Face-to-Face Interaction and the KTH Cooking Show

Jonas Beskow, Jens Edlund, Björn Granström,
Joakim Gustafson, and David House

KTH Speech Music and Hearing/Centre for Speech Technology, Lindstedtsvägen 24,
SE-100 44 Stockholm, Sweden
{beskow,edlund,bjorn,jocke,davidh}@speech.kth.se

Abstract. We share our experiences with integrating motion capture recordings in speech and dialogue research by describing (1) Spontal, a large project collecting 60 hours of video, audio and motion capture spontaneous dialogues, is described with special attention to motion capture and its pitfalls; (2) a tutorial where we use motion capture, speech synthesis and an animated talking head to allow students to create an active listener; and (3) brief preliminary results in the form of visualizations of motion capture data over time in a Spontal dialogue. We hope that given the lack of writings on the use of motion capture for speech research, these accounts will prove inspirational and informative.

Keywords: Face-to-face interaction, synchrony/convergence, motion capture.

1 Introduction

Human face-to-face interaction sets the ultimate example for spoken dialogue systems created to draw on a human metaphor [1]. As we are unlikely to generate flawless human behaviour within a foreseeable future, the target is often mitigated: "human enough that we respond to it as we respond to another human" [2]. Still, a number of aspects of face-to-face interaction – for example the temporal dynamics in the interactions and the relations between modalities – are largely unexplored. As an example, consider the well-known phenomenon that interlocutors are more similar to each other than to people they are not currently interacting with. It has been pointed out that this similarity may be better described as a dynamic process that develops throughout the interaction creating synchrony (e.g. [3, 4]). We may, using standard dictionary meanings of the words, distinguish between *synchrony* (i.e. things that happen at the same time or work at the same speed) and *convergence* (i.e. things that come from different directions and meet), as illustrated in Figure 2.

Whereas most studies of interlocutor similarity have focused on only a few data points per conversation (e.g. first half/second half) or even one data point per speaker, a much finer temporal resolution is needed to capture these dynamics. An understanding of these dynamics is crucial for our understanding of human communication. And for face-to-face interaction, all modalities must be taken into account. Many interactional signals, for example feedback and emphasis, seem to be expressed equally well in for example gestures, words, facial expressions or by prosodic means.

Although methods involving the analysis of large amounts of data have yielded unrivalled progress in other areas of speech technology, few data collections to date capture in full the multimodal nature of human face-to-face interaction. Motion capture

A. Esposito et al. (Eds.): COST 2102 Int. Training School 2009, LNCS 5967, pp. 157–168, 2010.

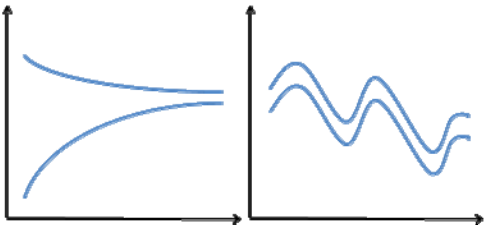

Fig. 1. Schematic illustrations of convergence (left pane) and synchrony (right pane)

has been used extensively by the entertainment business, and today the industry records scene specific motion capture data routinely. Motion capture has also been used to model animated agents in spoken dialogue systems, and some research data of acted dialogues have been recorded (e.g. CMU Graphics Lab Motion Capture Database). However, to our knowledge there is no sizeable database of unrestricted face-to-face conversations including motion capture data and high resolution video available for research. Nor have motion capture techniques been used to any significant extent in speech research. We are hoping that by gathering our experiences of working with motion capture in speech technology and dialogue research in this book chapter, we might inspire some to use the techniques and help others avoid pitfalls.

This book chapter presents a data collection effort, the Spontal project, in which 60 hours of unrestricted conversations between pairs of speakers is being recorded, with recordings capturing audio, hi-resolution video, and motion capture data (the latter is made possible in part because motion capture equipment has recently become considerably more available as well as affordable). Upon completion, the Spontal database will be the largest such data set to date. As the use of motion capture recordings as a means to study conversational behaviour is fairly new, this paper is an attempt to collect and share our experiences so far.

