Noname manuscript No.
(will be inserted by the editor)

D64: A Corpus of Richly Recorded Conversational Interaction

Catharine Oertel · Fred Cummins · Jens Edlund · Petra Wagner · Nick Campbell

Received: date / Accepted: date

Abstract In recent years there has been a substantial debate about the need for increasingly spontaneous, conversational corpora of spoken interaction that are not controlled or task directed. In parallel the need has arisen for the recording of multi-modal corpora which are not restricted to the audio domain alone. With a corpus that would fulfill both needs, it would be possible to investigate the natural coupling, not only in turn-taking and voice, but also in the movement of participants. In the following paper we describe the design and recording of such a corpus and we provide some illustrative examples of how such a corpus might be exploited in the study of dynamic interaction. The D64 corpus is a multimodal corpus recorded over two successive days. Each day resulted in approximately 4 hours of recordings. In total five participants took part in the recordings of whom two participants were female and three were male. Seven video cameras were used of which at least one was trained on each participant. The Optitrack motion capture kit was used in order to enrich information. The D64 corpus comprises annotations on conversational involvement, speech activity and pauses as well as information of the average degree of change in the movement of participants.

Keywords multimodality corpus · conversational involvement · spontaneous speech

Catharine Oertel
KTH Royal Institute of Technology Speech, Music and Hearing
Lindstedtsvägen 24, SE-100 44 Stockholm, Sweden
Tel.: +46-72 835 68 95
E-mail: catha@kth.se

1 Introduction

Developments in speech technology within the past decade have seen a shift of focus from the properties of speech as it is produced under controlled circumstances, to a broader concern with speech as it occurs in everyday contexts. With this shift has come an increasing focus on the properties of the context in which speech is produced, leading to the increased importance of rich multimodal recordings, including audio, video, and sometimes motion capture. We here present initial experiences with the D64 corpus, a rich multimedia corpus of conversational interaction recorded by the authors. We first discuss the motivation for recording the corpus, followed by a description of the recording situation itself.

Our initial work on this corpus has been mainly in the study of *conversational involvement*, as it evolves and is manifested among multiple participants in a multi-party context. We describe annotation procedures and report on several small studies that make use of both speech and non-speech indices of involvement. We close the paper with a discussion of the use of large, uncontrolled multimodal corpora for the study of conversational involvement.

Conversational involvement has been an active area of research since the 1960s (for an overview see [1]). Coker and Burgoon define conversational involvement as "the degree to which participants in a communicative exchange are cognitively and behaviourally engaged in the topic, relationship, and/or situation" [1]. In the studies presented here, we make use of a somewhat narrower interpretation this definition; we refer to conversational involvement only when interactants are in a conversation with each other.

If it were possible to build a system which is able to automatically predict the degree of involvement of interactants, it could be used for a variety of applications. One possible application lies in the aid of autistic children. If it was possible to equip children with a device, perhaps in the shape of a toy, that could signal to them when they should participate in a conversation, or when they should not this might help them to better integrate into society. A further possible application might lie in the field of human-machine communication. A statistical model of conversational involvement might be used in a third-party observers scenario such as in the case of a talk show moderator. A talking head such as the "Furhat talking head" [2] could be used to moderate a conversation. Everytime the conversation seems to become too intensive or just meander he could intervene and guide the conversation. Another possible application might be the time efficient query of very large multimodal databases [3] such as databases of video conferences. Applying a statistical model of involvement on the database might make it possible to automatically find the most important events within the conference. A model of conversational involvement might provide information on which participants are involved when and to which degree into the conversation. It might also provide information on the most important events in the conversation. The D64 database in general might also be used for studies on turn-taking, dialogue, gaze and gesture.

In this paper we are giving a new, extended viewpoint of the D64 corpus. We are including annotations and studies carried out on the D64 corpus and are relating them to each other.

2 Background

2.1 Speech and multimodal corpora

Face-to-face conversation is a fully embodied activity [4], i.e. the role of posture, eye gaze, torso movement, head rotation, hand and arm gestures all contribute to the dynamic establishment, maintenance, and dissolution of domains of joint attention [5–8]. The study of face-to-face conversation, then, requires multimodal corpora. But corpora also differ considerably in their design; there are those corpora which contain scripted or acted material such as [9] or [10], those containing task oriented dialogues such as [11], [12] or [13], meeting corpora such as [14], [15] or [16], corpora containing Wizard-of-Oz interactions [17], and corpora containing spontaneous, non-directed dyadic conversations recorded in laboratories such as [18], [19] or [20]. D64 fills a gap here, as none of these corpora is designed for the study of non-task driven conversational behaviour in a group of people interacting in a relaxed environment. A full description of the design considerations motivating the D64 corpus is provided in the section 3.

Table 1 Comparison of the D64-corpus in comparison to other studies.

corpus	**non-task directed**	**multi–party**	**long recording time**
D64	x	x	x
Cube G Corpus [9]			x
SaGA[11]			
AMI[14]		x	
IFADV[20]	x		

2.2 Conversational involvement

The pre-theoretical notion that participants engaged in conversational interaction do so with differing degrees of motivated involvement is not trivial to operationalize. Researchers have employed a range of methods for indexing involvement. Guerrero [21], for example, asked her interactants to fill out 8 7-point scales. The first five were designed to index how emotionally close the interactant was to his/her conversational partner and the final three were designed to index how carefully the interactant behaved towards his/her conversational partner with respect to impression management. In contrast, [22] used features associated in the literature with emotional over-involvement and

emotional under-involvement in order to detect those phases automatically in a video. Again in contrast, Wrede and Shriberg [3] chose a perceptual approach. They carried out a perception experiment in which they asked their subjects to identify "places in conversation where multiple participants get especially involved" [3]. They employed the notion of a "hot spot", or a locus of relatively high involvement of multiple interactants, and they implemented this by providing novice annotators with examples of different types of hotspots as previously annotated by an expert rater.

