"How to follow a conversation without listening to the words"

Nick CAMPBELL

National Institute for Information and Communications Technology,
nick@nict.go.jp

ATR Spoken Language Communication Research Laboratories
nick@atr.jp

Machine intelligence, ambient computing, sensitive devices, proactive machines

- "We need machines that better understand people, so that people can better understand machines."
- "Sensitive machines can be made pro-active, not needing to wait for instructions but being ready with an appropriate act."

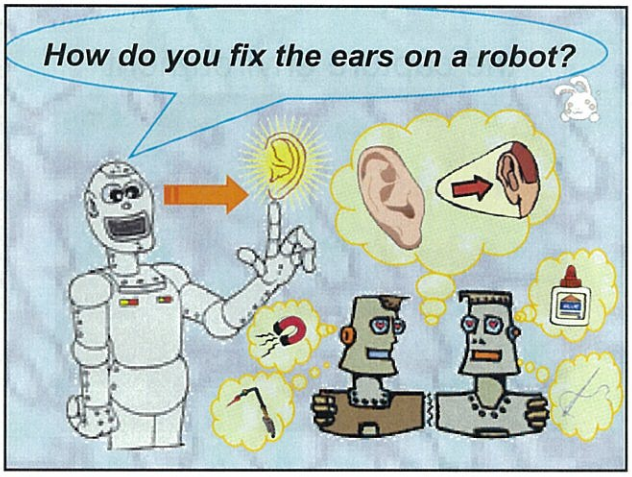
ヒューマンコミュニケーションの「場」が読めるロボットの研究開発 (041307003)

The "Robot's Ears" Project

研究代表者
ニック・キャンベル
国際電気通信基礎技術研究所 音声言語コミュニケーション研究所 音声音響処理室学研究室 主幹研究員
Nick CAMPBELL
ATR Spoken Language Communication Research Laboratories

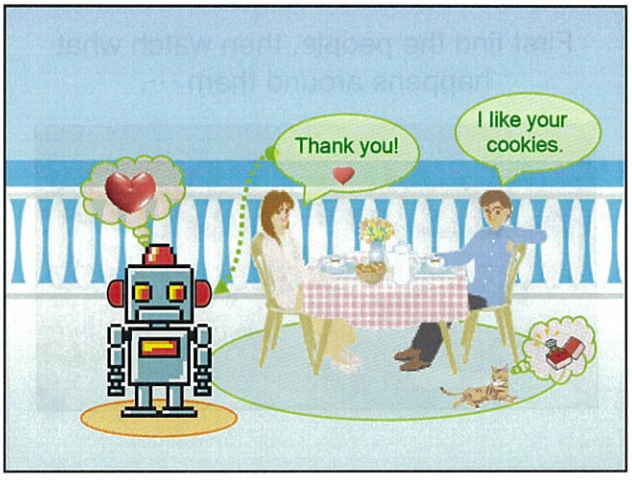
研究分担者
定延 利之†, 千原 國宏††, 井村 誠孝††
Sadanobu Toshiyuki, Chihara Kunihiro, Imura Masataka,
ダミアン ドゥシャン††, 岩橋 直人†††, 中川 明子†††
Damien Douxchamps, Iwahashi Naoto, Nakagawa Akiko
†神戸大学 国際文化学部 国際文化学科 情報コミュニケーション論講座
††奈良先端科学技術大学院大学 情報科学研究科 情報処理学専攻 像情報処理学講座
†††国際電気通信基礎技術研究所 音声言語コミュニケーション研究所 音声音響処理室学研究室

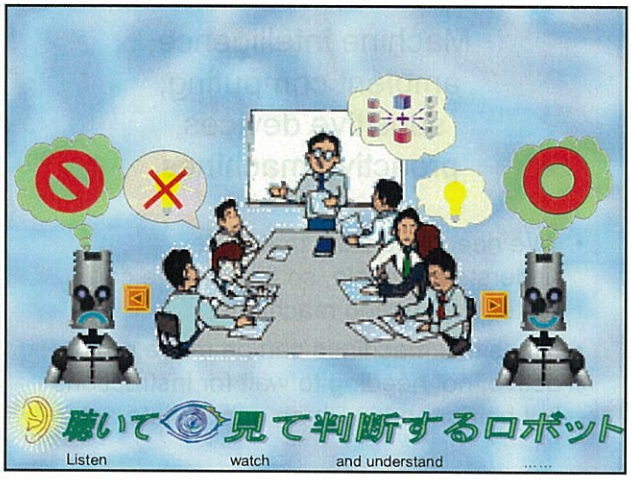
研究期間 平成16年度～平成18年度



You give them better eyes!

