

12:25 - 13:25 昼休み

13:25 - 14:35 (基調講演)

マルチモーダルな会話データの収集と処理

ニック・キャンベル

Nick Campbell

(ダブリン大学、元神戸大学客員教授)

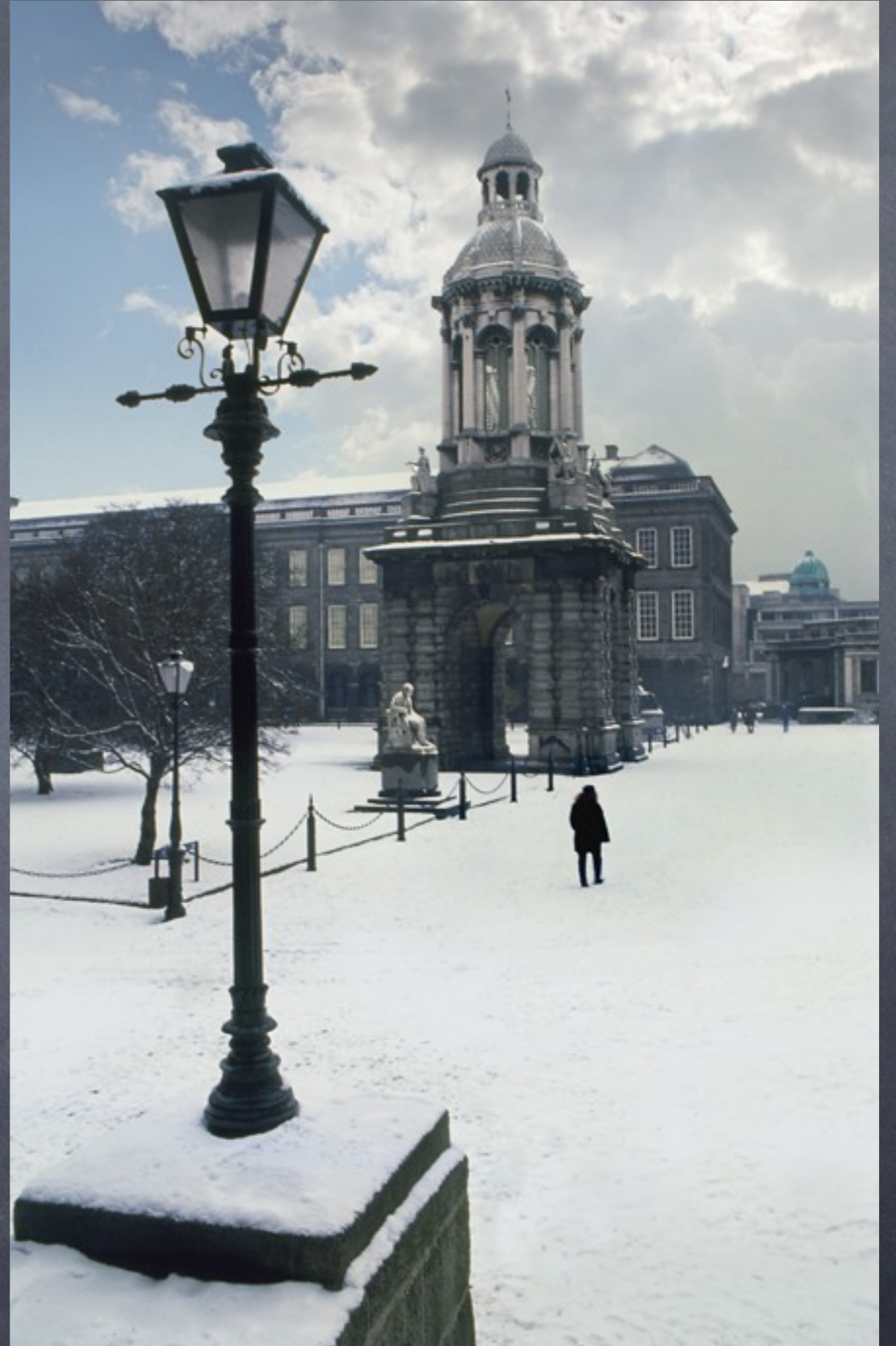
(nick@is.naist.jp)

nick@tcd.ie

Trinity College Dublin (since 1592)







scientists at work!



FASTNET - focus on actions in social talk



from 70 years ago ...

*QUANTITATIVE ANALYSIS OF THE INTERACTION OF
INDIVIDUALS*

BY ELIOT D. CHAPPLE

DEPARTMENT OF ANTHROPOLOGY AND DEPARTMENT OF INDUSTRIAL RESEARCH, HARVARD
UNIVERSITY

Communicated January 13, 1939

Up to the present time, no quantitative studies of human interaction have been undertaken, with the exception of observations by Dorothy Thomas and her co-workers on the behavior of children.¹ These studies were designed to develop an "index of personality" based on the percentage of total time spent by a child handling objects, interacting with people or playing alone. Only percentage figures were secured for interaction, and the authors were more interested in developing criteria for determining the reliability of observers.

This paper represents a preliminary account of quantitative results secured through the use of a crude recording apparatus, now being superseded by an accurate instrument. Although the investigation was primarily exploratory, it is believed that a discussion of the results obtained by use of the old instrument will be of interest as an indication of the

in Proc. Nat. Acad. Sci., V.25, N.1, 1939

historical note

In 1939, Eliot Chapple and his co-workers, adapted a manual typewriter* by fitting a small electric motor to the rubber shaft so that an operator could accurately record changes in subject activity over time, using a roll of adding-machine paper, for subsequent analysis.

* a machine pre-dating the word-processor, consisting mainly of levers, gears, and cogs, typically used for letter writing.

- Chapple observed the discourse actions of two individuals and obtained a series of durations of their actions, and studied how that sequence of durations was arranged
- flaws in his technique made it difficult to quantify overlapping speech and silent periods within the speech of one uninterrupted partner.
- his was the first recorded sequential analysis of human discourse behaviour
- an earlier study actually had made similar observations but with the goal of obtaining only percentage behavioural data

ethologists at work

“The first task of a human ethologist must be systematic description.

He must set out to see what behavioural structures the human being has

In doing this with people it would seem best to begin with those aspects of behaviour which are most likely to be shared with other animals

Thus while detailed analyses of language must eventually find a place in human ethology, these do not seem to be the best aspects of human behaviour with which to start”

(Adam Kendon)

machines watching people

Conversation Analysis and Discourse Analysis have also focussed on such sequence organisation

Our aim though is to produce a technology to observe 'the systematic, socially organised procedures underlying the ways in which social actors move into mutually ratified participation in an encounter,' which more recently Kendon has referred to as 'frame attunement'

another form of speech processing (not text-based)

- our goal: to produce a technology for tracking such discourse moves in conversational speech,
- we focus on the behaviour of the participants to make inferences about their discourse participation status, rather than focussing on the text or interpretation of their speech.

Current speech technology is based on text.

- People don't speak text, so there is often a mismatch between the expectations of the system and the performance of its users.
- Talk in social interaction frequently does involve the exchange of propositional content (which can be well expressed through text) but it also involves social networking and the expression of interpersonal relationships, as well as displays of emotion, affect, interest, etc.

Networking & social actions

- 'Networking' is an essential component of human interaction, and the content of a spoken conversation has as much to do with social bonding as with the transfer of propositional meaning
- we think that social interactions are the essential components of vocal communication, and that "actions, rather than words, are the prime units to be processed in a discourse"
- we aim for a paradigm shift in the way computers process speech, to incorporate 'speaking-style' information alongside 'message content' to provide a richer expression (or understanding) of an utterance.