Research in the 1960's was mainly concerned with the conceptualization and identification of non-verbal indicators of conversational involvement. Mehrabian in 1969 was the first to identify "touch, distance, forward lean, eye contact, and body orientation" [23] as non-verbal indicators of involvement. In later studies, this list of cues was extended to include "body orientation, gaze, kinesic animation, vocal animation, conversational fluency, general interest, vocal interest, attention, general composure, infrequent random movement, bodily relaxation, vocal relaxation, infrequent nervous vocalisations, smiling, facial pleasantness, vocal pleasantness, relaxed laughter, proxemic distancing, degree of forward lean, silences, postural congruence" [21]. Guerrero et al. investigated whether behaviours to express involvement changes when subjects are talking to same-sex friends, opposite-sex friends and romantic partners and found that "[while] there is considerable behavioral consistency across relational partners, there are also important differences due to the relational partner with whom one is interacting" [21].

In addition to non-verbal cues, the correlation between speech and conversational involvement has been analysed as well. Wrede and Shriberg [3], for example, found an increase in mean and range of the fundamental frequency (F0) in more activated speech as well as tense voice quality. Moreover, Crystal and Davy [24] reported that, in live cricket commentaries, the more the commentator was involved in reporting the action (i.e. at the action peak), the quicker the speech rate. This contrasts with the findings of Trouvain, that the perceptual impression of increasing speech rate exhibited by the commentator during a horse race was not based on an increase in the rate of articulation, but rather in the increased frequency of breath pauses [25].

In the following study we build on these results. The annotation scheme we employ in the studies presented here takes the findings from Guerrero, Altman Wrede and Shriberg into account. In contrast to Guerrero, however, we try to design an approach which is not based on the subjective impression of the participants but rather how third party-observer judge the involvement of the participant on a 5 second basis. In contrast to all those approaches we present an annotation scheme with detailed descriptions of each annotation level.

A novel aspect of the present study is that we want to build a statistical model which is able to predict different degrees of involvement of a group of people rather than individual people. For this we build upon, and extent, the studies summarised here and include these cues into our statistical model.

3 The Corpus

There is widespread agreement that the empirical investigation of conversational interaction demands multimodal data [26]. This is important, both in furthering our understanding of naturally occurring human-human interaction, and in the development of systems that are required to interact in a human-like fashion with human speakers [27]. Along with audio recordings, it is now commonplace to include video recordings of at least the faces of conversation participants [20]. Speech is, however, thoroughly embodied, and unfettered conversational behavior includes appropriate manual gesturing, torso positioning, head direction, gaze behavior, blinking, etc. Furthermore, conversation is often carried out in a dynamic context, with free movement of the participants, change over time in the set of conversational participants, and with an openness that is entirely lacking from most careful studio recordings.

The D64 Multimodal Conversational Corpus [28] was collected to facilitate the quasi-ethological observation of conversational behavior. It set out to transcend many of the limitations associated with the use of read speech, structured dialogues, and task-driven interactions. To this end, conversation was situated in a domestic environment, and the flow of conversation was not prescribed, but emerged in unscripted and unguided fashion among the participants. This goal necessitated recording speech over several hours on two occasions, allowing initial self-awareness of the recording set up to fade into the background as the conversations themselves unfolded. We first outline the recording setup, the planned model of distribution, and finally, some of our initial aspirations in the analysis of the rich data that results.

The recording setup for data collection built on the following premises:

(1) The context of speech production ought to allow for natural conversation. This requirement demands qualification. Speech is hugely variable as a function of communicative context and ethological situatedness. The speech produced into a lover's ear is markedly different from that employed in the supermarket, yet both are spontaneous and natural. Even speech elicited under tightly controlled laboratory circumstances is natural [29], in that it is the spontaneous reaction of a person to a specific speaking context. We established a speaking context in which many of the properties of conversational speech could emerge without obvious artifact, and without direct modification by the constraints we as researchers imposed. The ethological aspirations of the project were best served by creating a context in which participants were free to move with respect to others as they wished, and to comport themselves as the conversational flow dictated, with minimal interference. A necessary corollary of this arrangement is that unanticipated forms of interaction and unexpected events are entirely possible and, indeed, desirable. Fig. 1 illustrates a notional scale of "naturalness" and is to be interpreted with all these qualifications.

(2) Unlike most corpus recordings (e.g. map tasks, tourist information scenarios etc.), the chosen setup was not task oriented. No agenda or set of topics was provided. The motivation behind this was to allow the speakers to focus on

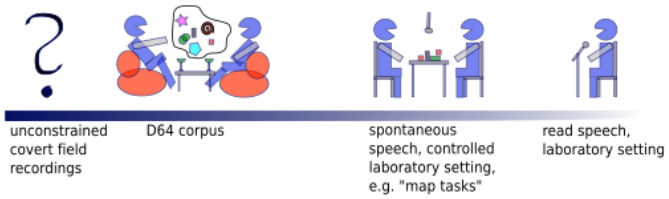

Fig. 1 Notional spectrum of observation scenarios ranging from highly controlled to relatively ethological. Qualifications of the concept of "naturalness" are provided in the text.

social interaction, rather than on the speech being produced. In task oriented dialogue, the linguistic exchange serves the collaborative achievement of a particular goal set by the task, e.g. to receive a particular kind of information or make an appointment. Clearly, social interaction does play an important role in task-oriented dialogue as well, but the removal of restriction with respect to topic and content facilitates a degree of social exchange that is quite distinct.