- By watching how people interact
- And listening to how they speak
- But not bothering much with what they say!





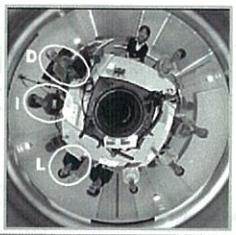
Following Meetings Data

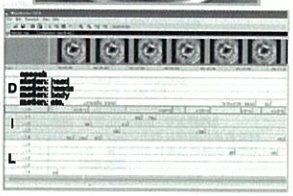
- 2-10 people sitting round a table
  - upper-body movement only
- who is participating how (to what degree)
  - talking / listening / thinking / waiting to speak
- basic primitives only
  - no speech recognition
  - no eye tracking
    (simple unobtrusive devices)

the capture environment

Annotation:

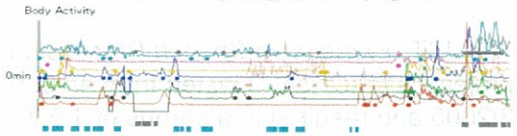
- speech
  - on/off
  - long/short
- motion
  - head
  - body
  - hands



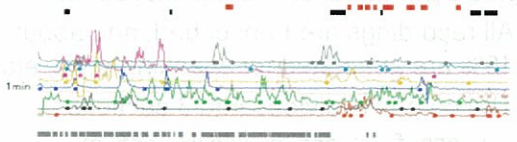


First find the people, then watch what happens around them ….

Tracking activity

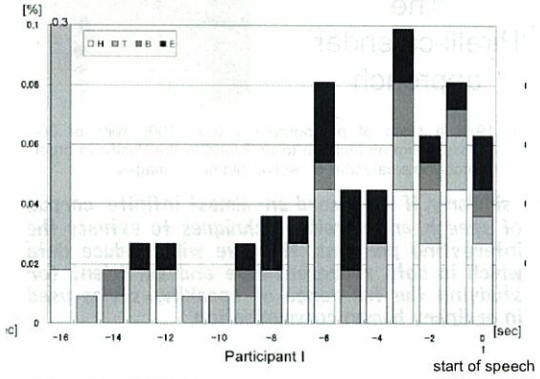


This is the output of the 360-degree camera that does the people-spotting

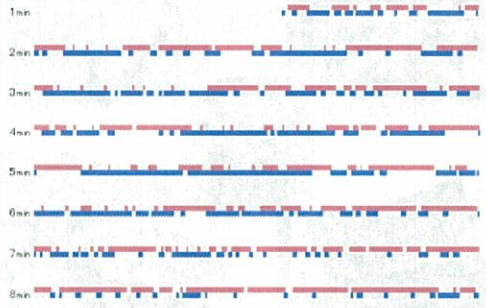


It tracks and measures movements in the head area, and body activity below

Movement before start of speech



Short and long utterances



Tracking speech

10073	うん	467	スーー	228	ううん	134	へー,ーー
9692	※S	455	スー	227	えっ	134	はいはいはいはい
8607	はい	450	んーー	226	へーー	134	そう,です
4216	laugh	446	うーーーん	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	はーい	393	はいはいはい	199	はーー	129	はー
1573	うーーん	387	あー,はい	193	ハー	129	い
1348	スー	372	ねん	192	その	127	はー
1139	ふん	369	ふーーん	190	え,えー	125	ハハハハハ
1098	あのー	369	だから	188	あ,あーー	119	はいはい
1084	あっ	368	あー,ん	187	ね	119	はーーー
981	はあい	366	ああ	180	ん,はい	114	ハハ
942	あの	345	あのーー	180	あのーーー	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	あーー	305	スーー	164	う,うーん	110	フフフ
701	あ	274	うん,うん,うん	161	はーー	110	その
630	あーーー	266	ハハハハ	160	※K	110	もう
613	あ,はい	266	て一	159	そうです,ねー	109	ふーーーん
592	うん,うん	266	え,ーー	151	あーーーー	108	はあ,ーー
555	あー	258	で	143	だから,ー	106	そうですねえ
500	えー	248	う	139	アハハハハ	105	えーん
469	え	242	へー	137	そう,そう,そう	104	いや

「表現豊かな発話音声のコンピュータ処理システム」

"Expressive Speech Processing"

the JST CREST-ESP Project

科学技術振興事業団報　第131号：
「高度メディア社会の生活情報技術」

The JST/CREST ESP project

Learning how people talk …

- by listening to LOTS of examples
- recorded in everyday conversations
- over a period of 5 years
- all transcribed and annotated
- and acoustically mapped

The 'Pirelli-calendar' approach

in 1970 a team of photographers took 1000 rolls of 36-exposure film on location to an island in the Pacific in order to produce a calendar of twelve (glamour) images.