FastNet's goals

- This new research project will generalise and extend previous findings from Japanese (JST/ESP) using new speech data of Irish and Irish-English conversations
- The academic goal of the research is to model this parallel channel of spoken communication, verifying its universality in human dialogue, but also illuminating the extent that it may take on language- & culture-specific forms
- The technical goal of this research is to produce speech technology specifically adapted to interactive or conversational speaking styles that will enable a friendlier and more efficient speech interface for public services, commerce, and entertainment

an 'intelligent' device

- The project aims to produce a device and methods for processing human speech (i.e., as part of a robot, an information-providing service, a translation service, or an entertainment system) which is able to process not just the text of that speech, but also able to interpret the intentions, or acts, of the speaker. It is not enough just to know what a person is saying; this research will help enable a machine to know what that person is doing with each utterance in a discourse.

the second channel

- The often fragmented and 'broken' nature of spoken language was long thought to be simply due to 'performance errors' while the underlying 'competence' of the speaker was supposedly better represented by a system such as that used for the written language.
- It now appears that this view of speech is incomplete, and that the frequent, repetitive, fragmentation of spoken language actually serves to carry a second channel of non-verbal information, essential to a proper understanding of the speaker's intentions.

patterns of sound
fragments of meaning

inter-activity

joint creation of
a meaningful social event

the JST ESP corpus

- Recent research sponsored by the Japan Science & Technology Agency (JST) enabled the collection of a very large 5-year in-situ corpus providing 1500 hours of manually-annotated natural conversational speech, and a realisation of how ordinary people use their voices in everyday social interaction.
- These data have yielded some surprising results, one of which is the large amount of non-verbal speech that is unobtrusively present in normal everyday conversation.

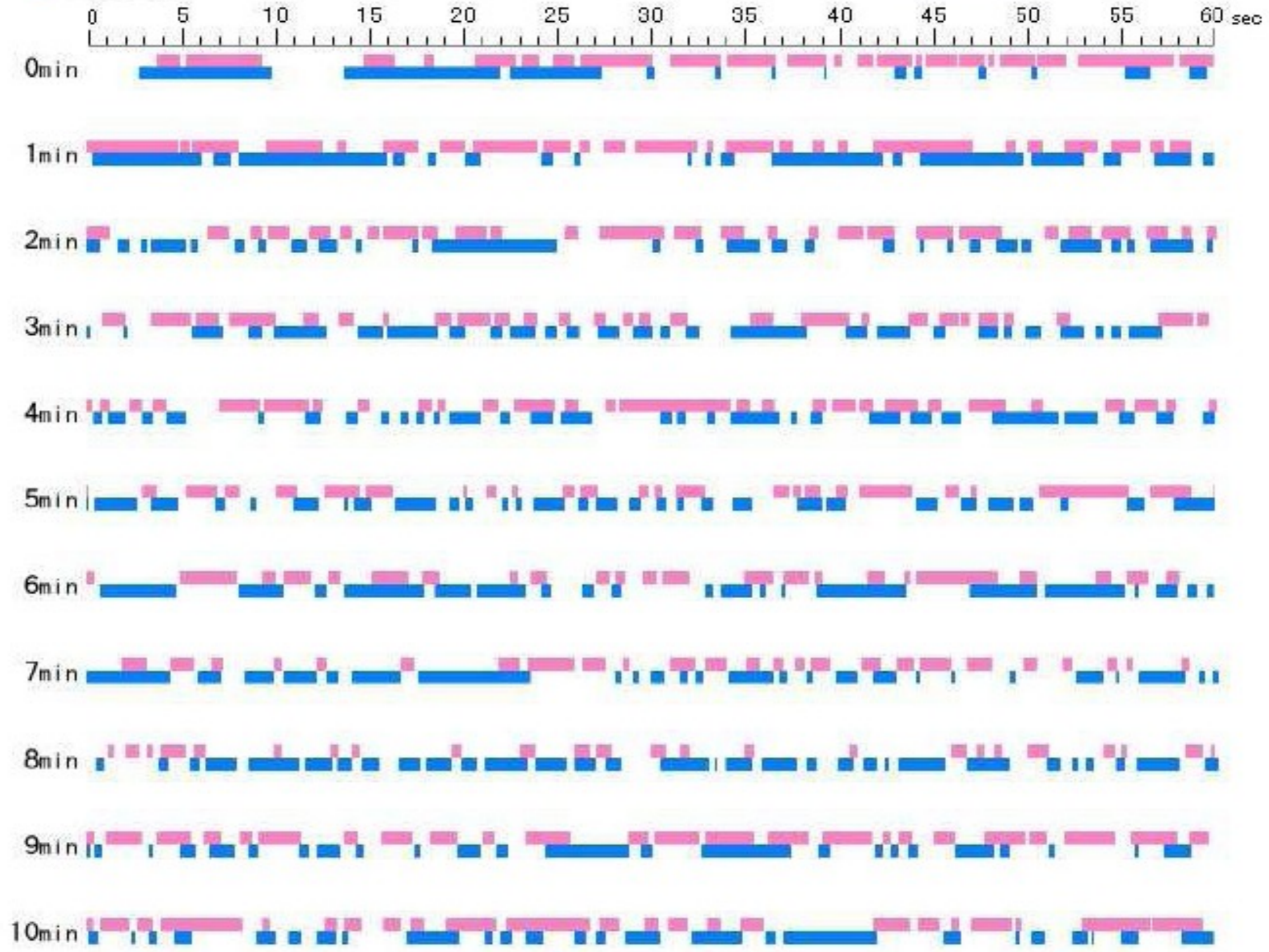
unobtrusive & ever-present



JMA_EFA_E06
EFA_JMA_E06

Conversation Chart

[Click here](#) to open LIST



activity patterns in telephone speech data

- patterns of interactive speech activity in two-party telephone speech data collected in the ATR/JST 'Expressive Speech' project
- *no constraints on the content of the conversations* (the partners were paid to talk to each other by telephone – with no face-to-face contact – for *30 minutes each week* for 3 months).
- All conversations were manually transcribed. The resulting text is VERY difficult to read! The speech is very fragmented.

frequent short sounds

- non-verbal speech signals

Counts of word usage according to conversational partner

JFA:	CFA	CMA	EFA	EMA	JFB	JMA
a,a-	143	145	88	89	138	170
ano	224	277	221	176	209	266
demo	41	24	31	17	89	134
e-	48	51	37	25	74	94
hai	2932	2234	2181	3239	72	33
un,un	1029	546	585	1190	909	1037

Frequency of 'hai' for JFA-JFB

26	JFA_JFB_J01
13	JFA_JFB_J02
7	JFA_JFB_J03
3	JFA_JFB_J04
1	JFA_JFB_J05
4	JFA_JFB_J06
3	JFA_JFB_J08
2	JFA_JFB_J09
3	JFA_JFB_J10
3	JFA_JFB_J12

the 100 most common utterances

10073	うん	467	ズー	228	ううん	134	へー
9692	@S	455	スー	227	えっ	134	はい.はい.はい.はい
8607	はい	450	んー	226	へー	134	そう.です
4216	laugh	446	うーーん	226	ハハハ	133	@E
3487	うーん	396	ねー	225	う.んー	133	あ.そう.な.ん.です.か
2906	ええ	395	あ.あー	200	そうですね	130	そう.な.ん.です.か
1702	はい	393	はい.はい.はい	199	ほー	129	はー
1573	うーん	387	あー.はい	193	ハー	129	い
1348	ズー	372	ねえ	192	その	127	ほー
1139	ふん	369	ふーん	190	え.えー	125	ハハハハハ
1098	あのー	369	だから	188	あ.あー	119	はい.はい
1084	あっ	368	あー.ん	187	ね	119	はー
981	はあい	366	ああ	180	ん.はい	114	ハハ
942	あの	345	あのー	180	あのー	113	は
941	ふーん	337	なんか	173	ん.ん	113	でー
910	そう	335	え	172	アハハハ	113	て
749	えー	311	でも	168	はいー	112	は.あー
714	あー	305	スー	164	う.うーん	110	フフフ
701	あ	274	うん.うん.うん	161	はー	110	そのー
630	あー	266	ハハハハ	160	@K	110	もう
613	あ.はい	266	てー	159	そう.です.ねー	109	ふーん
592	うん.うん	266	えー	151	あー	108	はあ.ー
555	あー	258	で	143	だから.ー	106	そうですね.え
500	んー	248	う	139	アハハハハ	105	んー.ん
469	ん	242	へー	137	そう.そう.そう	104	いや

- my most frequently used slide (no apologies!)