(3) Since the speakers knew that they would be recorded and filmed, our setup did not control for the observers paradox [30]. However, it had at least the following desirable properties:

- The conversation was interpersonal, with an active and involved other;
- It was both social and spontaneous;
- Participants were free to move around, or even leave;
- Speech was unprompted and unscripted, and the topics unrestricted;
- Recordings were made over a long period (8 hours over 2 days) thus helping to avoid stereotypical role playing, and reducing the chance of self-consciousness about the recording situation;
- Subjects shared many common interests, and subjective impressions of the interaction were that it was unforced.

Fig. 2 shows the domestic apartment room in which all recording was conducted. A mid-sized room with conventional furniture, with a sofa and some comfortable chairs arranged around a low coffee table was employed. Recordings were made over two days, each session being approximately 4 hours long, although the length of the corpus that will ultimately be made available has yet to be precisely determined. The first session was split into two two-hour sessions with an intervening lunch break, while the recording on the second day was continuous over 4 hours. Five participants (the first three and the last author and a friend) took part on Day 1, and just the 4 authors on Day 2. The participants ranged in age between early 20s and 60s. Two participants were native speakers of English (S1 and S3), one native speaker of German (S2), one native speaker of Swedish (S4) and one native speaker of Dutch (S5). All non-native participants had lived in an English speaking country for some time and had a high level of English competence.

In order to liven up proceedings somewhat, several bottles of wine were consumed during the latter half of recording on Day 2. Participants were free

to stand up, use the adjoining kitchen, change places, etc. throughout. In the same spirit, no attempt was made to constrain the topic of conversation, and subject matter varied widely from technological detail (inevitable under the circumstances) to pop culture and politics.

Fig. 2 The apartment room within which the D64 corpus was recorded.

As illustrated in Fig. 4, seven video cameras were employed. There was at least one camera trained on each participant (or one on the sofa as a whole, accommodating two participants). There were also two 360-degree cameras that captured the entire conversational field at a lower resolution. Audio was captured using both wireless microphones (both head-mounted and lapel), as well as a variety of strategically placed wide-field microphones. In addition, reflective markers (3 on the head, 1 on each elbow and shoulder and one on the sternum) were monitored by an array of 6 OptiTrack cameras. This data stream is however rather noisy and has yet to be processed. It is not used in any of the studies used as examples here.

The video cameras had the following spatial and temporal resolutions: The 360 degrees camera recorded 740X740 pixels at 30 frames per second. The hand cameras 2 and 4 (as can be seen in Fig. 4 have a resolution of 480x272 pixels and are recorded at 25 frames per second. Camera 3 is recorded with 920x1080 pixels at a frame rate of 25 frames per second.

In order to ensure subsequent synchronisation of all audio and video cameras after the recordings, we made sure to include loud claps at regular places over the course of the recordings. Synchronisation of all audio and video material has been carried out in the video editing tool *Final Cut Pro*.

4 Conversational Involvement

For a quantification of involvement three steps need to be accomplished:

1. Conversational involvement needs to be defined (see 2.2),

Fig. 3 representative views of three camera angles.

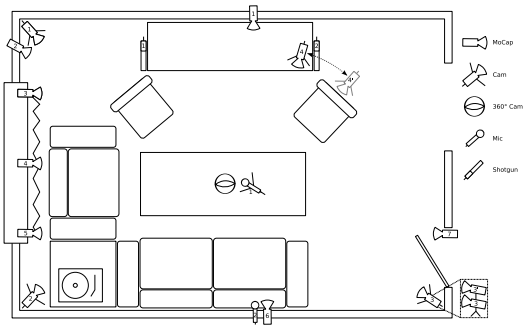

Fig. 4 representation of the room and the equipment used.

2. a suitable coding scheme needs to be found, and
3. a decision about which cues to use for the quantification needs to be made.

In the following section we focus on defining involvement. As when talking about the poorly defined notion of emotions, most people have a pre-theoretical idea of what concept is referred to when talking about conversational involvement. People are able to judge whether their conversational partner is interested, engaged or at least is following the conversation. Conversational interactants are sensitive to each other's conversational involvement constantly and change their conversational strategies accordingly. This process occurs automatically. Interactants do not typically think about precisely which behaviour in their conversational partner might trigger an impression of involvement in the conversation. In order to understand this phenomenon and be able to quantify it and make it usable for speech technology applications, it is necessary to find and understand how different behaviours conversational partners exhibit work together to give impressions of varying degrees of involvement in a conversation.