-> ***similarly, if we record an 'almost infinite' corpus of speech, and develop techniques to extract the interesting portions, then we will produce data which is both representative and sufficient for studying the full range of speaking-styles used in ordinary human communication.***

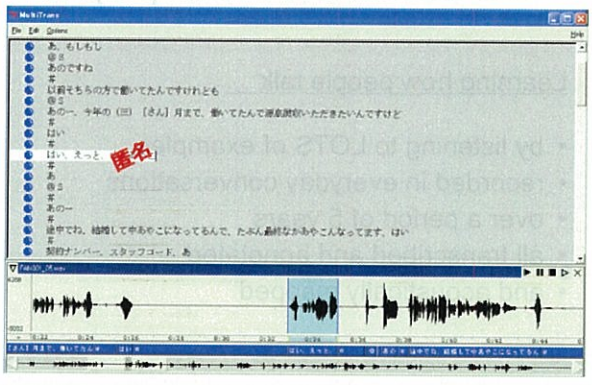
The JST/CREST 'ESP' corpus

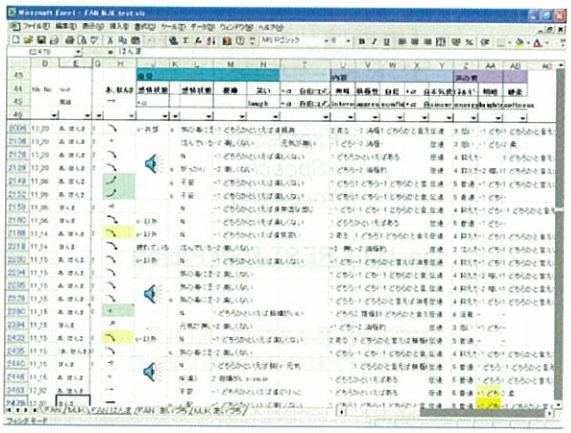
- The JST *"Expressive Speech Processing"* project *(ATR/CREST)* lasted from 4/2000 to 3/2005 and resulted in a corpus of 1,500 hours of natural conversational speech
- All recordings are transcribed, and about 10% are annotated for speaking-style, etc.
- The corpus is divided into 3 sections :
  i: ***esp_f***, ii: ***esp_c***, and iii: ***esp_m***

Sections of the ESP corpus

- **esp_f**
  - one female speaker, head-mounted mic, 600 hours of daily spoken interactions, emotion/speech-act/etc ...
- **esp_c**
  - 10 adult speakers, 5m 5f, 2 chinese, 2 english,
  - 30-minute telephone conversations x 10 weeks
  - all conversations in japanese, free content
- **esp_m**
  - multi-speaker, head-mounted microphones, variety of interaction settings (like esp_f but many more voices)

Transcription

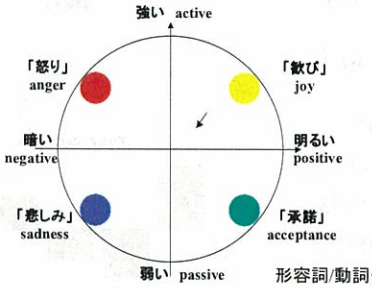




What's the difference?

- The words are (almost) the same
  - Yet the whole 'feeling' is different .....
- Shows speaker - listener relationships
- Transcription of the speech may be the same - but these small differences are very important!
- And 'GRUNTS' are very common!