ほんま

- no talk would be complete without it

Microsoft Excel - FAN NJK test.xls																		
ファイル(F) 編集(E) 表示(V) 挿入(I) 書式(O) ツール(T) データ(D) ウィンドウ(W) ヘルプ(H)																		
E2479 = ほんま																		
B	E	G	H	J	K	L	M	N	T	U	V	W	X	Y	Z	AA	AB	AC
43				自分					内容					声の質				
44	file No.	text	あ、ほんま	感情状態	感情状態	機嫌	笑い	+α 自由コメ	興味	積極性	自信	+α 自本気度	エネルギー	明暗	硬柔			
45		発話	→	+α			laugh	+α 自由コメ	interest	aggress	confid	+α 自sincer	energy	bright	softness			
46																		
2096	10_20	あ、ほんま	「	↘	s-共感	s	気の毒に見	-1 どちらかといえば楽親身	2 ある	-2 消極	1 どちらかと言え	伝達	3 低い	-1 どちら	1 どちらかと言え			
2106	10_20	あ、ほんま	「	↘			沈んでいる	-2 楽しくない	元気が無い	-1 どちら	-2 消極	-	伝達	3 低い	-1 どちら	2 柔		
2126	10_20	あ、ほんま	「	↘			N	-1 どちらかといえば楽しくない	1 どちらかといえばある			伝達	4 抑えた	-	1 どちらかと言え			
2129	10_20	あ、ほんま	「	↘		s	がっかり	-2 楽しくない	1 どちら	-2 消極的		伝達	4 抑えた	-2 暗い	1 どちらかと言え			
2149	11_06	あ、ほんま	「	↘		s	不安	-1 どちらかといえば楽しくない	1 どちら	1 どちら	-1 どちらかと言	伝達	5 普通	-1 どちら	1 どちらかと言え			
2152	11_06	あ、ほんま	「	↘		s	不安	-1 どちらかといえば楽しくない	1 どちら	1 どちら	-1 どちらかと言	伝達	5 普通	-1 どちら				
2159	11_06	ほんま	「	↘			N	-1 どちらかといえば楽素直な感じ	-1 どちら	-1 どちら	-1 どちらかと言	伝達	4 抑えた	-1 どちら	-1 どちらかと言え			
2160	11_06	ほんま	「	↘	s-以外		N	-1 どちらかといえば楽しくない	1 どちらかといえばある			伝達	5 普通	-1 どちら				
2188	11_14	あ、ほんま	「	↘	s-以外		N	-1 どちらかといえば楽気安い	2 ある	1 どちら	1 どちらかと言え	伝達	4 抑えた	-1 どちら	1 どちらかと言え			
2218	11_14	ほんま	「	↘			疲れている	沈んでいる	-2 楽しくない	-2 無い	-2 消極的		伝達	2 沈んだ	-1 どちら	1 どちらかと言え		
2280	11_15	あ、ほんま	「	↘	s-以外		N	-1 どちらかといえば楽しくない	-1 どちら	-1 どちらかと言え	消極	伝達	4 抑えた	-1 どちら	1 どちらかと言え			
2294	11_15	あ、ほんま	「	↘			気の毒に見	-2 楽しくない	1 どちら	-1 どちら	-1 どちらかと言	伝達	4 抑えた	-2 暗い	1 どちらかと言え			
2295	11_15	あ、ほんま	「	↘			気の毒に見	-2 楽しくない	1 どちら	-1 どちら	-1 どちらかと言	伝達	4 抑えた	-2 暗い	1 どちらかと言え			
2329	11_15	あ、ほんま	「	↘		s	気の毒に見	-2 楽しくない	1 どちら	-1 どちらかと言え	消極	伝達	4 抑えた	-1 どちら	1 どちらかと言え			
2390	11_15	あ、ほんま	「	↘			N	1 どちらかといえば機嫌がいい	1 どちら	2 積極的	1 どちらかと言	伝達	6 活発	-	-			
2394	11_15	ほんま	「	↘			元気が無い	-2 楽しくない	-1 どちら	-2 消極的		伝達	3 低い	-1 どちら				
2433	11_15	あ、ほんま	「	↘	s-以外		N	-1 どちらかといえば楽しくない	2 ある	1 どちらかと言え	ば積極	伝達	5 普通	-	-			
2435	11_15	(あ、ほんま)	「	↘		s	気の毒に見	-2 楽しくない	1 どちら	-1 どちら	-1 どちらかと言	伝達	4 抑えた	-1 どちら	1 どちらかと言え			
2440	11_15	ほんま	「	↘			N	1 どちらかといえば機嫌s-元気		1 どちらかと言	え	積極	伝達	5 普通	1 どちら	-1 どちらかと言		
2446	11_15	あ、ほんま	「	↘			N(楽)	2 機嫌がs-smile	1 どちらか	1 どちらか	1 どちらかと言	伝達	5 普通	1 どちら	1 どちらかと言			
2463	12_02	あ、ほんま	「	→			不安	-1 どちらかといえば楽ぶりっこ	1 どちらか	1 どちらか	1 どちらかと言	伝達	5 普通	-1 どちら	2 柔			
2479	12_02	ほんま	「	→		s	心配	-1 どちらかといえば楽しくない	1 どちら	1 どちら	-1 どちらかと言	伝達	5 普通	-1 どちら	-1 どちらかと言			

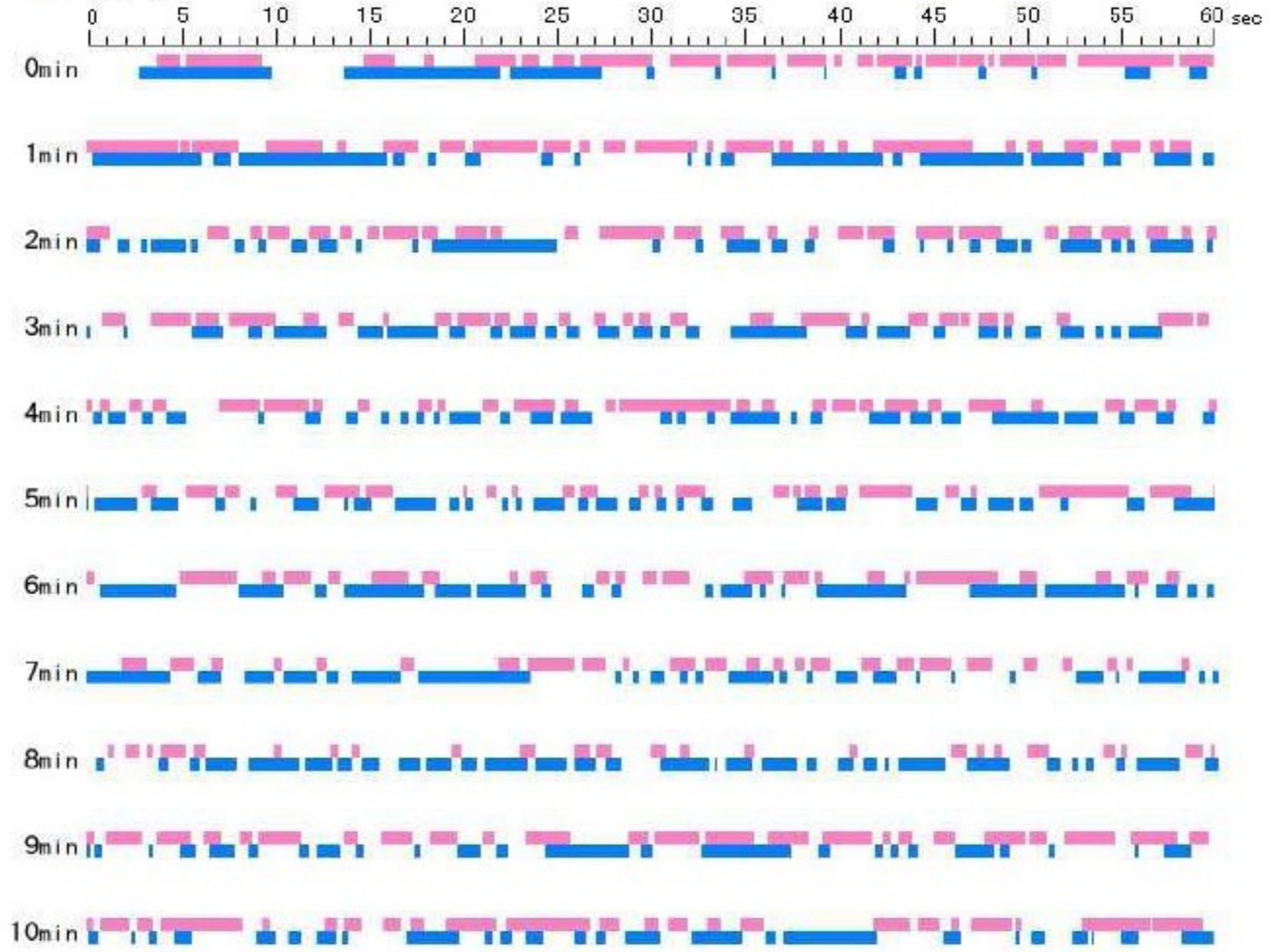
Common events facilitate simple comparisons

- these frequent simple sounds allow the listener to quickly estimate the affective states of the speaker
- they are simple and unobtrusive carriers of voice-quality and prosodic information
- interspersed regularly throughout the speech – i.e. not ‘ill-formed’, but ‘informed’!