4.1 Annotation scheme

In order to evaluate conversational involvement, Coker and Burgoon suggested a five dimensional matrix in which they assess the degree to which participants in conversation engage in smooth-flowing conversation , "the degree of animation and dynamism", "the tendency to be interested in, attentive to, and adaptive to the other in a conversation", the "immediacy" in the behaviour of conversants as well as their "social anxiety" [1]. We tried to capture these 5 dimension in a 10 point annotation scheme as described in [31]:

Involvement level 1 is reserved for cases with virtually no interaction and with interlocutors not taking notice of each other at all, but engaged in completely different pursuits. Involvement level 2 is a less extreme variant of involvement level 1. Involvement level 3 is annotated when subgroups emerge. For example, in a conversation with four participants, this would mean that two subgroups of two interlocutors each would be talking about different subjects and ignore the respectively other subgroup. Involvement level 4 is annotated when only one conversation is taking place while for involvement level 5 interlocutors also need to show mild interest in the conversation. Involvement level 6 is annotated when conditions for involvement level 5 are fulfilled and interlocutors encourage the turn-holder to carry on. Involvement level 7 is annotated when interlocutors show increased interest and actively contribute to the conversation. For involvement level 8, interlocutors must fulfill the conditions for involvement level 7 and contribute even more actively to the conversation. They might for example jointly, wholeheartedly laugh or totally freeze following a remark of one of the participants. Involvement level 9 is annotated when interlocutors show absolute, undivided interest in the conversation and each other and vehemently emphasize the points they want to make. Participants signal that they either strongly agree or disagree with the turn-holder. Involvement level 10 is an extreme variant of involvement level 9.

This annotation scheme avoids making references to specific cues which might be expected to lead to or lower involvement. In this way subjects are free to follow their relatively intuitive judgements of involvement. Giving subjects room for interpretation, however, opens the question of whether subjects will be able to agree on distinct levels of involvement. In order to evaluate, a perception study was conducted which will be described in the following subsection.

In order to verify the involvement annotation scheme a perception test was conducted. Short video sequences were extracted from the D64 corpus and displayed on a website. A total of 10 video pairs was presented to the subjects. Subjects had to decide in which video there was more involvement and in which there was less. The inter-rater reliability was found to be $\kappa = 0.56$ for 30 raters [32]. For all subsequent studies, involvement annotation of one expert rater, which has been validated by 30 naive participants, is used.

In the following sections we discuss experiments carried out on the quantification of involvement by means of those cues. Table 2 provides an overview over the corresponding sections used in each study. Each of the sections con-

sists of approximately 30 min of speech. In "Day 1 Section 1" participants engaged in casual social talk, in "Day 1 Section 2" participants were engaged in animated discussions about a topic which lay in their shared professional background, "Day 2 Section 1" contained light hearted social talk. "Day 1 Section 2" was mainly dyadic in nature whereas in "Day 1 Section 1" and "Day 2 Section 1" all participants contributed actively to the conversation. This paper has been written with the focus on highlighting the D64 corpus and its usability for studies on conversational phenomena such as conversational involvement. All experiments discussed in the following sections are included to illustrate how the D64 can be used. The studies are based on only a subset of the data. The sections have been chosen on a qualitative basis and are believed to be representative samples of the entire two days recording session. It is planned to extend the analysis of conversational involvement to the entire D64 corpus at a later stage.

Table 2 List of sections of the D64 corpus used in the following studies.

Study 1	**Study II**	**Study III**	**Study IV**
Day 1 Section 1			Day 1 Section 1
Day 1 Section 2	Day 1 Section 2	Day 1 Section 2	Day 1 Section 2
	Day 2 Section 1		

4.2 Study 1: Accommodation in Voice and Conversational Involvement

The goal of this study was to examine mimicry in the voice and its relation to conversational involvement. In this study mimicry is defined, following Burgoon, as "The situation where the observed behaviours of the two interactants although dissimilar at the start of the interaction are moving towards behavioural matching" [33]. The aim of the study was to measure not only mimicry in the whole interaction but also to capture the dynamic changes which might occur in mimicry behaviour over the course of the conversation.

It was hypothesised that prosodic features of interactants become more correlated during periods in which the estimated conversational involvement was higher [34]. The authors interpreted the correlation of prosodic features during conversational interaction as a (weak) index of speaker mimicry.

For this study Section "Day1 Section 2" and "Day 2 Section 1" were used; both sections were mainly dyadic in nature. One male participant (S1) was involved in both dialogues. In the first section he conversed with a colleague of the same sex (S3), whereas in the second dialogue he talked to a female student (S2). The first section, participants were engaged in animated discussions about a topic that lay in their shared professional background. The

second section chosen for this study was extracted from the second day of recordings and contained light hearted social talk.

In order to calculate the mimicry strength (I) in prosodic cues, a 20 second window, with 10 second overlap among successive windows was used to extract corresponding values of median intensity and intensity variability, pitch range, pitch ceiling and mean pause duration of both inter-actants from the acoustic signal. This procedure resulted in 300 data points. Acoustic measurements were obtained using the phonetic software Praat. Pitch level and span were measured by calculating the F0-median and the log2(F0max -F0min) respectively. The F0-median is given on a linear scale (i.e. Hertz) while F0-maxmin is given on a logarithmic scale (i.e. octave). Silent pauses were detected automatically and corrected manually. Filled pauses, laughter and overlaps were excluded from the analyses. The intensity of the voice was expressed as the root mean square (RMS) intensity (rms-Int) and standard deviation Intensity (sd-Int). In order to account for speaker differences in prosodic parameters a log- (except for f_0-span) as well as an additional z-score transformation were applied. Fisher's transformation was applied to the transformation in order to decide where (I) is significant. In a first step (I) and F(I) were calculated for the entire conversation. In a second step it was aimed to identify temporal variations in mimicry strength by calculating the Pearson's correlation coefficient on a series of overlapping windows (20 points) using a time step 5 % of the series' length. In a final step the degree of mimicry strength was correlated with the degree of conversational involvement. Figure 5 illustrates how conversational involvement and mimicry in intensity evolve over time.

