D64- a Corpus of richly-recorded Conversational Interactions

Catharine Oertel, Fred Cummins,
Nick Campbell, Jens Edlund, and Petra Wagner
U. Bielefeld, University College Dublin, Trinity College Dublin, KTH, U. Bielefeld

nick@tcd.ie

INTRODUCTION

- we recorded a corpus
- a VERY big corpus
- of VERY natural spoken interactions .....
- with lots of gear
- and now we are trying to cope with the mess!

THE INITIAL MESS

SPOKEN INTERACTION

- people always talk
- whenever they sit down together something social happens
- we are interested in how that happens (as a process)
- so we sat down and recorded ourselves
- with mocap (x6) audio (x12) video (x5) 360-degree (x2) etc
- and talked for 2 days!

CORPUS COLLECTION

PEOPLE & CONDITIONS

- there was 1 naive volunteer amongst 5 participants
- (ethics-committee-agreed release form were signed by all)
- NO constraints were imposed on the content of the conversations - but all participants have the right to ask for the removal of any section of the corpus at any time
- and we are crowd-sourcing the annotations . . . . . .

CHATTING OVER DRINKS

A BRIEF HISTORY OF THE PROJECT

- atr/jst-crest esp corpus (Expressive Speech Processing)
- scope robot's ears project (round-table meetings)
- nov07 - friends sitting around a table in the lab
- aug09 - friends coming round for a sushi meal at home
- dec09 - friends in a room in dublin (drinkin' coffee & wine)
- fastnet! (focus on actions in social talk (network-enabling))

THE RECORDINGS

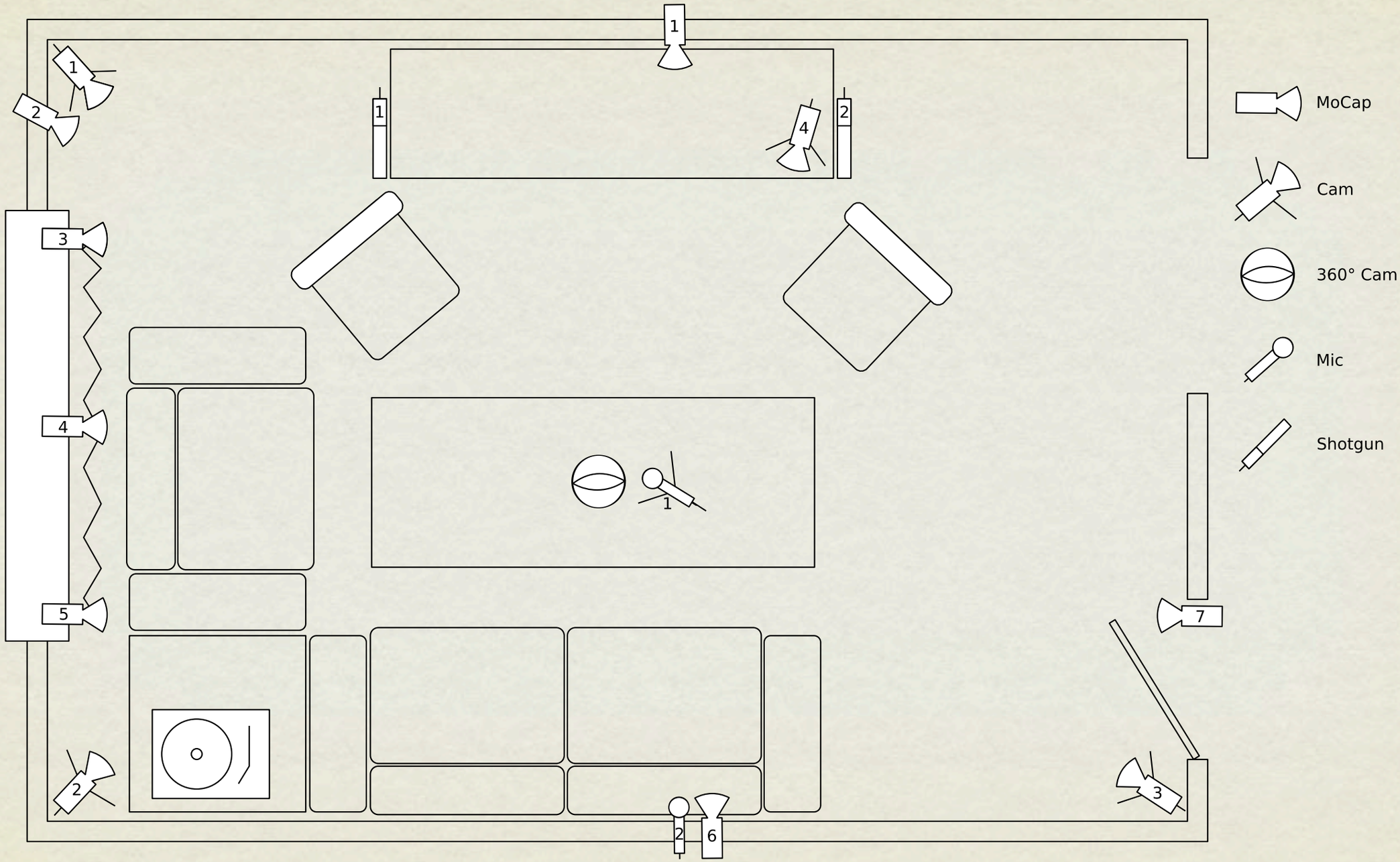
- D64 is the hotel room where Jens was staying (in Dublin)
- we wired it up - it was a mess - then we hid the wires
- we wore radio mics and were free to move around
- there was a kitchen area so we also ate & drank ..... wine too!
- we recorded a morning and an afternoon session on day 1
- and an all-day session on day 2 ending with a meal (& wine)

