Spontal: a Swedish spontaneous dialogue corpus of audio, video and motion capture

Jens Edlund, Jonas Beskow, Kjell Elenius, Kahl Hellmer, Sofia Strömbergsson & David House
KTH Speech Music and Hearing
SE-100 44 Stockholm
E-mail: {edlund, beskow, kjell, hellmer, sofia, davidh}@speech.kth.se

Abstract

We present the Spontal database of spontaneous Swedish dialogues. 120 dialogues of at least 30 minutes each have been captured in high-quality audio, high-resolution video and with a motion capture system. The corpus is currently being processed and annotated, and will be made available for research at the end of the project.

1. Introduction

Arguably, speech was conceived in a face-to-face interaction setting, and spoken dialogue is the cradle in which it evolved. As of this day, every-day face-to-face communicative interaction is the context in which most of our language use occurs. Vocalizations as well as gesture involving both the face and the body are important in much spoken communication, and in face-to-face conversation in particular. The Spontal corpus, a new Swedish dialogue corpus, provides recordings of audio, video, and three-dimensional motion capture of in excess of 60 hours of spontaneous dialogues. With its 120 half-hour multi-modal dialogue recordings, Spontal is the first and largest resource of its kind in Sweden. To our knowledge, it is also unique in the world at the point of writing.

The Spontal project is financed by the Swedish Research Council (VR 2006-7482), and takes as its point of departure the fact that 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 Spontal corpus is an attempt to remedy this situation for Swedish. The corpus is rich enough to capture important variations among speakers and speaking styles to meet the demands of current research of conversational speech.

2. Scope

The main corpus contains 60 hours of dialog consisting of 120 half-hour sessions. 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 balancing 15 dialogues of each configuration: 15x2x2x2 for a total of 120 dialogues. Currently (March, 2010), 116 of the 120 dialogues have been recorded. The remaining four will contain more precise capture of facial gestures, and are scheduled for recording during the first half of 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 such-like.

In addition to the main corpus, a number of smaller, specialized corpora are being recorded using basically the same configuration. Currently, this includes non-Swedish corpora, such as Spontal-N (Sikveland et al, 2010), a number of English recordings and single recordings in several other languages (e.g. Hindi, Syrian, and Italian); as well as a number of dialogues where additional sensors are used to capture more data. Part of the configuration was also used in the recording of the d64 database in Dublin (Oertel et al, 2010).

3. Recording configuration

3.1 Technical specification

In the base configuration, the recordings are comprised of high-quality audio, high-definition video, and motion capture data. Figure 1 shows an overview of the recording studio and Figure 2 shows the studio at the beginning of a session.

3.1.1. Audio

Each subject is recorded on two of microphones – a Bruel & Kjaer 4003 omni-directional goose-neck at 1m distance, and a head-mounted Beyerdynamic Opus 54 cardioid, which was replaced by a head-mounted Sennheiser ME 3-ew cardoid for the final 40 recordings. This combination of microphones is used to achieve optimal recording quality (B&K), while ensuring that we have recordings with a minimum of leakage from one speaker to the other (Beyerdynamic/Sennheiser). Mixer consoles

Figure 1: The recording studio. Video cameras are on tripods to the left and right in the picture. The turntable used for synchronization is visible just right of center, with the B&K microphones in microphone stands on either side. Motion capture cameras are barely visible, mounted along the walls just under the ceiling.

Figure 2: The recording studio during the instruction phase of a session. Subjects are wearing headmounted microphones and head bands with Optitrack markers, and additional markers are fixed to torsos, arms and hands.

are used as microphone pre-amplifiers and to supply phantom power to the microphones. The output of the consoles is connected to an M-Audio interface and recorded with the free audio software Audacity (Mazzoni 2008) in 4 channels 48 KHz/24 Bit linear PCM wave files on a 4 x Intel 2.4 MHz processor PC.

3.1.2. Video

Two JVC HD Everio GZ-HD7 high definition video cameras (1920x1080i resolution, 26.6 MBps bitrate) were placed with a good view of each subject, approximately level with their heads. The cameras are set so that the capture in full view a person with arms reaching out to each side.

3.1.3. Motion capture

In addition, six infrared OptiTrack cameras from Naturalpoint captures the participant's head, torso, and arm movements. The OptiTrack data consists of unsorted point clouds, and requires post processing to use. and Figure 3 shows the placement of motion capture markers in detail.

3.1.4. Synchronization

As each of the three systems used is susceptible to error, we take measures to ensure that the recordings can be synchronized even if frames are dropped or recordings partially lost due to for example hardware failure. The synchronization mechanisms evolved during the recordings as new obstacles appeared, but were stable from about 50% of the recordings and onwards, as follows. A record player is included in the setup. The turntable is placed between the subjects and to the side, in full view of both video and motion capture cameras. A marker is placed on the turntable, which rotates with a constant speed (33 rpm). A LP record is also placed on the turn-table. The record has been deliberately scratched at such a place that the pickup will hit the scratch each time the marker passes the pickup. The sound from the turntable is recorded on a separate fifth channel on the M-Audio interface. This setup enables high-accuracy synchronization of the frame rate of each of the systems in post processing.