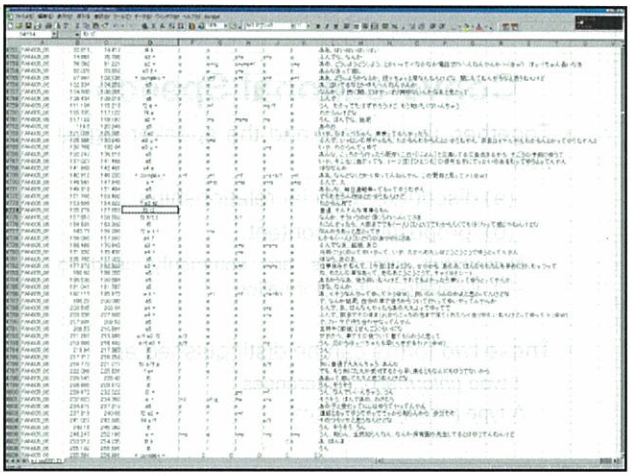
"Emotion" Labelling



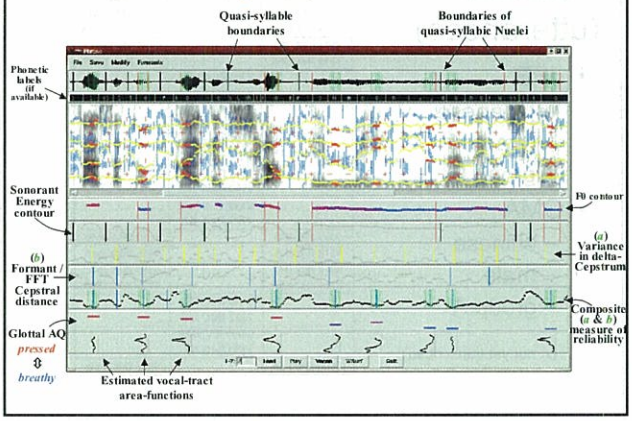
形容詞/動詞一語で
ふつう？

Discourse Act Labelling

a	あいさつ	greeting	o	反論	argue
b	会話終了	closing	p	提案、申し出	suggest, offer
c	自己紹介	introduce-self	q	気づき	notice
d	話題紹介	introduce-topic	s	つなぎ	connector
e	情報提供	give-information	r	依頼、命令	request-action
f	意見、希望	give-opinion	t	文句	complain
g	応答肯定	affirm	u	褒める	flatter
h	応答否定	negate	w	独り言	talking-to-self
i	受け入れ	accept	x	言い詰まり	disfluency
j	拒絶	reject	y	演技	acting
k	了解、理解、納得	acknowledge	z	繰り返し	repeat
l	割り込み, 相づち	interject	r*	要求	request (a~z)
m	感謝	thank	v*	確認を与える	verify (a~z)
n	謝罪	apologize	*?	よく分からない場合	



Acoustic Analysis / Visualisation tool



An example of how speaking-style can change

- 13,604 conversational utterances
- 1 female Japanese speaker (age 32-35)
- listener/speech-act/emotion labels

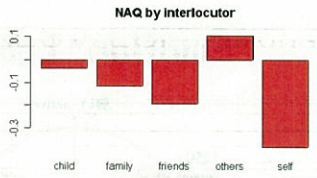
- Interlocutor:

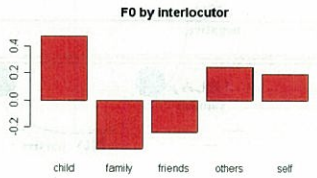
Child	Family	Friends	Others	Self
139	3623	9044	632	116

Voice quality

Talking to:

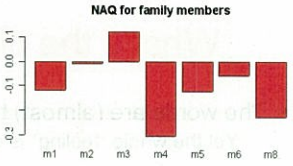
- child
- family
- friends
- others
- self

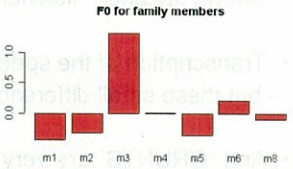




NAQ & F0 by family

m1 - mother
m2 - father
m3 - baby girl
m4 - husband
m5 - big sister
m6 - nephew
m8 - aunt





Conversational Speech

- Together, the listener and the speaker signal both
  - (a) discourse & social relationships
  - (b) propositional content.
  - Non-verbal utterances are commonly used to signal paralinguistic information.
- These two forms can be distinguished as
  - I-type *(information)* utterances
  - A-type *(affect)* utterances

esp_c: paired conversations

- 10 speakers, 5 male, 5 female
  - 2 Chinese, 2 English-native-speakers
  - all conversations in Japanese
- 3 groups (A,B, and C)
  - A – talking with foreigners
  - B – talking with strangers
  - C – talking with family members
  - all talking with each other … for 10 sessions

groupings for natural dialogues

```
female          male

( cfa efa    cma ema )      (foreign)
  /              \           Group A
 jfa      -      jma
  |               |
 jfb             jmb         Group B
  |               |
 jfc      -      jmc
  |               |           Group C
(fam)           (fam)       (intimate)
```