The study showed that prosodic cues can be used to measure and detect mimicry in speech. Mimicry calculation for the whole interaction did not yield significant results for the interaction in "Day 1 Section 2" between speakers S1 and S2. Yet mimicry was found for "Day 2 Section 1". Speakers S1 and S3 modulate their voice intensity level and variation (rms-Intensity, p=0.01539, sd-Intensity, p=0.00212), mean pause duration (dpauses, p=0.00256) as well as the ceiling of their pitch range (f0-max, p=0.01044) to match each other. Concerning temporal variations, S1 and S2 mimic each other in voice intensity (one phase of mimicry). Speaker S1 and S3 show two phases of mimicry in terms of f0- max and mean pause duration ($p<0.05$). Temporal variations of mean(I) for the interaction S1/S2 show a trend towards mimicry from point 9 towards the end, the mimicry strength being significant from point 17 to 20. Temporal variations of mean(I) for interaction S1/S3 enables the detection of one phase of mimicry, from the beginning of the interaction to point 9 ($p<0.05$) (Each point contains the information of 10 windows a 20 seconds; points are ordered according to when they occur in the corpus).

Conversational involvement is found to be strongly correlated with mimicry strength for the interaction S1/S2 (Rho=0.8889; $p = 0.0013$). It was concluded that the more as S1 and S2 are involved in the interaction with each other, the more they tend to mimic each other's speech prosody. It was therefore argued that the absence or presence of mimicry in speech prosody can serve as

a cue for the detection of degrees of conversational involvement in spontaneous conversation.

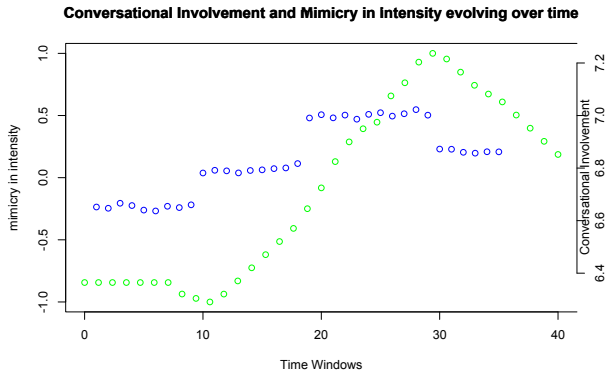

Fig. 5 Conversational Involvement (green) and Mimicry in rms-Intensity (blue) evolving over time.

4.3 Study 2: Gaze

The purpose of this study was to investigate whether the degree of mutual gaze is correlated with the degree of conversational involvement [35].

For this study "Day 2 Section 2" was used. This section was chosen as the conversations were mainly dyadic in nature and therefore better suited for the analysis of mutual gaze in conversation. Mutual gaze was manually annotated for two participants according to the annotation scheme proposed by Cummins [8]. Here, a binary distinction is made between gaze directed at the partner's face, and gaze directed anywhere else. Mutual gaze was calculated as the proportion of the overall duration in which Speaker 1 and Speaker 2 simultaneously looked at each other. For this a 20 second sliding window with 10 second overlaps among successive windows was chosen. The hypothesis was that the higher the amount of time partners spend in mutual gaze the greater the level of conversational involvement.

The proportion of time spent in mutual gaze was strongly correlated with the estimate of conversational involvement. Fig. 6 illustrates the relation between estimated involvement (X-axis) and the proportion of time spend in mutual gaze (Y-axis).The blue lines illustrate two separate linear regressions. It can be seen that the data are dichotomous towards the right hand side of the plot (greater involvement). This is readily accounted for, as Speaker 1, one of the principal participants, was making frequent use of her laptop during the conversation. The data was separated into two sets, one in which the laptop was being used, and one in which it was not being used. The relation between

estimated conversational involvement and proportion of mutual gaze for each set separately. The correlation between mutual gaze and involvement is R= 0.96 for the time speaker 1 uses her laptop and R=0.93 for the time she does not.

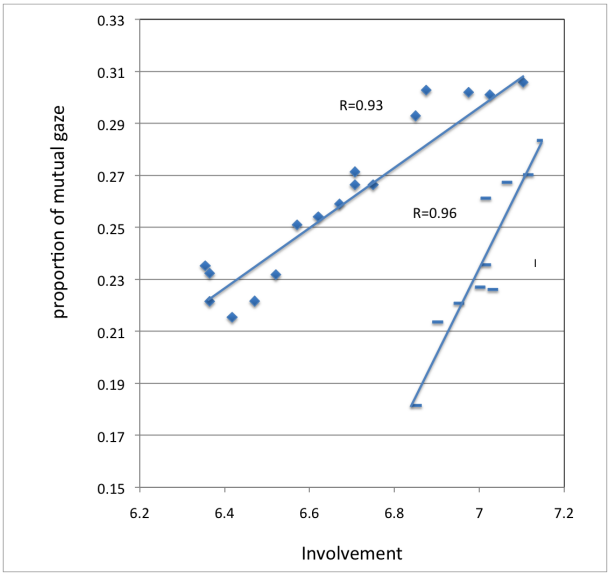

Fig. 6 Mutual gaze for the whole interaction as a function of estimated involvement for 20 second intervals with a moving window of 10 seconds.

4.4 Study 3: Automatic Prediction

Given the found interaction of acoustics and gaze with involvement, the purpose of this study was to investigate to what degree it is possible to predict the degree of conversational involvement based on acoustic and visual cues [35],[31]. A further aim was to investigate whether either acoustic cues, visual cues or a combination of the two lead to better prediction results. Table 3 shows the features used in the acoustic and visual model.