Following the presentation of the Spontal project, we show how the data collection techniques used in Spontal can be used to create small demonstrations and exercises. The motion capture equipment used in the project is quite portable, which has given us the opportunity to bring it to various locations and run tutorials and hands-on exercises on a number of occasions. On these occasions, part of the process is normally prepared in advance, be it for technical reasons, for expedience or simply to save participants some of the nitty-gritty. As this bears similarity with the format commonly used in televised cooking shows, we refer to these events as *cooking shows*. The exercises share several combined goals: (1) to showcase motion capture technology in a spoken dialogue context, as there is a great interest for this, (2) to teach how motion capture data can be used to investigate and to model face-to-face dialogue, especially since some of the experiences involved are hard-earned and many of the mistakes need not be repeated, and (3) to actually research face-to-face dialogue – the data collected during the exercises is valuable for hypothesis generation, as are comments and insights provided by students. Here, we describe exercises aimed at modelling *active listeners and listening speakers*, which were held at VISPP in 2008 and in part at the COST Spring School in Dublin in 2009.

We conclude this overview – or *cook book*, as we half jokingly call it – by presenting preliminary analyses investigating synchrony of movement to serve as an illustration of

how motion capture data can provide support for theories and intuitions regarding the dynamics of spoken dialogue.

2 Animated Talking Heads

Before moving on to motion capture, we will describe in brief the setting in which multimodal conversation is modelled at KTH Speech, Music and Hearing and the Centre for Speech Technology using in animated talking heads. The talking head developed at KTH is based on text-to-speech synthesis. Acoustic speech synthesis is generated from a text representation in synchrony with visual articulator movements of the lips, tongue and jaw. Linguistic information in the text is used to generate visual cues for relevant prosodic categories such as prominence, phrasing and emphasis. These cues generally take the form of eyebrow and head movements. Facial gestures can also be used as conversational gestures to signal such things as positive or negative feedback, control of the dialogue flow, and the internal state of a dialogue system (for a summary of spoken dialogue systems at KTH utilizing audiovisual synthesis, see [5]). More recently, we have been exploring data-driven methods to model articulation and facial parameters of major importance for conveying social signals and emotions [6].

Animated synthetic talking faces and characters have been developed using a number of different techniques and for a variety of purposes for more than two decades. Historically, our approach is based on parameterised, deformable 3D facial models, controlled by rules within a text-to-speech framework [7]. The rules generate the parameter tracks for the face from a representation of the text, taking coarticulation into account [8]. Several face models have been developed for different applications, some of them can be seen in Figure 1. All can be parametrically controlled by the same articulation rules.

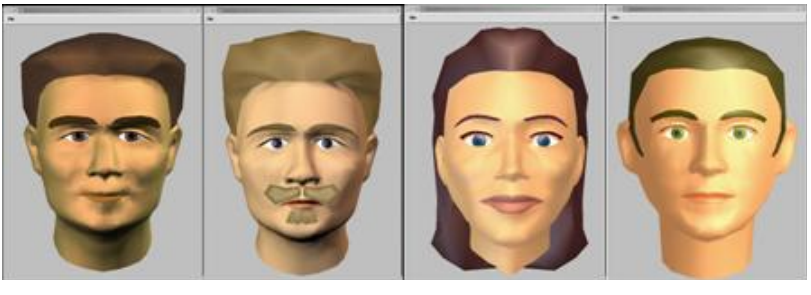

Fig. 2. Different versions of the KTH talking head

3 Spontal

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 goal of the project is to create 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.

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 (November, 2009), about 75% of the database has been recorded. The remainder is scheduled for recording during 2010. All subjects permit, in writing (1) that the recordings be used for scientific analysis, (2) that the analyses be published in scientific writings and (3) that the recordings can be replayed in front of audiences at scientific meetings for demonstration and illustration purposes.

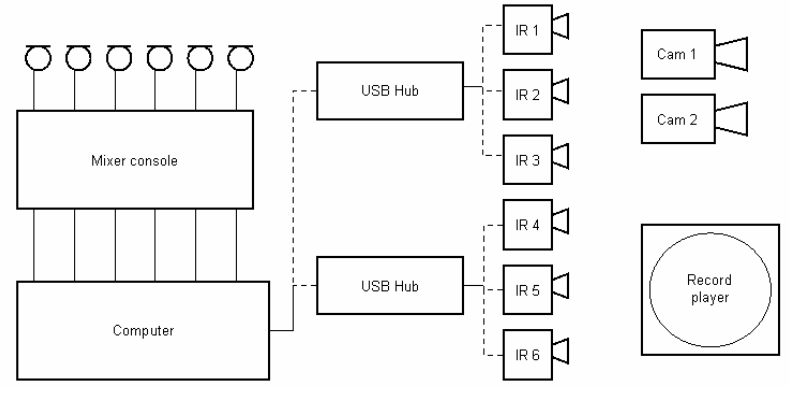

Fig. 3. Schematic over the recording setup used in Spontal

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.

