understanding of the vocal and visual aspects of conversation, we are lacking in data with which we can make more precise measurements. There is currently very little data with which we can measure with precision multimodal aspects such as the timing relationships between vocal signals and facial and body gestures, but also acoustic properties that are specific to conversation, as opposed to read speech or monologue, such as the acoustics involved in floor negotiation, feedback and grounding, and resolution of misunderstandings.

The goal of the Spontal project is to address this situation through the creation of a Swedish multimodal spontaneous speech database rich enough to capture important variations among speakers and speaking styles to meet the demands of current research of conversational speech.

Scope

60 hours of dialog consisting of 120 half-hour sessions will be recorded in the project. Each session consists of three consecutive 10 minute blocks. The subjects are all 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. This balance will result in 15 dialogs of each configuration: 15x2x2x2 for a total of 120 dialogs. Currently (April, 2009), about 33% of the database has been recorded. The remainder is scheduled for recording during 2010. All subjects permit, in writing (1) that the recordings are used for scientific analysis, (2) that the analyses are published in scientific writings and (3) that the recordings can be replayed in front of audiences at scientific conferences and suchlike.

In the base configuration, the recordings are comprised of high-quality audio and high-definition video, with about 5% of the recordings also making use of a motion capture system using infra-red cameras and reflective markers for recording facial gestures in 3D. In addition, the motion capture system is used on virtually all recordings to capture body and head gestures, although resources to treat and annotate this data have yet to be allocated.

Instruction and scenarios

Subjects are told that they are allowed to talk about absolutely anything they want at any point 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. The subjects may then hold, examine and

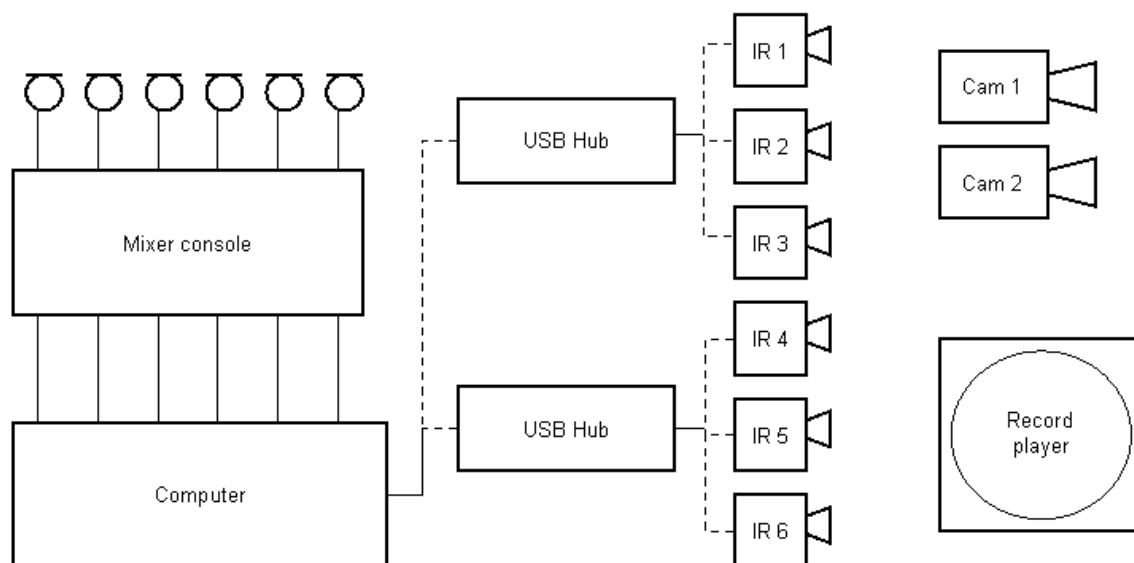


Figure 1. Setup of the recording equipment used to create the Spontal database.

discuss the objects taken from the box, but they may also chose to continue whatever discussion they were engaged in or talk about something entirely different.

Technical specifications

The audio is recorded on four channels using a matched pair of Bruel & Kjaer 4003 omni-directional microphones for high audio quality, and two Beyerdynamic Opus 54 cardioid headset microphones to enable subject separation for transcription and dialog analysis. The two omni-directional Bruel & Kjaer microphones are placed approximately 1 meter from each subject. Two JVC HD Everio GZ-HD7 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 both of the participating subjects. They are placed about 1.5 meters behind the subjects to minimize interference. The cameras record in mpeg-2 encoded full HD with the resolution 1920x1080i and a bitrate of 26.6 Mbps. To ensure audio, video and motion-capture synchronization during post processing, a record player is included in the setup. The turntable is placed between the subjects and a bit to the side, in full view of the motion capture cameras. The marker that is placed near the edge on the platter rotates with a constant speed (33 rpm) and enables high-accuracy synchronization of the

frame rate in post processing. The recording setup is illustrated in Figure 1.