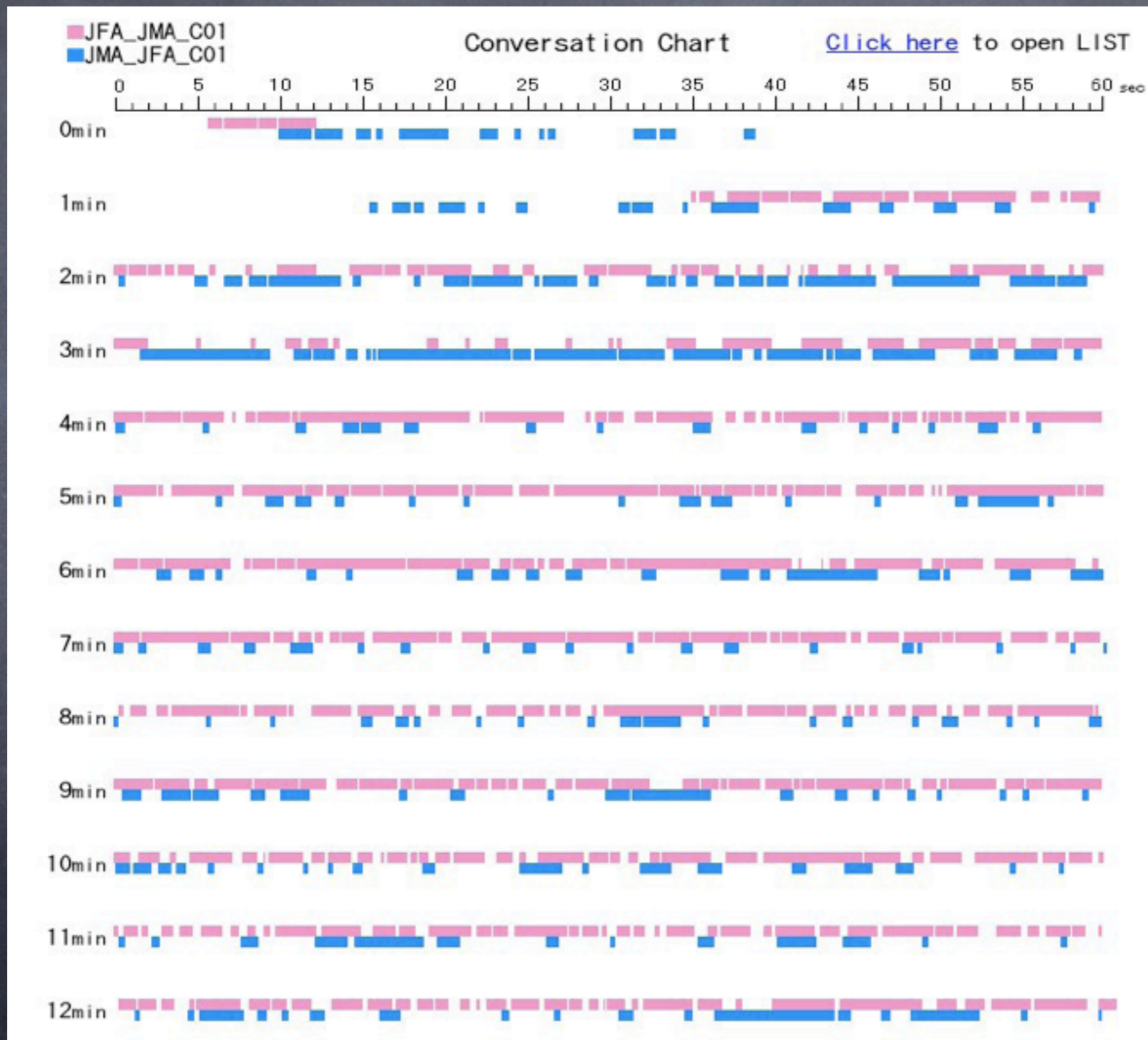
JMA_EFA_E06
EFA_JMA_E06

Conversation Chart

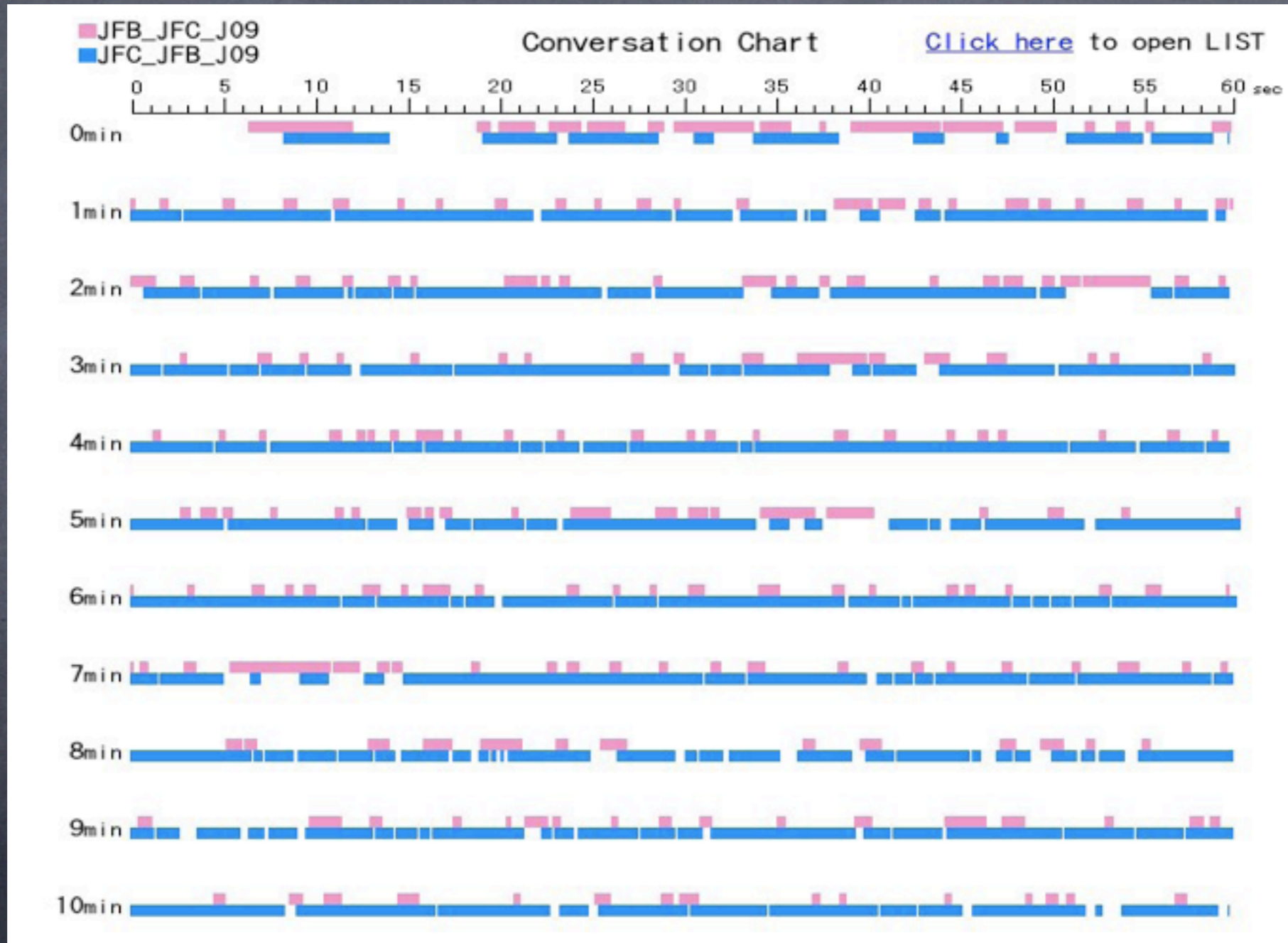
[Click here](#) to open LIST



Speech Activity in the first 13 minutes of the first conversation between Japanese female JFA (pink) and her male partner JMA (blue)