Here a distinction was made between two models; the acoustic model and the visual model. The acoustic model is based on "Day 1 Section 1" and "Day 1 Section 2". The visual model is only based on "Day 1 Section 2" alone. Furthermore, a distinction was made between a two class involvement model (Model I) and a three class involvement Model (Model II). In Model I, the first class contained data exhibiting low involvement (level 4, 5 and 6), and the second class contained data of high involvement (level 7, 8 and 9). Model II contained a class of low (level 4, 5 and 6) and class of high involvement (level 8 and 9). Moreover, an intermediate class (level 7) of involvement was

Table 3 List of features used in the acoustic and visual models.

acoustic model	visual model
f_0-median	(mutual) gaze
f_0-range	blinking rate
f_0-max	
f_0-sd	
f_0-min	
intensity	

introduced due to high proportion of annotations obtained for involvement level 7. For the prediction experiment support vector machines (SVM) with radial basis function kernels [36][37] were used.

Further, early (feature level) and late (decision level multiplication fusion) fusion approaches, making use of the two modalities provided, for the prediction of involvement within session II [38] were used. Early level fusion combines the extracted unimodal features to a multimodal representation of the observations before classification. In this case the alignment of the observations in the different modalities is crucial for the training of a single multimodal classifier. In contrast to early fusion, late fusion, or decision level fusion, combines the decisions of multiple unimodal classifiers. Typical combination schemes comprise majority vote, or multiplication fusion [39]. The two fusion schemes investigated, early and late fusion, differ with respect to the time at which the information of the different modalities is combined.

Table 4 Prediction results for involvement. ERR = Error rate reduction.

	Model I (two classes)		**Model II (three classes)**	
	mean acc.	**ERR**	**mean acc.**	**ERR**
Early fusion	0.7440	**0.11**	0.6820	**0.30**
Late fusion	0.7420	0.11	0.6420	0.26
Audio only	0.6940	0.06	0.5060	0.12
Video only	0.6640	0.03	0.6060	0.22

Table 4 illustrates the results of the experiments. Error rate reduction (ERR) is calculated as an improvement in accuracy rate from a hypothesised classifier relying on the a priori probability of the most likely class (for Model I that is class 2 with 0.63; for Model II that is class 2 with 0.38). The results are based on a standard 10 fold cross validation with a 90/10 split of the data.

Concerning Model I the best performance is achieved for an early fusion of both audio and video data (ERR = 0.11; accuracy = 0.7440). The late fusion of the audio and video data has a similar ERR of 0.11 and an accuracy of 0.7420. Video only produced lower accuracy and only a small reduction in error rate (ERR = 0.03; accuracy = 0.66). The single modality approaches

are both significantly outperformed by both of the fusion approaches in paired t-tests over the ten fold cross validation (late fusion vs. audio only $p = .011$; late fusion vs. video only $p = .023$; early fusion vs. audio only $p = .008$; early fusion vs. video only $p = .002$).

The best performance overall in terms of ERR is achieved for Model II using an early fusion of both acoustic and visual data (ERR = 0.30; accuracy = 0.6820). The late fusion of the acoustic and video data has a ERR of 0.26 and an accuracy of 0.64. The least accurate results are achieved for audio data only (ERR = 0.12; accuracy = 0.50). The early fusion again outperforms the single modalities significantly (early fusion vs. audio only $p < .00$; early fusion vs. video only $p = .002$), but late fusion only outperforms the audio only approach ($p < .001$). Further, video only outperforms audio only in the paired t-test with a p-value $< .001$.

In order to test how well the model generalises it was trained on session II and tested on session I and achieved a prediction accuracy of 0.5830 (ERR = 0.20) for Model II, which shows a good generalisation performance.

5 Discussion

We have described the motivation for recording an ethologically situated corpus of free conversational speech and its usefulness in studying conversational dynamics. The corpus is different from other speech corpora in many ways, but perhaps its most salient characteristic is the unscripted, long lasting, and animated nature of the conversations. In the landscape of already existing multimodal-corpora, it is probably most similar to the CRDO corpus [18] and the Spontal corpus [19] in that it contains spontaneous, non-task-directed conversations and to meeting corpora such as [14], [15] or [16] in number of participants. The corpus portrays the conversational behaviour of people interacting with up to 4 different people, differing in age, gender, social status, cultural background as well as degree of extroversion. It captures the conversational behaviour of people in both the listener and speaker role, and it portrays those same people in more relaxed as well as more formal situations. A significant advantage of the D64 corpus is that it is based on conversations of people who are intrinsically motivated to contribute to the conversation. This suggests a use for the corpus in the field of dialogue modeling that can not be obtained from structured task corpora. Moreover, the provision of multiple points of view in parallel video recordings allows considerations of embodied engagement in conversation in a manner not possible with audio only corpora.

There are inherent limitations to the approach taken, and there are contingent limitations that could be overcome in future recordings. An inherent limitation lies in the small number of interactants. Five people do not constitute a representative sample of any population, and the instinctive jump to generalization must be suitably constrained. One way to employ such a singular data set is to use it as the basis for the grounding of hypotheses for future studies, obtained using different participants. Limitations to video and audio

resolution are evident, though these fall into the class of contingent, rather than inherent, limitations. Audio recordings are of suitable quality for transcription and phonetic analysis, but not, for example, for voice source analysis. As we recorded in an apartment rather than an especially equipped recording booth the acoustic signal, despite the use of high quality microphones, is, to a certain extent noisy and can therefore not be used for voice source analysis. Video recordings support observation of gross body movement. Concerning gaze annotation only cameras 3 and 4 are suitable for gaze annotations. These cameras are the ones used for the analysis presented in this paper. The other cameras do not allow for gaze annotations due to the recording angle. In some rare instances participants pass in front of the camera such sections were excluded from the here presented analysis of gaze.