ALL-SEEING EYE

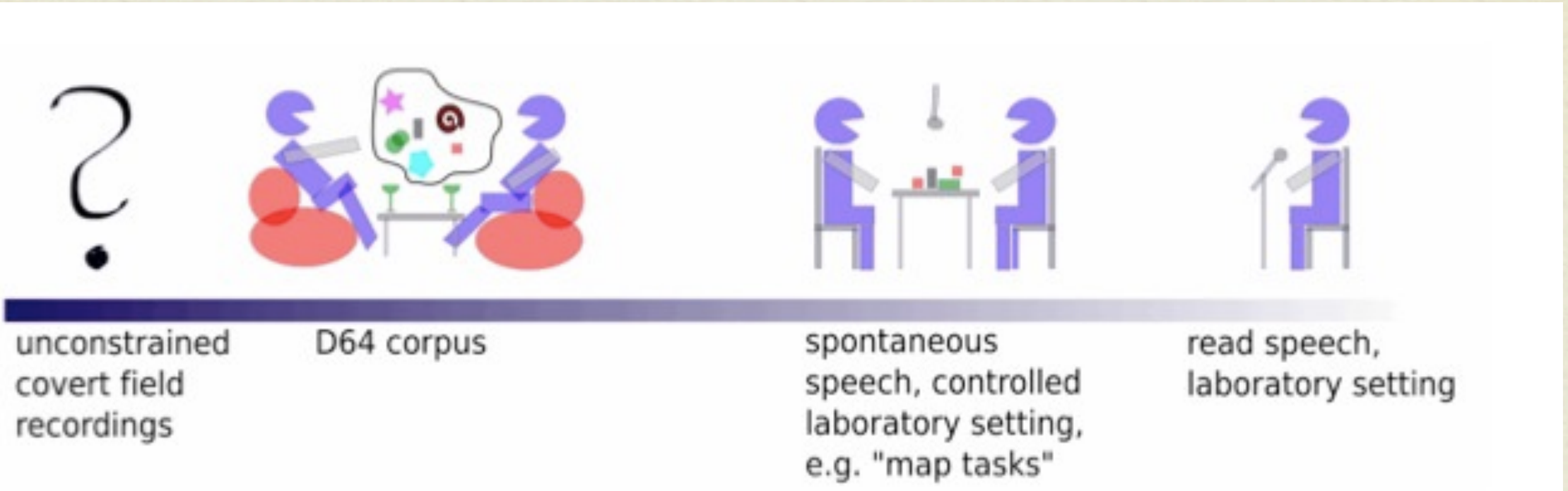
WIRED UP! (TRACKED)

CLEARED UP & WORKING

THE ROOM LAYOUT



ON A SCALE OF ‘REALITY’



- this is NOT read-speech nor scripted or prompted
- it is a spontaneous social interaction and has HIGH ‘reality’