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 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.

Audio is recorded on four channels using two omni-directional microphones for high audio quality, and two head-set microphones to enable subject separation for transcription and dialog analysis. Video is recorded on two high definition video cameras 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. 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. A marker placed near the edge on the turntable rotates with a constant speed (33 rpm), enabling high-accuracy synchronization and validation of the frame rate in post processing. The recording setup is illustrated in Figure 3.

Motion capture data is recorded on six NaturalPoint Optitrack FLEX:V100 cameras connected to the same computer over two externally powered USB hubs. The recording software is created in-house specifically for the purpose. Each subject is rigged with 12 reflective markers as seen in Fig 4.

1. Head – Center
2. Head – Left
3. Head – Right
4. Chest
5. Shoulder – Left
6. Shoulder – Right
7. Elbow – Left
8. Elbow – Right
9. Wrist – Left
10. Wrist – Right
11. Hand – Left
12. Hand – Right

Fig. 4. A Spontal subject rigged with 12 reflective markers

Markers 1-3 are mounted on 1.25" (3.175 cm) tall marker bases onto a plastic diadem (tiara) and fitted on the head of each subject, directed upwards. Marker 4 is placed in the center of the sternum, between 3 and 10 cm below the suprasternal (jugular) notch, depending on the clothing and any facial hair. Markers 5 and 6 are

placed on the outermost parts of the shoulders, as close to the acromion (bone in the shoulder above the deltoid muscle) as possible. Markers 7 and 8 are placed as close to the center of the elbow as possible. Markers 9 and 10 are placed just above the center of the wrist to avoid interference with movements of the hand. Markers 11 and 12 are placed close to the knuckles of the index finger and middle finger. Markers 4-12 are attached using adhesive pads. Finally one marker is placed on the turntable, making for a total of 25 markers. All markers are passive reflective markers with a diameter of 1.111 cm (7/16").

The motion capture data is saved directly to disk with no data manipulation or management at all during the recording session. There is no indexing or tracking of the recorded markers, so each frame contains 25 markers but there is no information as to which marker is placed where on the subject, nor which markers were or were not present in the previous frame. Markers may also be obscured from the cameras by the subject herself causing the marker to disappear for some number of frames. Ghost-markers can also appear in one or more frames – these are not actual markers but reflections on shiny objects that the cameras interpret as a marker. Another issue is that the system is not constantly running at 100 FPS. The frame rate varies between 99,8 and 100,2 FPS, and occasionally it varies around 64 FPS. If left alone, this problem causes significant synchronization issues rendering the motion capture data useless. However, the steady rotation of the turntable included in the setup can be used to control post process re-sampling.

Figure 5 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 turntable is visible in the left part of the left pane and the right part of the right pane. As in Figure 4, we see the reflective markers for the motion capture system on the hands, arms, shoulders, trunk and head of the subject.

Note that 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 jewellery, and other shiny objects are masked with masking tape.

Fig. 5. A single frame from the video recording

The Spontal database is currently being transcribed orthographically. Basic gesture and dialogue-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 variation.

Our recording experience so far have presented us with a number of more or less unexpected technical challenges which have been overcome. These include selection and installation of a light source for the video recordings which does not interfere with the motion capture cameras; shiny objects, eyes and eyeglasses which create spurious reflections interfering with the motion capture data; problems with USB power for the motion capture system; and synchronization problems finally solved by the use of the turntable. Fortunately, we have encountered far less obstacles related to our human subjects. Engaging in spontaneous and unstructured dialogues have presented no problems at all, nor have we seen any hesitation regarding coming up with topics of discussion during the interactions. Dividing the half-hour sessions into 10 minute sections and informing the subjects about the elapsed time after each section may reassure the subjects that all is proceeding well, and the introduction of the box after 20 minutes creates diversity in the conversational topics. Our transcribers' general impression so far is that all dialogues are spontaneous and unforced. It seems that the subjects quickly become rather unaware of the audio, video and motion capture equipment and busily proceed with their dialogues. The same observation has also been offered freely, albeit with some surprise, by many subjects after participating in a recording.