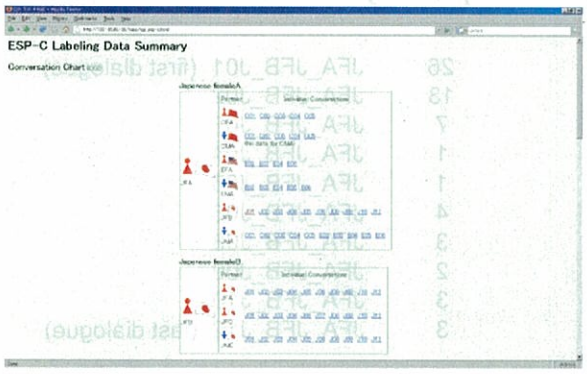
counts (utterances)

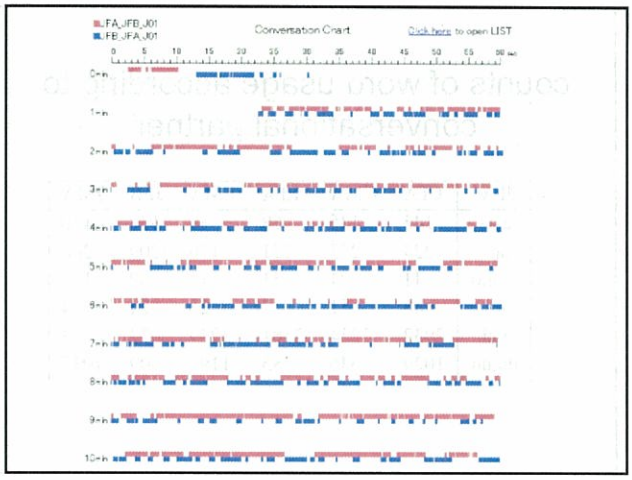
- JFB
  - japanese
  - female
  - group B
- CMA
  - chinese
  - male
  - group A

(x: not yet transcribed)

both female	mixed	both male
6425 EFA JFA 7359 JFA EFA		9348 EMA JMA 7433 JMA EMA
8827 CFA JFA 9145 JFA CFA		x (cma jma) 7530 JMA CMA
	9236 EFA JMA 8499 JMA EFA	
	7557 JMA CFA x (cfa-jma)	
	8237 JFA CMA x (cma-jfa)	
	8416 JFA EMA 8560 EMA JFA	
	10068 JFA JMA 7701 JMA JFA	
9069 JFA JFB 9378 JFB JFA		8614 JMA JMB 9465 JMB JMA
8044 JFB JFC 8234 JFC JFB		6983 JMB JMC 7735 JMC JMB
	7686 JFB JMC 7222 JMC JFB	
	10005 JFC JMB 7980 JMB JFC	
13900 JFC Fam		9961 JMC Fam

the top-level web pages





the structure of spoken language

- one of the main goals of this work is to model the structure of spoken dialogues
- to understand why they appear 'broken'
- to model the different types of information (not just linguistic) that are carried by the frequent (non-verbal) sounds and by differences in the structure of speech

a sample conversation

the raw data

- transcribed speech
- start & end times
- filenames also show speaker, partner, & conversation number
- one utterance / line
- noises and non-speech sounds also transcribed

```
CFA JFA C01 200.369 0.491 #
CFA JFA C01 200.860 0.808 laugh
CFA JFA C01 201.668 0.869 あと.は
CFA JFA C01 202.537 1.099 変わり.まし.た
CFA JFA C01 203.636 1.868 laugh
CFA JFA C01 205.504 0.670 うん
CFA JFA C01 206.174 0.744 #
CFA JFA C01 206.918 0.917 はい
CFA JFA C01 207.835 2.691 #
CFA JFA C01 210.526 0.602 はい
CFA JFA C01 211.128 2.791 #
CFA JFA C01 213.919 0.749 #S
CFA JFA C01 214.668 2.685 そう.です.結構.し
う.西.有閑.なり.ました
CFA JFA C01 217.353 0.785 はい
CFA JFA C01 218.138 0.561 #
CFA JFA C01 218.699 0.731 はい
CFA JFA C01 219.430 1.384 #
CFA JFA C01 220.814 1.088 行って.ます
CFA JFA C01 221.902 0.738 #
CFA JFA C01 222.640 0.784 はい
CFA JFA C01 223.424 1.107 #
CFA JFA C01 224.531 1.356 あの.一.歳.です
CFA JFA C01 225.887 0.525 #
CFA JFA C01 226.412 0.600 はい
CFA JFA C01 227.012 2.795 #
CFA JFA C01 229.807 0.443 はい
CFA JFA C01 230.250 0.941 #
```