SOCIAL INTERACTION

TALKING WHILE EATING

DATA HANDLING

Multimodal Expressive data (d64)

◆◆◆audio◆◆◆◆

unless you really need the original recording, try the mp3 first - for easier downloads!

	day1 am		day1 pm		day2	
left field	(big)	mp3	(big)	mp3	(big)	mp3
right field	(big)	mp3	(big)	mp3	(big)	mp3
mix1	-	a b	-	x	-	a b c d
mix2	-	a b	-	y	-	a b c d
Nike	(big)	mp3	(big)	mp3	-	-
Catha	(big)	mp3	(big)	mp3	(big)	mp3
Jens	(big)	mp3	(big)	mp3	(big)	mp3
Fred	(big)	mp3	(big)	mp3	(big)	mp3
Nick	(big)	mp3	(big)	mp3	(big)	mp3
table	???	???	???	???	(big)	mp3
sofa	-	a b	-	y	-	a b c d
sony-b66	-	a b	-	y	-	a b c d
head-mount					(big)	nick
sync	(big)	-	(big)	-	(big)	-

these tables were filled from logic rather than listening ...
from wiring diagrams rather than ears-on experience .....
they are bound to be wrong!!

◆◆◆video◆◆◆◆

most of these are compressed to mp4 - should download MUCH faster than the originals

	day1 am		day1 pm		day2	
sofa	all	cut	-	cut	-	a ??
armchair	a b c	-	a b c d e	-	-	a b c d e f g h i j k
chair	uncut	-	-	-	-	clearing up
sofa left	all	-	-	-	uncut	-
fred	-	1 2	-	3 4	-	5 6 7 8

see note above .. same applies here!

◆◆◆extras◆◆◆◆

	day1 am	day1 pm	day2
loops		(fred)	
360-degree camera	a b c d e f	a b c d e f g h i j k l m n o p q r s t u v w x y z aa ab ac ad ae	a b c d e f g
other round	1 2	3a 3b	4 5
mocap	set 1	set 2	set 3
4-track bkp audio	?1 ?2	x1 x2 x3 x4 x5 x6 x7 x8 x9	xa xb xc xd xe

◆◆◆photos◆◆◆◆

see here (see separate page)

AN ANALYSIS

- we were particularly interested here in ‘virtual spaces’
- we were all occupying the same physical space, but in different cognitive spaces - or attentional states
  - can distances be measured in these spaces?
  - does ‘proximity’ correlate with ‘arousal’?

AROUSAL & DISTANCE

- we hope to measure the ‘push&pull’ of social interaction . . . (cf Murray & Trevarthen - mothers & babies over video)
- what are the dynamics of spoken interaction?
- are there fundamental ‘laws’ or ‘forces’ that drive it?
- AROUSAL: when the group gets ‘heated’ - what happens?
- SOCIAL DISTANCE: when two people interact - is it ‘focus’?

THE EBB & FLOW OF JOINT INTERACTION

- simultaneous & real-time movements of several participants at once - like puppets being pulled by a common string . . .

SUMMARY

- this is a massively multi-modal 8-hour corpus
- this large-scale over-sampling provides very rich data
- it is available for research purposes (contact nick@tcd.ie)
- web-based interfaces are being developed and we hope that the availability of this multi-faceted data will encourage further research into the mechanisms of spoken-interaction

CONCLUSION

- this paper has presented a new corpus of multimodal data
- it features spontaneous multi-party social interaction
- captured in a multi-featured way - motion, audio, and video
- we are currently assembling and annotating the material
- it will be made available under www.speech-data.jp/mmx
- MMX is Multi-Modal-Expressive (2010 in roman numerals)

ACKOWLEDGMENTS

- This work has been supported by grants to Nick Campbell from the Visiting Professorships & Fellowships Benefaction Fund from Trinity College Dublin, and the Kaken-B Fund for Advanced Research from the Japanese Ministry of Information, Science & Technology, and also Science Foundation Ireland, Stokes Professorship Award 07/SK/I1218. Jens Edlund is supported by The Swedish Research Council KFI Grant for large databases (VR 2006-7482). Catharine Oertel is supported by the German BMBF female professors programme (Professorinnenprogramm) awarded to Petra Wagner.

- Finally, thanks to Nike Stam for her generous participation ;-)

THANK YOU FOR LISTENING