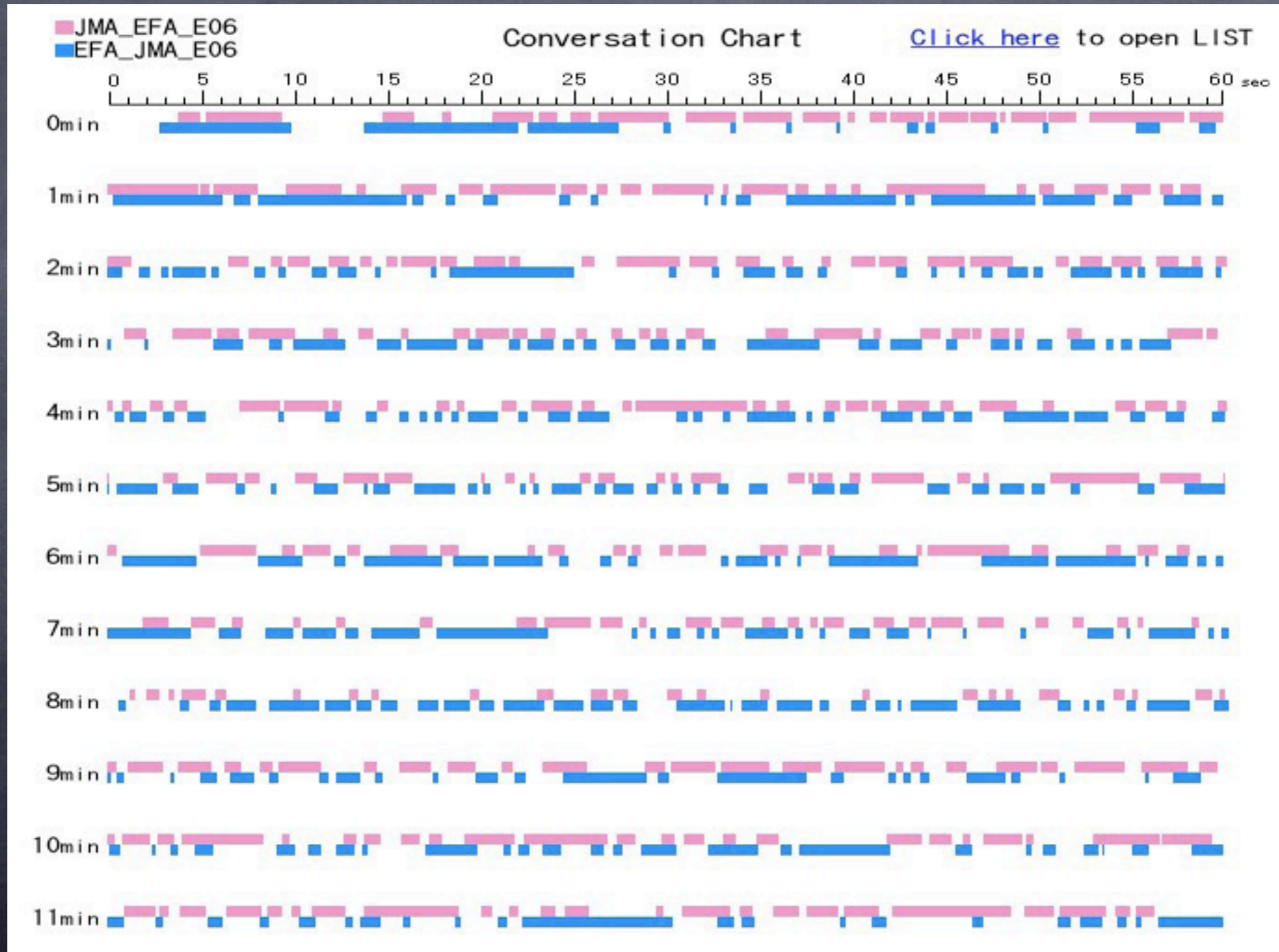


An extreme case : Speech Activity in the first 13 minutes of the last conversation between Japanese female JFB (pink) and her Japanese female partner JFC (blue)



Here, flow would be very high for blue, and very low for pink

Speech Activity in the first 13 minutes of the last conversation between Japanese male JMA (pink) and female English-native-speaker partner EFA (blue)



research issues for fastnet

- eliciting representative speech samples
 - (maybe a little whiskey will help here?)
- annotating relevant features of the discourse
- developing subtle sensor devices
- tracking movements & synchrony
- modelling voice quality & prosodies (Firthian)