Figure 2 shows a frame from each of the two video cameras aligned next to each other, so that the two dialog partners are both visible. The opposing video camera can be seen in the centre of the image, and a number of tripods holding the motion capture cameras are visible. The synchronization turn-table is visible in the left part of the left pane and the right part of the right pane. The table between the subjects is covered in textiles, a necessary precaution as the motion capture system is sensitive to reflecting surfaces. For the same reason, subjects are asked to remove any jewelry, and other shiny objects are masked with masking tape.

Figure 3 shows a single frame from the video recording and the corresponding motion-capture data from a Spontal dialog. As in Figure 2, we see the reflective markers for the motion-capture system on the hands, arms, shoulders, trunk and head of the subject. Figure 4 is a 3D data plot of the motion capture data from the same frame, with connecting lines between the markers on the subject's body.



Figure 2. Example showing one frame from the two video cameras taken from the Spontal database.

Annotation

The Spontal database is currently being transcribed orthographically. Basic gesture and dialog-level annotation will also be added (e.g. turn-taking and feedback). Additionally, automatic annotation and validation methods are being developed and tested within the project. The transcription activities are being performed in parallel with the recording phase of the project with special annotation tools written for the project facilitating this process.

Specifically, the project aims at annotation that is both efficient, coherent, and to the largest extent possible objective. To achieve this, automatic methods are used wherever possible. The orthographic transcription, for example, follows a strict method: (1) automatic speech/non-speech segmentation, (2) orthographic transcription of resulting speech segments, (3) validation by a second transcriber, (4) automatic phone segmentation based on the orthographic transcriptions. Pronunciation variability is not annotated by the transcribers, but is left for the automatic segmentation stage (4), which uses a pronunciation lexicon capturing most standard variations.



Figure 3. A single frame from one of the video cameras.

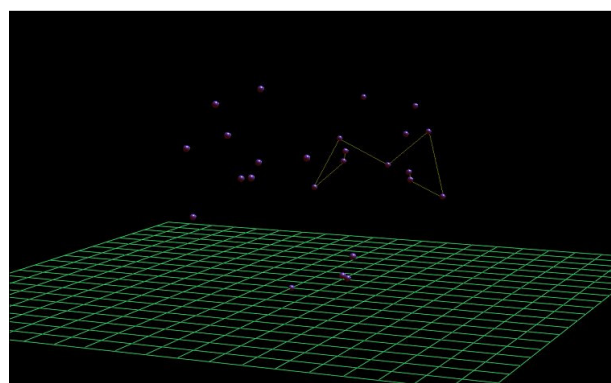


Figure 4. 3D representation of the motion capture data corresponding to the video frame shown in Figure 3.

Concluding remarks

A number of important contemporary trends in speech research raise demands for large speech corpora. A shining example is the study of everyday spoken language in dialog which has many characteristics that differ from written language or scripted speech. Detailed analysis of spontaneous speech can also be fruitful for phonetic studies of prosody as well as reduced and hypoarticulated speech. The Spontal database will make it possible to test hypotheses on the visual and verbal features employed in communicative behavior covering a variety of functions. To increase our understanding of traditional prosodic functions such as prominence lending and grouping and phrasing, the database will enable researchers to study visual and acoustic interaction over several subjects and dialog partners. Moreover, dialog functions such as the signaling of turn-taking, feedback, attitudes and emotion can be studied from a multimodal, dialog perspective.

In addition to basic research, one important application area of the database is to gain

knowledge to use in creating an animated talking agent (talking head) capable of displaying realistic communicative behavior with the long-term aim of using such an agent in conversational spoken language systems.

The project is planned to extend through 2010 at which time the recordings and basic orthographic transcription will be completed, after which the database will be made freely available for research purposes.

Acknowledgements

The work presented here is funded by the Swedish Research Council, KFI - Grant for large databases (VR 2006-7482). It is performed at KTH Speech Music and Hearing (TMH) and the Centre for Speech Technology (CTT) within the School of Computer Science and Communication.