A similar analogue system is used to synchronize the start and end of recordings. A box containing a green diode and an infrared diode is placed next to the turntable. When powered, the diodes light up and can be detected by automatic means both in the video and the motion capture. The power switch is in the control room, and is also connected to a sine tone generation. When the switch is

Figure 3: Each subject wears markers on hands, wrists, elbows, shoulders, sternum, and three markers mounted on a head-worn tiara.

flipped, the diodes light up and the sine tone is forwarded to a sixth channel on the M-Audio interface. We use three blinks at the start and end of each session, and two blinks to signify the start of each ten minute block. One blink is used to show that the recording staff has left audio comments on the sixth channel, which doubles as a synchronisation channel and an acoustic notebook.

3.2 Instructions and scenario

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. They were also told to not feel any pressure to keep the conversation going at all times, and that in case they felt like it, it would be ok to sit in silence as well. They recording studio is equipped with a intercom system which subjects are instructed to use to contact the recording staff if they feel inconvenienced. They are also told that they may stand up and leave at any point, should they feel the need to.

Figure 5: Two subjects lifting up the wooden box from the floor.

3.3 Physical objects

A recording is 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 over the intercom, 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. Figure 5 shows two subjects picking the box up. The box contains objects whose identity or function is not immediately obvious. The subjects may then hold, examine and discuss the objects taken from the box, but they may also chose to continue whatever discussion they are engaged in or talk about something entirely different.

3.4 Annotation

The Spontal database is currently being transcribed orthographically. The orthographic transcription includes orthographic words as well as labels for events such as breathing, coughing, laughing, and interactional tokens for turn management and feedback, such as eh, hm, okey, uh-huh. Care is taken to make the transcriptions quick to do and at the same time as consistent as possible.

Automatic methods are used wherever possible. The first step is to segment the audio into speech and non-speech segments in each of the speaker channels. Annotators then transcribe the speech chunks. At the next stage, the transcriptions are fed to a forced alignment system which provides rudimentary phonetic transcriptions with onset and offset times. These are more reliable than the original speech/non-speech labels, and are used to modify the time stamps.

Video is also segmented and mapped to the speech data using automatic methods developed within the Spontal project. The start and end times are found by automatically locating frames where the green diode is lit. The position of the participants' heads are pointed out manually, but in a manner that makes it very quick: the annotator is shown an average still picture of a number of pictures extracted from the video, which makes it simple to point out where a participant's head was during most of the time in the recording. The head position info is fed to a video processor which extracts close-ups and smaller low-resolution videos that are used for overviews.

The motion capture data will be treated in a similar manner. It is, however, not part of the original Spontal project, and its processing will take longer to finish.

Higher-level manual annotation is outside the scope of the project. Parts of the data is however already being used in several newly inaugurated research projects, which will result in various annotations. Researchers using the data once it is available are naturally encouraged to share their annotations as well.

4. 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 corpus will make it possible to test hypotheses on the visual and verbal features employed in communicative behaviour covering a variety of functions. To increase our understanding of traditional prosodic functions such as prominence lending and grouping and phrasing, the corpus will enable researchers to study visual and acoustic interaction over several subjects and dialogue partners. Moreover, dialogue functions such as

the signalling of turn-taking, feedback, attitudes and emotion can be studied from a multimodal dialog perspective.

Although the recordings were only recently completed, and the preparation and basic annotation of the corpus is still in progress, interesting results have already been generated. The material has been used to demonstrate automatic techniques for measuring synchrony and convergence in dialogue (Edlund, Heldner & Hirschberg, 2009) and perception studies on the limits of perception of overlaps and gaps in turntaking as well as of question intonation are currently being undertaken.

We have also made several noteworthy observations throughout the recording process. Although subjects were not given any task or topic, nobody reported any difficulty in coming up with things to talk about. Our initial examinations of the data reveal that people speak about a great variety of things, from what seems like regular work meetings through surprisingly open-hearted gossip to common dinner table topics such as "where did you grow up" and "what do you do for a living".

Subjects were told that they may end a session at any time should they feel uncomfortable. They were also told that, should they feel that they had said something inappropriate, they could (at a later date) go through the material together with the recording staff and delete the offending utterance. Nobody chose to do either.

Finally, a large portion of subjects asked without prompting if they could participate again. Taken together, we take these observations as an indication that our subjects felt relaxed and untroubled by the whole recording situation, and we have good hope that the recorded data is representative of conversations people have in their everyday life. This is also the impression reported by our annotators on initial examination of the corpus.

The project is planned to extend through 2010, when automatic processing and basic orthographic transcription will be completed. At this point the database will be made freely available for research purposes.

5. 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.

6. References

The following examples (of fictitious references) illustrate the basic format required for conference proceedings, books, journals, articles, Ph.D. theses, and chapters of books respectively:

Oertel, C., Cummins, F., Campbell, N., Edlund, J., & Wagner, P. (2010). D64: A corpus of richly recorded conversational interaction. In *Proceedings of LREC 2010 Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*. Valetta, Malta.

Sikveland, R-O., Öttl, A., Amdal, I., Ernestus, M., Svendsen, T., & Edlund, J. (2010). Spontal-N: A Corpus of Interactional Spoken Norwegian. In *Proceedings of LREC*. Valetta, Malta.

Edlund, J., Heldner, M., & Hirschberg, J. (2009). Pause and gap length in face-to-face interaction. In *Proceedings of Interspeech 2009*. Brighton, UK.