4 Modelling an Active Listener

The following section describes, in some detail, a three-session cooking show first held at the VISPP Summer School 2008 in Kuressaare, Estonia. VISPP in general focuses on speech variation in perception and production, and the summer school had talk-in-interaction, expressive speech and multimodal communication as its theme. Our intention in including it here is to show how motion capture data can be used not only to efficiently model aspects of face-to-face interaction, but also to illustrate as well as test intuitions and theories of such interaction.

The summer school attracted in excess of 30 participants, all PhD students, but of varied background. For this reason, we opted to investigate a behaviour that is recognized by everyone: the responses a listener provides to a narrative, often in the form of grunts and head-nods. The goal of the exercise was to increase the students' awareness of this mechanism by creating a rudimentary multimodal automatic active listener, using the KTH talking together with analyses and syntheses the students themselves created.

4.1 Data Recording

The students were initially divided in three groups, each of which did a recording. As the exercise aims to model an active listener, subjects are instructed to take different roles. One subject is the speaker (S) and is instructed to tell a story. Each group was asked to pick a member who enjoys narrating to take the role of S. The second subject is the active listener (AL). AL's role is to listen to S and provide feedback whenever she deems it suitable. Each group were asked to pick an attentive person for the role of AL.

A blacked-out hotel room served as the recording set. With the specific goal of this exercise, it is sufficient to record the motions of AL alone, but the audio from both subjects is required, and it needs to be reasonably well separated. Both subjects were rigged with close talking microphones, and AL was equipped with reflectors on the upper and lower lip and on each eyebrow – enough to model head pose, mouth opening and eyebrow movements. Limiting motion capture to one of the subjects allows us to use four cameras only (strictly speaking, three would suffice, but an extra camera improves results notably), which is a great relief in the confined space provided by a hotel room filled with spectators. In order to minimize the pressure on the subjects, no video equipment was used.

S was asked to think of a topic in advance, and be prepared to speak on it for five minutes. This presented no problems for the S in any of the three groups, although the groups occasionally found it hard not to giggle or become otherwise involved in the interaction.

The actual recordings took place sequentially over a 90 minute session, allowing each group 30 minutes. Each group was taken in to the recording set and given a full demonstration of the system, including a full calibration session, for the first ten minutes. The subjects were then rigged with reflectors and microphones and seated opposite each other, and S proceeded to narrate for five minutes. The last ten minutes were spent storing the recorded data and removing the equipment from the subjects while the group asked questions.

4.2 Analysis

In the next 90 minute session, students were divided into groups of two or three. Each group had students that had belonged to the same group on the recording session.

It is widely held that feedback responses occur more frequently after certain prosodic patterns – a notable example is Ward's description of how to decide where to insert feedback in a conversation automatically using pitch extraction (Ward, 1996). The analysis session was intended to give the students a feel for how pitch movements may be related to the occurrence of feedback from a listener. Since a considerable proportion of the students had no experience in phonetics, a graphical method of drawing templates to match pitch was used.

Students created pitch templates by (1) looking at pitch contours from S in Wavesurfer [9], and (2) selecting examples of contours preceding feedback (AL's audio track was included, as was a 3D rendition of the motion captured head movements) they felt were typical. Using a custom designed add-on to Wavesurfer, (3) these contours were extracted to a standard black/white image. They then (4) edited the image

in a standard image editing programme, making the contour wider in order to make it trigger more readily when applied to unseen curves (5). The end product of the analysis exercise was a number of templates – black/white images which were used to match a pitch contour segment. Pitch values inside the white area of the image scored positive and those inside the black areas scored negative, and a threshold on the total value was used to trigger feedback.

4.3 Resynthesis

The resynthesis session aimed at selecting audiovisual sequences for reproduction as canned speech and synchronised 3D-animated head pose, eyebrow and lip movement. Students were placed in the same groups as in the second session and were asked to select typical examples of feedback – auditory, visual and audiovisual. The audio was then replayed with gestures reproduced in an animated talking head. Students reported having no trouble finding good examples.

4.4 Final Session

On the plenary session of the final day, we applied one set of trigger templates and one set of audiovisual feedback segments to one of the narrations, resulting in an automated active listener that – according to the highly subjective views of the audience – did well. The take-home message is that given support and tools, it is possible for laymen to create a multimodal listener providing automatic feedback in less than a day's work, and hopefully gain some insights along the way.