the hundred most common utterances

10073	うん	467	スー	228	ううん	134	へー.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	へーー	134	そう.です
4216	laugh	446	うーーーん	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	はーい	393	はいはいはい	199	ほーー	129	はー
1573	うーーん	387	あーはい	193	ハー	129	い
1348	スー	372	ねえ	192	その	127	ほー
1139	ふん	369	ふーーん	190	え.えー	125	ハハハハハ
1098	あのー	369	だから	188	あ.あーー	119	はいはい
1084	あっ	368	あーん	187	ね	119	はーーー
981	はあい	366	ああ	180	え.はい	114	ハハ
942	あの	345	あの.ーー	180	あのーーー	113	は
941	ふーん	337	なんか	173	え.え	113	て.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	あーー	305	スーー	164	う.うーん	110	フフ
701	あ	274	うん.うん.うん	161	はーー	110	そのー
630	あーーー	266	ハハハハ	160	@K	110	もう
613	あ.はい	266	で.ー	159	そう.です.ねー	109	ふーーーん
592	うん.うん	266	え.ーー	151	あーーーー	108	はあ.ーー
555	あー	258	で	143	だから.ー	106	そうですわえ
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

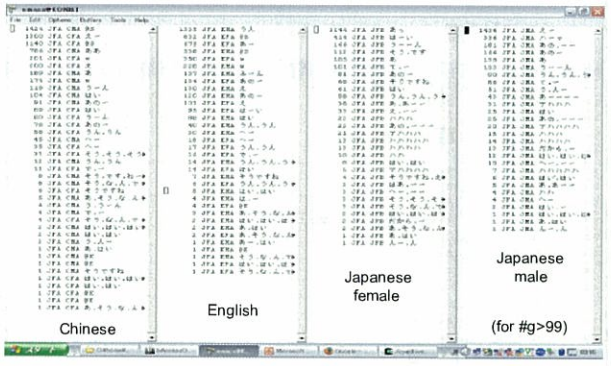
counts of word usage according to conversational partner

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 (first dialogue)
13	JFA_JFB_J02
7	JFA_JFB_J03
1	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 (last dialogue)

Frequency of 'grunts' per partner

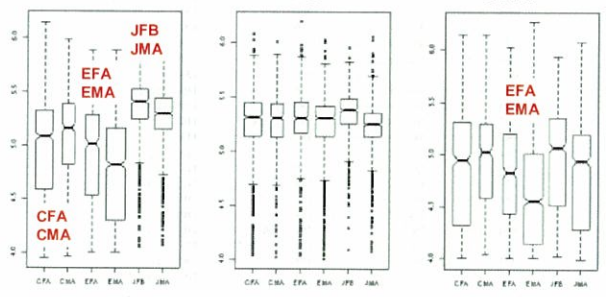


some acoustic measures for JFA

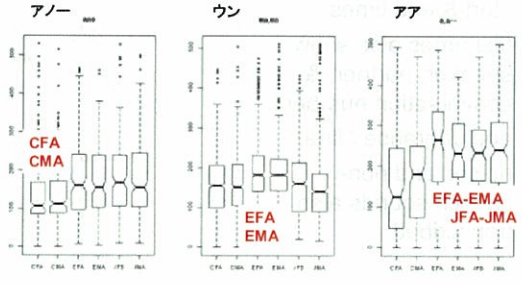
"a,a–"	CFA	CMA	EFA	EMA	JFB	JMA
f0r	125	181	266	232	234	241
f0m	201	214	220	192	206	198
pwr	28	29	29	28	31	31
pwm	38	39	36	35	42	41
"un,un"	CFA	CMA	EFA	EMA	JFB	JMA
f0r	154	152	182	181	161	141
f0m	172	175	162	145	198	174
pwr	28	29	27	26	29	27
pwm	37	40	36	35	42	39
"ano"	CFA	CMA	EFA	EMA	JFB	JMA
f0r	106	113	161	154	169	155
f0m	131	136	142	133	156	149
pwr	27	28	28	27	31	29
pwm	38	40	37	36	42	39