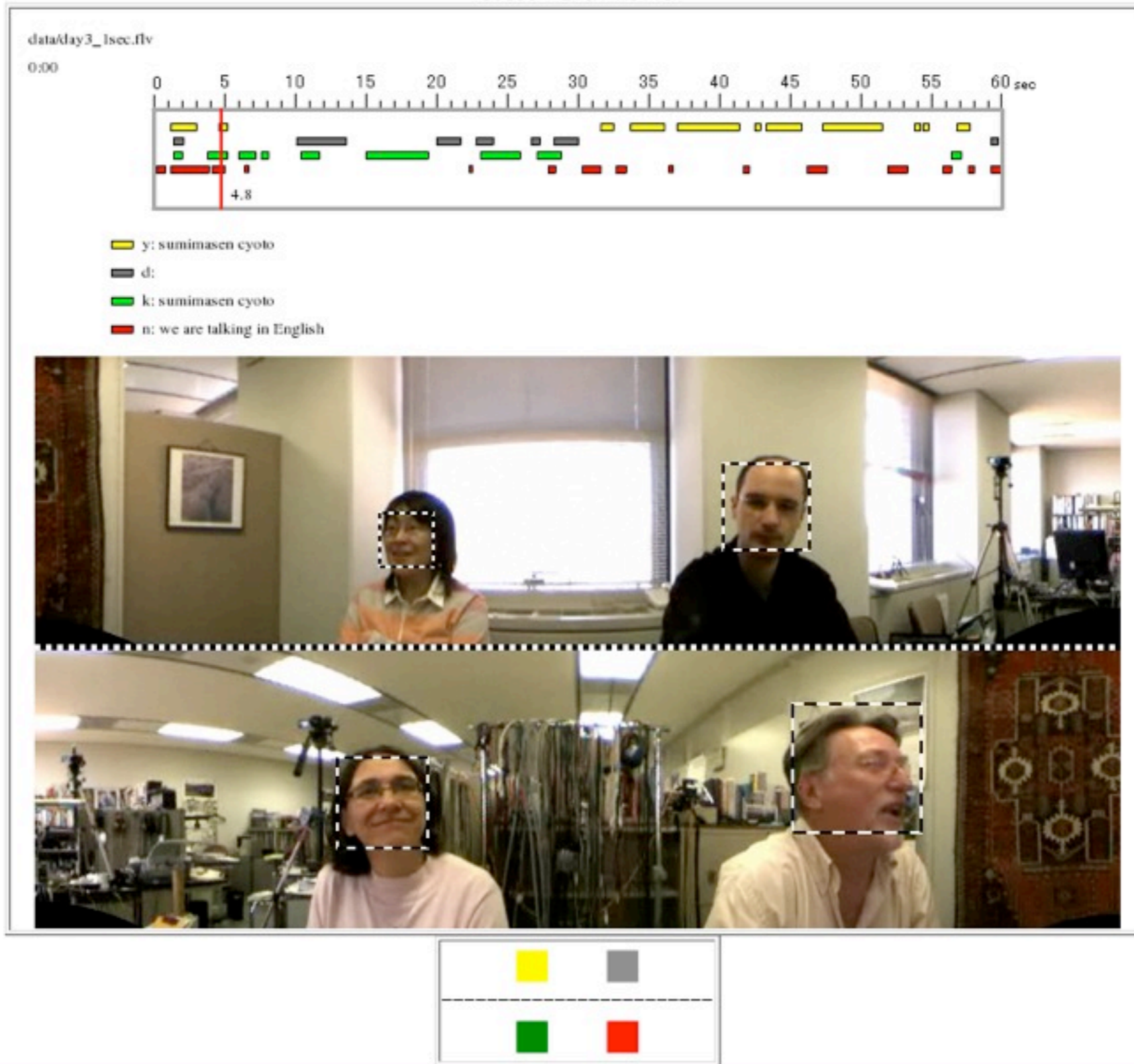
multimodal data

- processing large volumes of multimodal data
 - examples available on the Social Signal Processing website SSPNet and at <http://www.speech-data.jp> (with special thanks to Sadanobu sensei!)



interactive speech

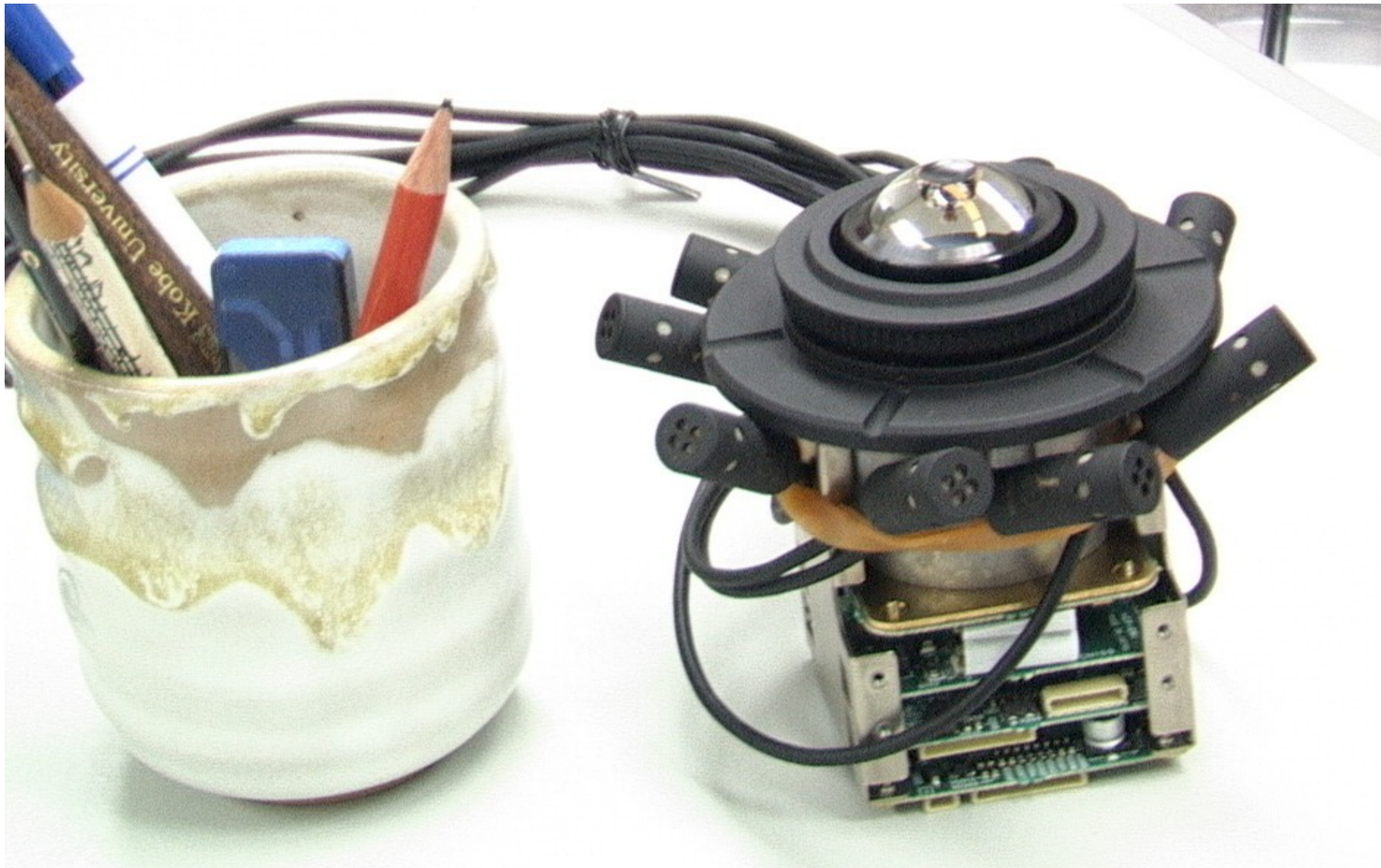
Conversation chart & face Video
60sec FLASH movie



A portable meetings recorder



SONY RPU-C251 (desktop version)





Mini-devices Real-time processing

a multimodal corpus


NOV 07 Data Summary

http://feast.atr.jp/non-verbal/project/html_files/tab/nov07/


BBC iPlayer NAIST Google Late Junction ich wrh Apple Google Maps YouTube Wikipedia News (287) Popular

MPG data with subtitles '07/11/05 - '07/11/07


DAY 1



DAY 2



DAY 3



Conversation Chart & LIST

data	length		CHART	Emotion CHART		LIST	TOPIC LIST	MPG data
DAY 1	34:35	4 persons with FACE trace	CHART	labeller A	labeller B	LIST	TOPIC	DOWNLOAD
DAY 2	01:22:15	5 persons	CHART	labeller A	labeller B	LIST	TOPIC	DOWNLOAD
DAY 3	01:22:45	4 persons	CHART	labeller A	labeller B	LIST	TOPIC	DOWNLOAD

1min flash movie with Head motion X graph

[Day 3](#) 4 persons with FACE trace

1min flash movie with Conversation Chart

[Day 3](#) 4 persons with FACE trace

[Video & Audio data LIST '07/11/05 - '07/11/07](#)

topic-level annotation

topic number	topic title	video file counter	what happened at the change of topic	who's mainly talking	who's listening/reacting	mood: heated/quiet
10	Manzai	0:00:00	Izumi starts talking about how different the said topic "Rakugo" and "Manzai" are.	Izumi	Christina interested and Damien explains his image of Manzai. Nick almost just listening.	interested and heated
20	"Oubeika"	0:01:23	Damien starts picking up from the former topic.	Damien explains the meaning of "Oubeika" how it is used. Izumi explains the technique of Manzai that makes the expression "Oubeika" effective in Manzai as well as its variation.	Damien actively asks questions and Christina gives her comprehension. Nick seems very interested.	interested and heated
30	Damien used "Oubeika" in kyoto	0:05:06	Damien explains what happened in Kyoto.	Damien said "Oubeika" on the street.	Everybody seems very interested and laughed a lot.	very funny
40	to get to know fragement of cultural background	0:06:44	Damien draws the topic to general understading.	Damien and Christina	Nick and Izumi mainly listen a bit quietly and later gives their understanding.	a bit quiet
		0:08:48	Izumi starts talking about	Izumi explains with the original action	Nick also knows a bit about him. Damian	heated but doesn't last long