The small studies we have presented provide illustrative examples of how a rich topic, such as conversational involvement, can begin to be operationalized through the use of the richly multimodal data.

6 Conclusion

The main implication of the studies, as they concern this article, is that D64 indeed contains very data rich in dynamics, with sections in which a lot is happening and sections in which the conversation merely meanders. It is of particular interest for those researchers working on conversational behaviour analysis and dialogue modelling, in particular complex conversational behaviour influenced by situation, personality and conversational dynamics, as shown for the analysis of conversational involvement.

Final data synchronization of all audio and video data is currently underway. The final release of the corpus is planned to contain a master video, with all single videos merged into one video, as well as a master audio file which will have all head mounted microphones merged into one single wave file. In addition to the master files, all single audio and video files will be provided as well with information on their respective offsets. Speaker activity annotations, gaze and pause annotations will also be included. The release will be made available online, and free to download for all interested researchers under the Creative Commons Attribution-Noncommercial-Share Alike License.

Acknowledgements This study is in part supported by the Swedish Research Council project “The Rhythm of Conversation” (contract 2009-1766). Catharine Oertel wishes to acknowledge the Irish Research Council for Science, Engineering & Technology - Embark Initiative. Petra Wagner is being supported by the BMBF Professorinnenprogramm for young female professors. Nick Campbell wishes to acknowledge the FASTNET project - Focus on Action in Social Talk: Network Enabling Technology funded by Science Foundation Ireland (SFI) 09/IN.1/I2631.

References

1. D. A. Coker and J. K. Burgoon, "The Nature of Conversational Involvement," *Human Communication Research*, vol. 13, no. 4, pp. 463—-494, 1987.
2. S. Al Moubayed, J. Beskow, G. Skantze, and B. Granström, *Cognitive behavioural systems- Lecture Notes in Computer Science*. Springer, 2012, ch. Furhat: A Back-projected Human-like Robot Head for Multiparty Human-Machine Interaction.
3. B. Wrede and E. Shriberg, "Spotting "Hot Spots" in Meetings: Human Judgements and Prosodic Cues." in *Proceedings of Eurospeech 2003*, Geneva, 2003, pp. 2805–2808.
4. J. Cassell, T. Bickmore, L. M. Billinghurst, K. Campbell, H. Chang, V. lmsson, and H. Yan, "Embodiment in conversational interfaces: Rea," in *SIGCHI Conference on Human Factors in Computing Systems: the CHI is the limit,*, ACM New York, NY, USA, 1999, pp. 520–527.
5. D. Baldwin, "Understanding the link between joint attention and language," *Joint Attention: Its Origins and Role in Development*, pp. 131–158, 1995.
6. D. C. Richardson, R. Dale, and N. Z. Kirkham, "The art of conversation is coordination: Common ground and the coupling of eye movements during dialogue," *Psychological Science*, vol. 18, no. 5, pp. 407–413, 2007.
7. K. Shockley, D. Richardson, and R. Dale, "Conversation and coordinative structures," *Topics in Cognitive Science*, vol. 1, no. 2, pp. 305–319, 2009.
8. F. Cummins, "Gaze and blinking in dyadic conversation: A study in coordinated behavior among individuals," *Language & Cognitive Processes*, 2011.
9. M. Rehm, Y. Nakano, H.-H. Huang, A.-A. Lipi, and Y. Yamaoka, "Creating a standardized corpus of multimodal interactions for enculturating conversational interfaces," in *IUI-Workshop on Enculturating Interfaces*, Gran Canaria, 2008.
10. M. ZELEZNY, Z. KRNOUL, P. CISAR, and J. MATOUSEK, "Design,implementation and evaluation of the Czech realistic audio-visual speech synthesis," *Signal Processing*, vol. 83, no. 12, pp. 3657–3673, 2006.
11. A. Lücking, K. Bergman, F. Hahn, S. Kopp, and H. Rieser, "Bielefeld Speech and Gesture Alignment Corpus (SaGA)," in *LREC 2010.Workshop on Multimodal Corpora*, Valetta, Malta, 2010, pp. 92–98.
12. X. Sun, J. Lichtenauer, M. F. Valstar, A. Nijholt, and M. Pantic, "A Multimodal Database for Mimicry Analysis," in *4th Bi-Annual International Conference of the HUMAINE Association on Affective Computing and Intelligent Interaction (ACII2011)*, Memphis, Tennessee, USA, 2011.
13. D. Herrera, D. Novick, D. Jan, and D. Traum, "The UTEP-ICT Cross-Cultural Multiparty Multimodal Dialog Corpus," in *LREC 2010.Workshop on Multimodal Corpora.*, Valetta, Malta, 2010.
14. J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus," in *Language Resources and Evaluation*, 2007, pp. 181–190.
15. L. Chen, R. Travis-Rose, F. Parrill, X. Han, J. Tu, Z. Huang, M. Harper, F. Quek, D. McNeill, R. Tuttle, and T. Huang, "VACE Multimodal Meeting Corpus," *Lecture Notes in Computer Science*, vol. 3869, pp. 40–51, 2006.
16. N. Mana, B. Lepri, P. Chippendale, A. Cappelletti, F. Pianesi, P. Svaizer, and M. Zancanaro, "Multimodal Corpus of Multi-Party Meetings for Automatic Social Behavior Analysis and Personality Traits Detection." in *Workshop on Tagging, Mining and Retrieval of Human-Related Activity Information*, Nagoya, Japan., 2007, pp. 9–14.
17. G. McKeown, M. F. Valstar, R. Cowie, M. Pantic, and M. Schroeder, "The SEMAINE database: Annotated multimodal records of emotionally coloured conversations between a person and a limited agent," in *IEEE Transactions on Affective Computing*, 2012.
18. R. Betrand, P. Blache, R. Espesser, G. Ferre, C. Meunier, B. Priego-Valverde, and S. Rauzy, "Le CID- Corpus of Interactional Data- Annotation et Exploitation Multimodale de Parole Conversationelle." *Phonétique et Phonologie*, vol. 49, no. 3, 2008.
19. J. Edlund, J. Beskow, K. Elenius, K. Hellmer, S. Strmbergsson, and D. House, "Spontal: a swedish spontaneous dialogue corpus of audio, video and motion capture," in *Proc. of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner, and D. Tapias, Eds., Valetta, Malta, may 2010, pp. 2992 – 2995. [Online]. Available: http://www.speech.kth.se/prod/publications/files/3399.pdf