mean f0 values for JFA's うんうん

start mid end



pitch range variation in 3 utterances from speaker JFA differs according to interlocutor



utterance frequencies for JMA

a 296	a- 368	a– 693	a— 608	a.a- 390	a-.hai 386
a.hai 577	a-.n 368	ano 337	ano- 494	a!! 927	demo 272
e- 665	e– 254	ee 2679	fun 642	fu-n 625	fu–n 273
ha.ai 978	hai 7295	ha-i 1657	hai.hai.hai 378	n(umm) 265	n- 456
n– 410	nanka 273	ne- 367	nee 284	@S 3382	sou 810
su- 429	su– 296	un 3717	u-n 2401	u–n 1243	u—n 333
un.un 351	@W 3041	zu- 1348	zu– 467		

F0 varies according to partner …

mean f0 max f0 min f0

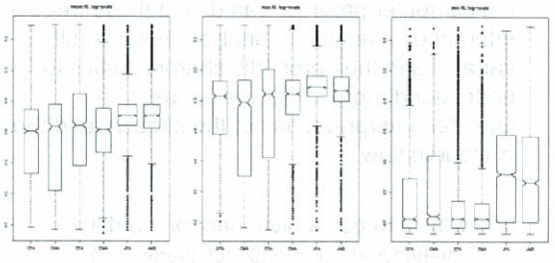
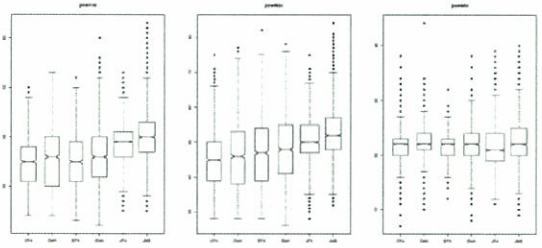


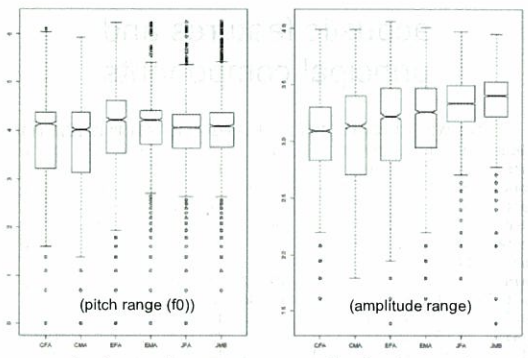
Figure 1: Plots of mean, maximum and minimum f0 values observed in the data of each of the interlocutors. The box-plots show median and interquartile values, with whiskers extending to 1.5 times the interquartile range. All F0 measurements are converted to their log values for ease of comparison

as does signal amplitude (energy)

mean max min

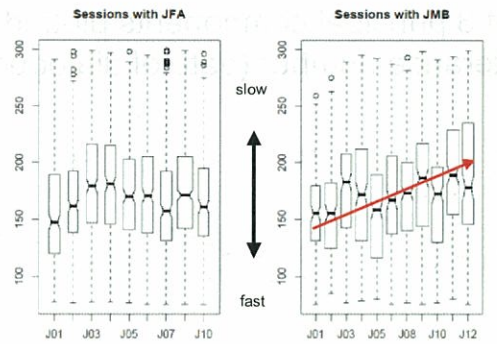


ure 2: Plots of mean, maximum and minimum rms amplitude (speech signal power) values for each of the interlocutors



… and also the ranges of pitch and power

Figure 3: Ranges of fundamental frequency and power measurements for the affective utterances, plotted by interlocutor. As above, C,E,J stand for Chinese, English, Japanese respectively, and F,M represent female and male interlocutors. Both f0 and power are plotted as log values.



speech rate also varies <u>across time</u>

Figure 4: Speaking rate changes over weekly sessions

talk's focus : **frequent utterances**

- utterances that are frequently repeated can be used as *comparison points* …
- small differences in voice quality and other prosodic parameters can carry significant information about the speaker state
- i.e., *even if we do not know the speaker well*, we can use common speech sounds as an indicator of changes in that persons <u>affective state(s)</u> and <u>discourse intentions</u>

Detecting frequent utterances

- if a computer program can detect these frequent utterances, then it can analyse these small variations in their prosodic characteristics to understand more about the speaker, the intended interpretations of the speech, and the discourse flow
- one way is to use a dictionary or word list
  (one-hundred words >50% of utterances!)