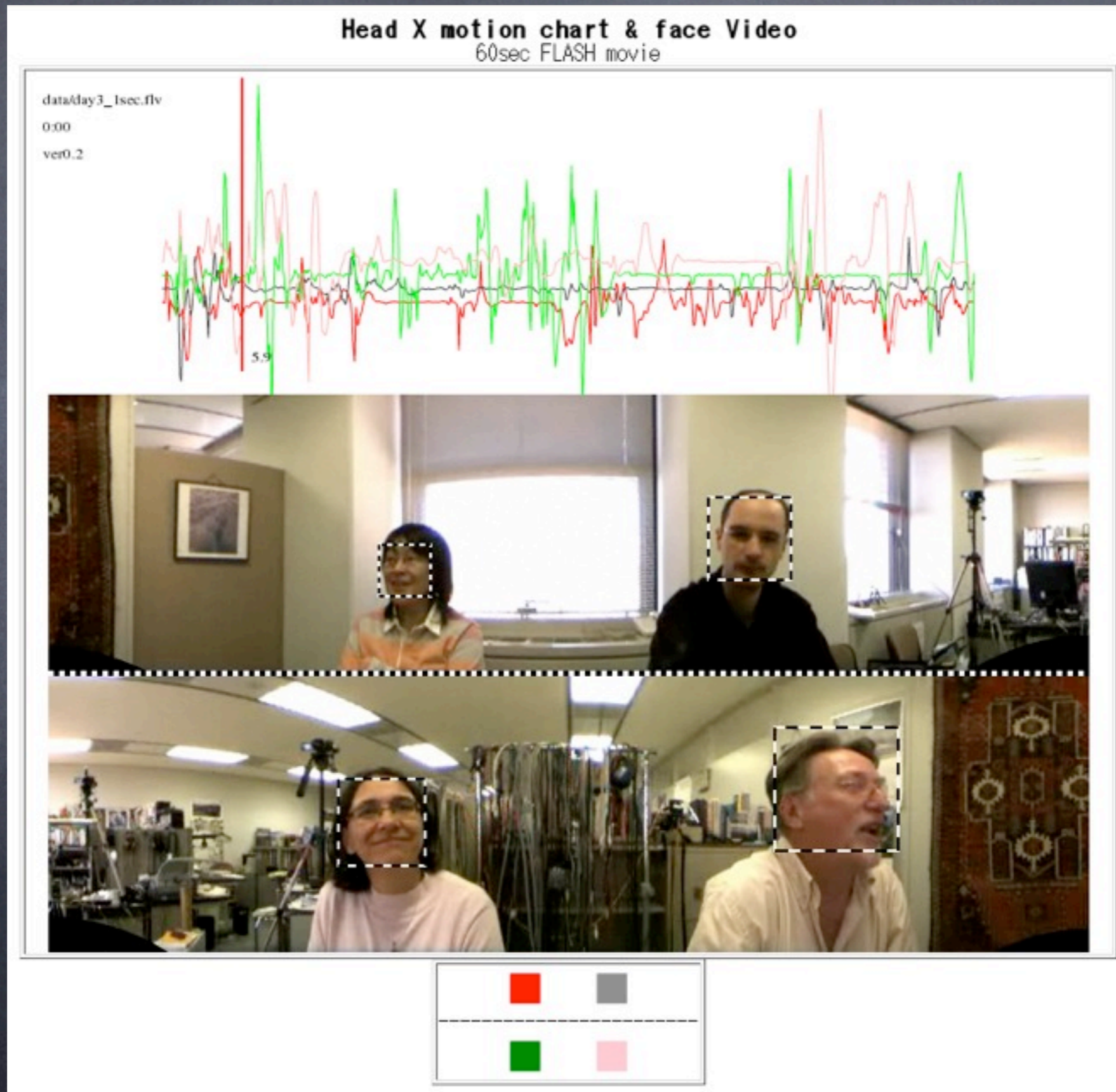
actions rather than words

participants actively engage in the discourse in an overlapping and complementary manner

our focus is on contributory and participatory discourse actions, rather than on the cognitive attention states of the listener.

These are physical observables that can easily be measured.

tracking head movements



body movement

- not surprisingly, there were significant positive correlations between their own head and body movement for all participants:

Correlation	P1	P2	P3	P4
head/body	0.797	0.809	0.808	0.722

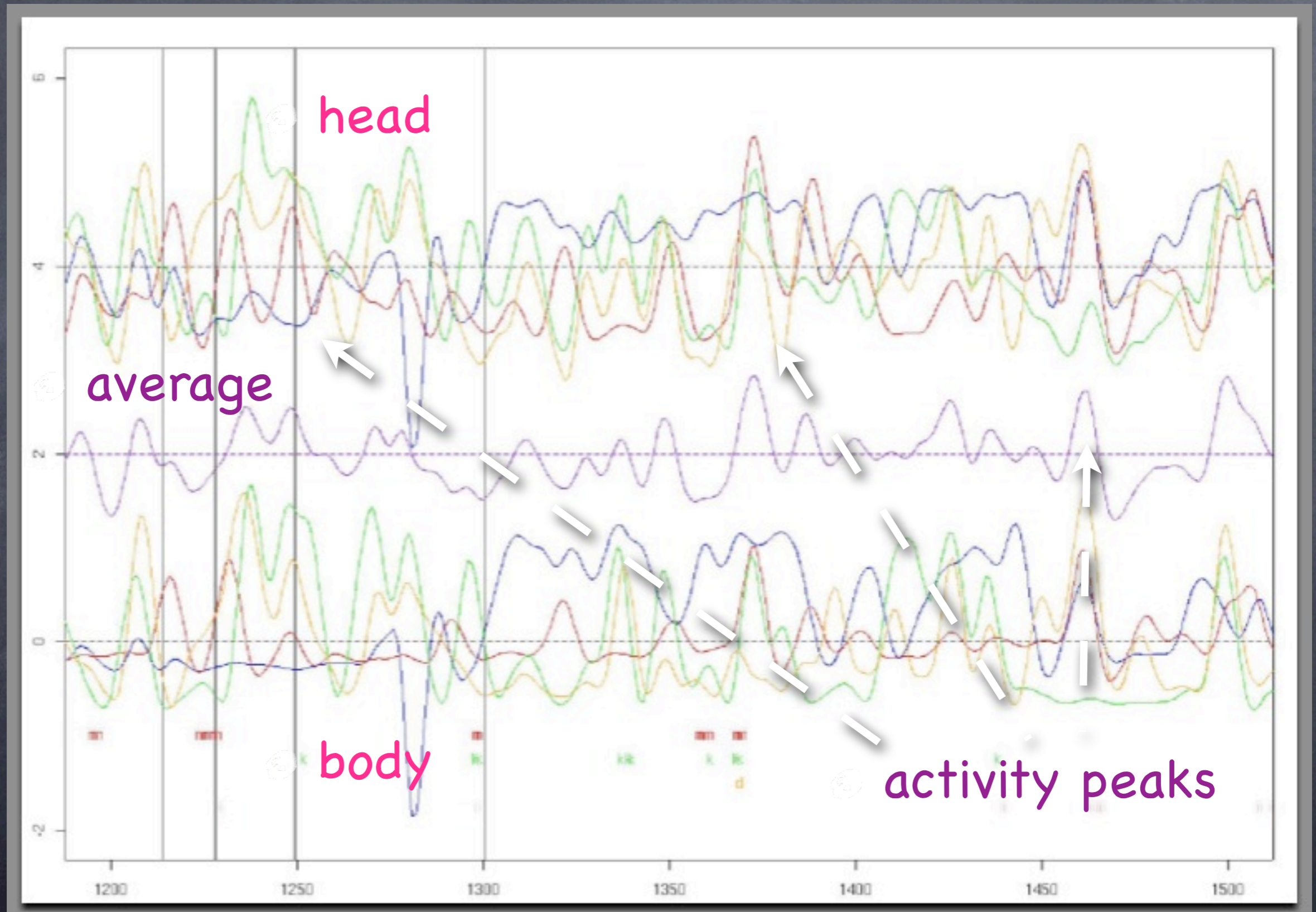
Table 1 Head - body correlations within speaker

body and head also synchronise between participants

Body	b1	b2	b3	b4
P1	-	0.289	0.082	0.436
P2	-	-	-0.308	-0.036
P3	-	-	-	0.408
Head	h1	h2	h3	h4
P1	-	0.53	0.233	0.239
P2	-	-	0.081	-0.204
P3	-	-	-	0.221

Table 2 Head and Body correlations between 4 partners. Note especially P1-b4 and P1-h2.