20. R. van Son, W. Wesseling, E. Sanders, and H. van Den Heuvel, "The IFADV corpus: A free dialog corpus," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, Marrakech, Morocco, 2008, pp. 501–508.
21. L. K. Guerrero, "Nonverbal Involvement Across Interactions with Same-Sex Friends and Romantic Partners: Consistency or Change ?" *Journal of Social and Personal Relationships*, vol. 14, no. 1, pp. 31–58, 1997.
22. U. Altman, R. Hermkes, and L.-M. Alisch, "Analysis of Nonverbal Involvement in Dyadic Interactions," in *Verbal and Nonverbal Communication Behaviours, LNAI 4775*, A. Esposito, Ed. Heidelberg: Springer-Verlag Berlin Heidelberg, 2007, pp. 37–50.
23. A. Mehrabian, "Methods and design: Some referents and measures of nonverbal behaviour," *Behavior Research Methods & Instruments*, vol. 1, pp. 203–207, 1969.
24. D. Crystal and D. Davy, *Investigating English Style*. London: Longman Group., 1969.
25. J. Trouvain and W. Barry, "The prosody of excitement in horse race commentaries," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.
26. D. Massaro and J. Beskow., "Multimodal speech perception: A paradigm for speech science," in *Multimodality in Language and Speech Systems*, B. Granström, D. House, and I. Karlsson, Eds. Dordrecht, The Netherlands: Kluwer Academic Publishers, 2002, pp. 45–71.
27. J. Edlund, J. Gustafson, M. Heldner, and A. Hjalmarsson, "Towards human-like spoken dialogue systems." *Speech Communication*, vol. 50, no. 8-9, pp. 630–645, 2008.
28. C. Oertel, F. Cummins, N. Campbell, J. Edlund, and P. Wagner, "D64: A corpus of richly recorded conversational interaction," in *Proceedings of LREC 2010; Workshop on Multimodal Corpora*, Valetta, 2010, pp. 27–30.
29. Y. Xu, "In defense of lab speech," *Journal of Phonetics*, vol. 38, pp. 329–336, 2010.
30. W. Labov, "Some further steps in narrative analysis," *Journal of Narrative & Life History*, vol. 7, no. 1-4, pp. 395–415, 1997.
31. C. Oertel, C. De Looze, S. Scherer, A. Windmann, P. Wagner, and N. Campbell, *Towards the Automatic Detection of Involvement in Conversation.* Berlin/Heidelberg: Springer-Verlag, 2011, pp. 163 – 170.
32. C. Oertel, "Identification of Cues for the Automatic Detection of Hotspots," Master's Thesis, Bielefeld University, 2010.
33. J. Burgoon, S. L.A., and L. Dillman, *Interpersonal Adaptation: Dyadic Interaction Patterns*. Cambridge: Cambridge University Press, 1995.
34. C. De Looze, C. Oertel, S. Rauzy, and N. Campbell, "Measuring Dynamics of Mimicry by Means of Prosodic Cues in Conversational Speech," in *ICPhS XVII*, Hong Kong, China, 2011, pp. 1294–1297.
35. C. Oertel, S. Scherer, and N. Campbell, "On the use of multimodal cues for the prediction of involvement in spontaneous conversation," in *Interspeech 2011*, 2011, pp. 1541–1544.
36. K. P. Bennett and C. Campbell, "Support vector machines: hype or hallelujah?" *ACM SIGKDD Explorations Newsletter*, vol. 2, no. 2, pp. 1–13, 2000.
37. B. Scholkopf and A. Smola, *Learning with kernels*. MIT Press, 2002.
38. F. Schwenker, S. Scherer, M. Schmidt, M. Schels, and M. Glodek, "Multiple classifier systems for the recognition of human emotions," in *9th International Workshop on Multiple Classifier Systems (MCS 2010)*, N. El Gayar, J. Kittler, and F. Roli, Eds. Springer, 2010, pp. 315–324.
39. L. Kuncheva, *Combining pattern classifiers: methods and algorithms*. Wiley, 2004.