5 A Glimpse at Synchrony in Motion Capture Data

As mentioned above, similarity between interlocutors is a process that can be established and that can develop during a dialogue. 3 employ an automatic approach to the analysis of video to quantify coincidental head movements (synchrony) between therapist and patient during psychotherapeutic sessions. Their approach employs Motion Energy Analysis (MEA) which is based on an image differencing algorithm that takes into consideration differences in the grey-scale distribution changes between subsequent video frames. By using motion capture data collected in the Spontal project, we are able to perform a similar automatic analysis of movement synchrony and produce a number of interesting and useful measurements which can contribute to new ways of analyzing dialogue behaviour.

As an illustration of how motion capture data can be used – even with very little processing – we picked a Spontal dialogue at random and prepared the motion capture data as follows. Note that we present no numerical evidence here. Such analyses would be moot, as the Spontal data needs to be validated and possibly resynchronized before it can be used as scientific evidence proper, as mentioned above.

For each frame, any data point whose position was within 0.5 metres from a point between the participants was discarded, creating a blind area of one metre width right between the speakers. This was done as a precaution to exclude the possibility that a motion capture track gets confused with another between the speakers. This precaution is well motivated for this type of study, as mixing tracks between speakers would

instantly create artificial synchrony. This treatment has a relatively small effect on the data, since the speakers are seated further apart than one metre, and the only frames that are lost are those where a speaker leans heavily forward. In the dialogue we used, it is less than 3% of the frames, divided in two continuous sequences.

For each sequence of two frames, the Euclidean distance between the points in each track was calculated. Tracks situated on one side of the blind area (thus belonging to the same speaker) were averaged, as were the tracks belonging to the other speaker. The result is a measure of *average marker movement* in the time between the two adjacent frames, per speaker. As mentioned before, the motion capture equipment aims at 100Hz, but this varies somewhat. However, as the synchronization between the cameras is highly accurate, we do not need to worry – each frame contains data sampled at the same time, which is all we need know for a rough investigation of synchrony.

Next, the two resulting time series of speaker average marker movement were down-sampled and smoothed using a 3000 frames (30s) long median filter with a step-length of 100 frames, resulting in heavily smoothed 1Hz data.

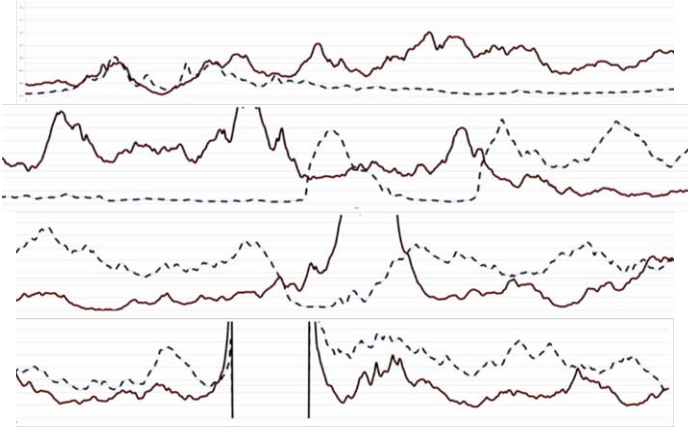

Fig. 6. Plot of Z-normalized average marker movement per speaker over time for the entire dialogue. Speaker A is represented by a solid line, speaker B by a dashed line. Finally, the data in both graphs is Z-normalized over speaker to make the plots easier to read and to facilitate comparison.

We note in Figure 6 that A and B show synchrony over a couple of segments, for example the first few minutes and in the bottom panel. When comparing to the video and audio, we note that the higher activity for A between 3 to 10 minutes into the dialogue, approximately, corresponds to a segment where she narrates a connected story, after which B takes over the narration, which corresponds to the higher activity of B at the end of the second and beginning of third panel. The gap in the final panel is caused by the Spontal setup: about 20 minutes into each recording, the subjects are asked to bend down and pick up a box that sits on the floor beneath them, place it on the table between them, and open it. While they do this, and while they are investigating the contents of the box, they lean into the blind area.

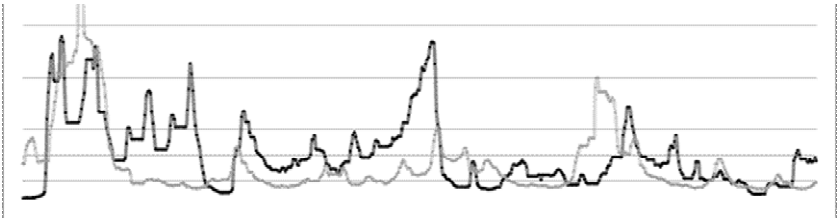

Fig. 7. Plot of Z-normalized average marker movement per speaker over time for the one minute of the dialogue. Speaker A is represented by a black line, speaker B by a grey line.