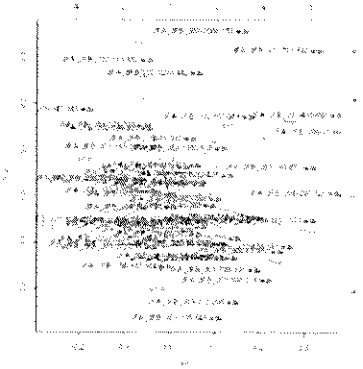
discussion

- Prosody varies according to the nature of relationships between speaker and partner
- By tracking frequent repeats, we can detect small changes in these relationships
- Small changes can be easily detected by comparison between frequent similar segments that are close in time

acoustic features and principal components

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14
fmean	-50	14	25	-20	11	-3	-4	-7	8	3	14	27	71	0
fmax	-48	10	23	0	9	1	30	-26	18	18	36	-37	-46	0
fmin	-29	11	17	-46	12	-26	-27	15	7	-14	-62	-9	-26	0
fpct	-6	20	-13	-22	-40	46	-62	-25	11	18	12	-9	-2	0
fvcd	-8	-23	-39	-29	-7	6	36	-61	-15	9	-28	29	-7	0
pmean	-36	-26	-31	29	-16	-7	-7	-7	-17	-28	-16	-59	30	0
pmax	-43	-12	0	42	-5	2	-27	1	-17	-30	8	56	-34	0
pmin	-20	-26	-37	7	-12	-34	-10	32	29	64	2	11	-1	0
ppct	-16	16	-15	-44	-51	30	45	46	23	-27	-6	8	-5	0
hlh2	-13	-30	32	4	6	53	8	25	-44	41	-26	-5	1	0
hla3	9	-50	34	-12	-28	-9	-4	-9	20	-12	7	1	1	-67
hl	5	-57	11	-14	14	19	-7	-1	43	-21	8	0	0	60
a3	-8	0	-39	-1	63	40	-5	12	27	-10	0	-1	-1	-44
dn	5	18	22	54	-6	12	7	-27	49	14	-51	5	9	0

Principal Components Analysis



PCA in R (r-project statistical software):
prcomp(un, retx=T, center=T, scale.=T)

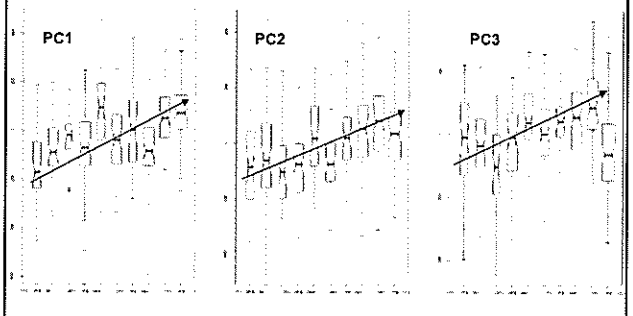
3 components: 50%, 7 components: 80%

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	1.65	1.53	1.38	1.32	1.12	0.96	0.89	0.83
Proportion of Variance	0.19	0.16	0.13	0.12	0.08	0.06	0.05	0.04
Cumulative Proportion	0.19	0.36	0.49	0.62	0.71	0.78	0.83	0.88

	PC9	PC10	PC11	PC12	PC13	PC14
Standard deviation	0.74	0.71	0.61	0.29	0.23	0.0004
Proportion of Variance	0.03	0.03	0.026	0.006	0.004	0.0001
Cumulative Proportion	0.92	0.96	0.98	0.99	1.00	1.00

First 3 principal components plotted by utterance number (date of dialogue)



conclusion

- this paper has presented an initial analysis of the jst/crest esp_c corpus
- which is balanced according to speaker familiarity, sex, and ease of interaction
- showing that speakers modify their speaking habits according (perhaps?) to familiarity with the interlocutor
- and that the "ill-formed" nature of speech serves to carry important social information

… more conclusion

- … and has presented a simple program for analysing the transcriptions of the conversations to
  - differentiate between frequent "wrappers" and their linguistic "fillers"
  - to provide a measure of the differences in "social prosody" in conversational interactions
- this is work in progress …
  - so I hope to benefit from your helpful comments and suggestions

some initial findings : acoustic differences

- people change their speaking styles when talking with different people
  - this is plain common sense (nothing new!) but nice to be able to measure/quantify
- people also vary significantly their pitch ranges and voice quality settings, even for the same utterance

Latest devices

SONY RPU-C251 (desktop version)

thank you for listening