traces for 4 participants

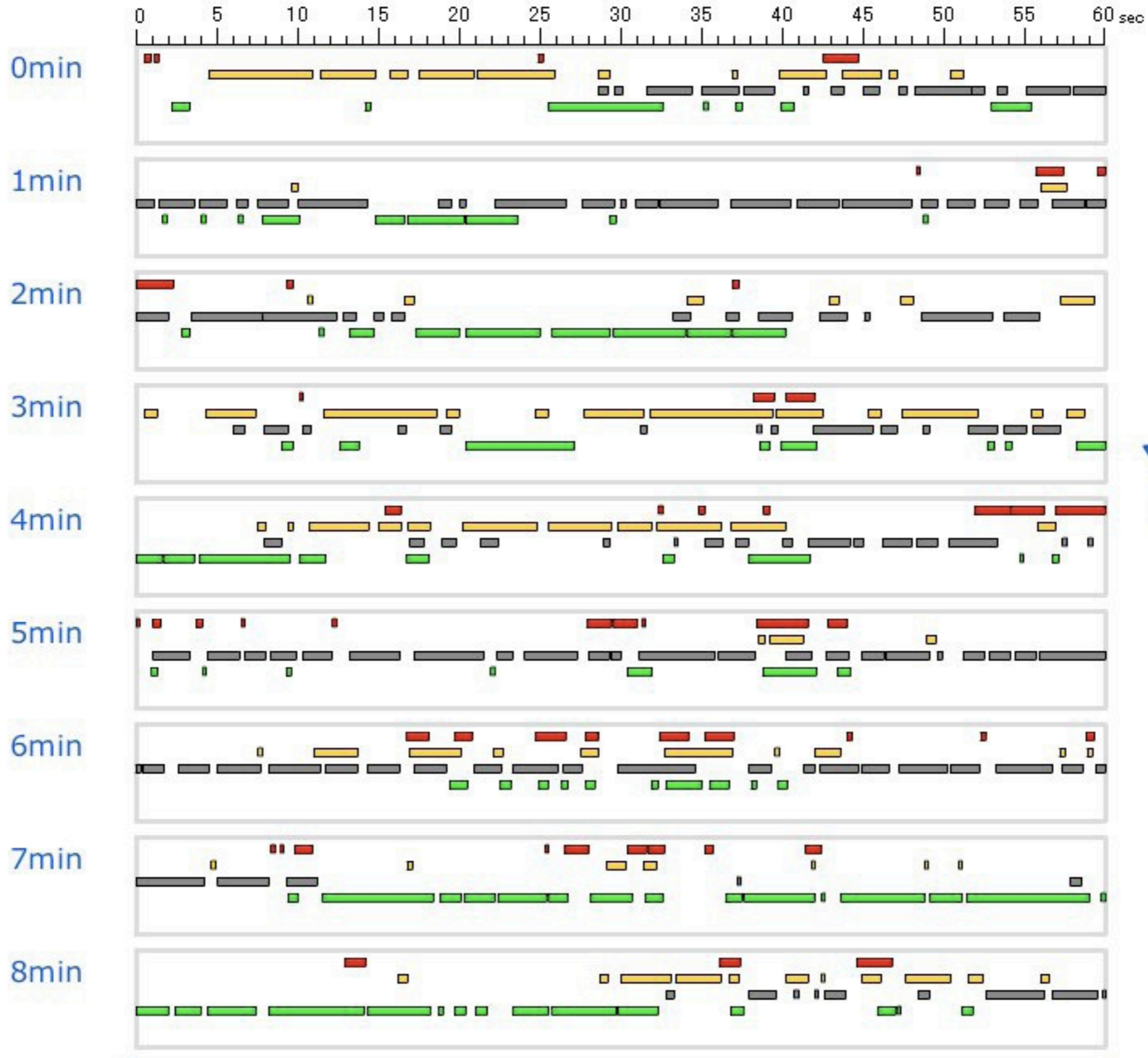


a different view of interaction modelling

- i.e., not processing discourse talk as content
- but the “dance” ... a socially evolving event
- multifaceted, multidimensional, and integrated
- loosely based around a framework of synchrony
- engagement, entrainment, mutual cooperation

ver 3.01 : day1_face.flv

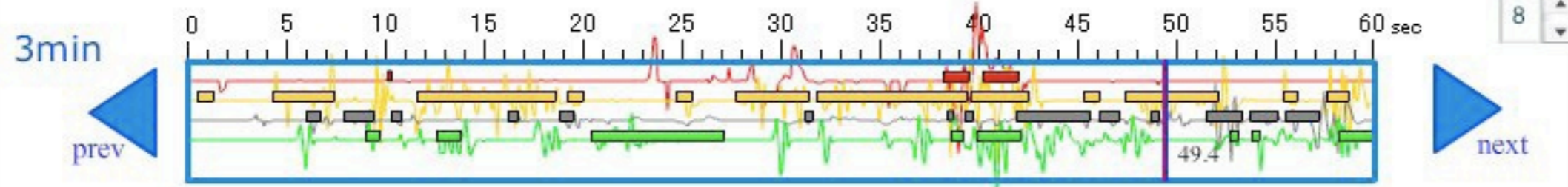
0:00



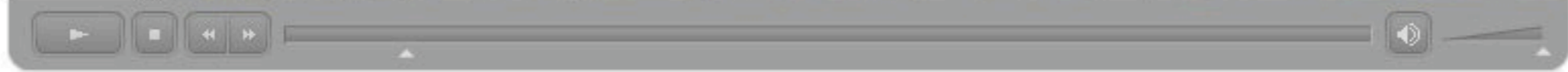
ver 3.01 : day1_face.flv

3:49

LIST All View



- n:
- y: "is that ""Oubeika"", "" Oubeika"", but they never mean ""Oubeika"""
- d:
- k:



- Human speech communication differs from written communication in several ways, the most important differences being in the use of intonation, speaking rate, and phonation style to indicate speaker states, attitudes, and intentions, both towards the listener and with respect to the discourse. It successfully integrates the two channels, linguistic and interpersonal, of speech information.
- Current speech technology, on the other hand, is still largely based on more formal styles of speech that are closer to the written mode than to interactive conversational speech, yet domestic users of the technology expect it to be able to respond to their normal modes of everyday conversational speech interaction

back to the future

- so what next?
- how are we to collect 'real' data?
- what aspects can we 'control'?
- what levels should we observe?
- what features should we measure?
-

perfect communication devices

- People are almost perfect communication devices. We have evolved language and speech, and later writing systems, to command and control, inform and entertain, and to generally socialise with each other
- Speech is perhaps the oldest form of human communication, yet speech is still only partly understood from the point of view of information technology
- There is a strong linguistic component in speech which is well-understood, but there is also a second channel – expressed by tone of voice and manner of speaking – which conveys very important but subtle information about the speaker, the discourse, and the hearer, that is still little understood

Studio techniques for eliciting natural variations in conversational speech

- A key innovative element of the research will be to develop methods that allow for the efficient collection of conversational speech data without the need for extensive recordings.
- This will require development of both capture devices (cameras and recorders) and capture environments (equivalent to a recording studio) that encourage participants to relax, interact informally, and maximise the range of speaking styles and formats.

evaluation

- the signal processing results will be tested in the context of speech synthesis (i.e., in close collaboration with the concurrent Irish synthesis project), for the provision of an interactive “chatty” style of speech as will be required for conversational interfaces.
- The prototype interface thus developed will also be evaluated through teaching use in Irish-language classrooms, but we envisage its incorporation into more sophisticated commercial applications, such as machine interpretation, robotics, and customer-services.
- Ongoing testing will go hand in hand with further development and refinement of the analysis/recognition and of synthesis modules.

Acknowledgements

Much of this work was carried out in Japan while I was employed by NiCT, the National Institute of Information and Communications Technology, and by ATR, the Advanced Telecommunications Research Institute in Kyoto.

It is being continued at Trinity College, Dublin, with funding from the Science Foundation Ireland (SFI), with partial support via NiCT through the Japanese Government 'kaken' funding.

Particular thanks to Tabatan for her skillful programming.

We are recruiting --- Please join us!

- why no handouts?
- a pdf of this talk ... and much of our data can be found at www.speech-data.jp
- help yourself! nick@tcd.ie