We also looked at this dialogue with a a 20 frames (200ms) long 20th percentile filter with a step-length of 10 frames, resulting in 10Hz data with very brief peaks removed. In this view, considerably more synchrony is visible on the micro-level, but the graph is too detailed and messy to reproduce here in its entirety. Figure 7 shows a small (1 min, starting at 11.5 minutes into the dialogue) section of that graph.

6 Summary

Studies of face-to-face human interaction are fundamental to basic research on speech in use, as it is arguably the most common and the oldest type of language use. They are equally important to speech technology development, at least to the extent that we aim to create human-like spoken dialogue systems – for example systems aiming to capitalize on the fact that humans already know how to communicate with each other. Face-to-face interaction is by its very nature multimodal, and we feel that there is a lack of substantial multimodal datasets. In particular, audio and video alone may not be sufficient to allow us to investigate in detail the relations between gesture, speech, and facial expressions. Adding motion capture provides precision data in three dimensions that we think will prove valuable in this respect.

As the use of motion capture in speech research is as of yet uncommon, we have presented here some of our experiences in recording face-to-face interactions with audio, video and motion capture. The technical setup for an ongoing large Swedish multimodal database project (Spontal) provided insights into the possibilities and pitfalls of large-scale dialogue recordings involving motion capture. The report on a tutorial in which students created an automated active listener using audio and motion-capture analysis from dyads engaged in narrations showed that current motion capture equipment is relatively portable and can be used for demonstrations, teaching and exploratory research in smaller, ad hoc settings. Finally, our very preliminary data visualizations exemplified how motion capture data may be used to produce interesting and useful measurements which can contribute to new ways of analyzing dialogue behaviour.

We note, naturally, that the tutorial recording settings do not meet research requirements – it is more than likely that the situation produced both subjective biases and artefacts of the extraordinary dialogue situation. Similarly, the analyses of synchrony presented last do not include statistics. The data is preliminary and not yet validated, which would make any statistical significance found void. In spite of this, we hope that given the lack of writing on the use of motion capture for speech research, our early accounts will be found inspirational and worthwhile.

References

[1] Edlund, J., Gustafson, J., Heldner, M., Hjalmarsson, A.: Towards human-like spoken dialogue systems. Speech Communication 50(8-9), 630–645 (2008)
[2] Cassell, J.: Body language: lessons from the near-human. In: Riskin, J. (ed.) Genesis Redux: Essays on the history and philosophy of artificial life, pp. 346–374. University of Chicago Press, Chicago (2007)
[3] Keller, E., Tschacher, W.: Prosodic and gestural expression of interactional agreement. In: Esposito, A., Faundez-Zauny, M., Keller, E., Marinaro, M. (eds.) Verbal and nonverbal communication behaviours, pp. 85–98. Springer, Berlin (2007)
[4] Edlund, J., Heldner, M., Hirschberg, J.: Pause and gap length in face-to-face interaction. In: Proc. of Interspeech 2009, Brighton, UK (2009)
[5] Gustafson, J.: Developing multimodal spoken dialogue systems. Empirical studies of spoken human-computer interaction. Doctoral dissertation, KTH, Department of Speech, Music and Hearing, Stockholm (2002)
[6] Beskow, J., Granström, B., House, D.: Analysis and synthesis of multimodal verbal and non-verbal interaction for animated interface agents. In: Esposito, A., Faundez-Zanuy, M., Keller, E., Marinaro, M. (eds.) Verbal and Nonverbal Communication Behaviours, pp. 250–263. Springer, Berlin (2007)
[7] Carlson, R., Granström, B.: Speech synthesis. In: Hardcastle, W.J., Laver, J. (eds.) The Handbook of Phonetic Science, pp. 768–788. Blackwell Publ. Ltd., Oxford (1997)
[8] Beskow, J.: Rule-based visual speech synthesis. In: Pardo, J. (ed.) Proc of the 4th European Conference on Speech Communication and Technology (EUROSPEECH 1995), Madrid, pp. 299–302 (1995)
[9] Sjölander, K., Beskow, J.: WaveSurfer - an open source speech tool. In: Yuan, B., Huang, T., Tang, X. (eds.) Proceedings of ICSLP 2000, 6th Intl Conf on Spoken Language Processing, pp. 464–467. China Military Friendship Publish, Beijing (2000)