APPROACHES TO CONVERSATIONAL SPEECH RHYTHM: SPEECH ACTIVITY IN TWO-PERSON TELEPHONE DIALOGES

Nick Campbell

National Institute of Information and Communications Technology
& ATR Spoken Language Communication Research Labs
Keihanna Science City, Kyoto 619-0288, Japan
nick@nict.go.jp, nick@atr.jp

ABSTRACT

This paper examines speech activity patterns in telephone dialogues and illustrates some details of their timing organisation. It is shown that partners participate actively, even when listening, through frequent use of speech overlaps and backchannel utterances.

Keywords: Spontaneous speech, telephone conversations, speech overlaps, dialogue structure

1. INTRODUCTION

One of the basic and frequently repeated assumptions of conversation analysis is that people talk in turns, and that usually only one person talks at a time [1]. Levinson has defined conversational speech as “a kind of talk in which two or more participants freely alternate in speaking” [2]. This alternation accords well with our sense of the rhythm of a conversation, but it is not well supported by a quantitative analysis of a large number of telephone conversations where people were paid “just to talk” to each other [3]. Goffman [4] differentiates unfocussed interaction, where participants are simply concerned with the “management of sheer and mere copresence”, from focussed interaction where persons “openly cooperate to sustain a single focus of attention”. The present study examines data wherein people pass the time by chatting with each other about a variety of topics. Timing details derived from time-aligned transcriptions of these recordings reveal considerable overlapping speech in the discourse. The paper illustrates the structure of these patterns in the interactive conversations.

2. DATA

One hundred thirty-minute telephone conversations were recorded over a period of several months, with paid volunteers coming to an office building in a large city in Western Japan once a week to talk with specific partners in a separate part of the same building over an office telephone. While talking, they wore a head-mounted Sennheiser HMD-410 close-talking dynamic microphone and recorded their speech directly to DAT (digital audio tape) at a sampling rate of 48kHz. They did not see their partners or socialise with them outside of the recording sessions. Partner combinations were controlled for sex, age, and familiarity, and all recordings were transcribed and time-aligned for subsequent analysis.

In all, ten people took part as speakers in these recordings, five male and five female. Six were Japanese, two Chinese, and two native speakers of American or Australian English. All conversations were held in Japanese. The non-native speakers were living and working in Japan, competent in Japanese, but not at a level approaching native-speaker fluency. Partners were initially strangers to each other, but became friends over the period of the recordings. There were no constraints on the content of the conversations other than that they should occupy the full thirty-minute time slot. Recordings continued for a maximum of ten sessions between each pair, or five for the non-native speakers.

3. ANALYSIS

The speech data were transferred to a computer and transcribed manually to provide a time-aligned record of what was spoken when, by who and to whom. A computer program was written to align the conversations and to calculate the amount of time each person spent silent or talking during the 30-minute sessions. Four classes of activity were distinguished: both partners silent, both talking at the same time, and one or the other partner talking while the other was silent, presumably listening. These numbers were stored in a file which also recorded for each utterance the length in milliseconds of that utterance, the duration of the pause preceding it, the duration of the previous utterance, and the duration of the pause preceding that. Similar durations were stored for the conversation partner, to facilitate a prediction of the pause length, or time of utterance initiation, relative to the previous utterances by both speaker and partner.

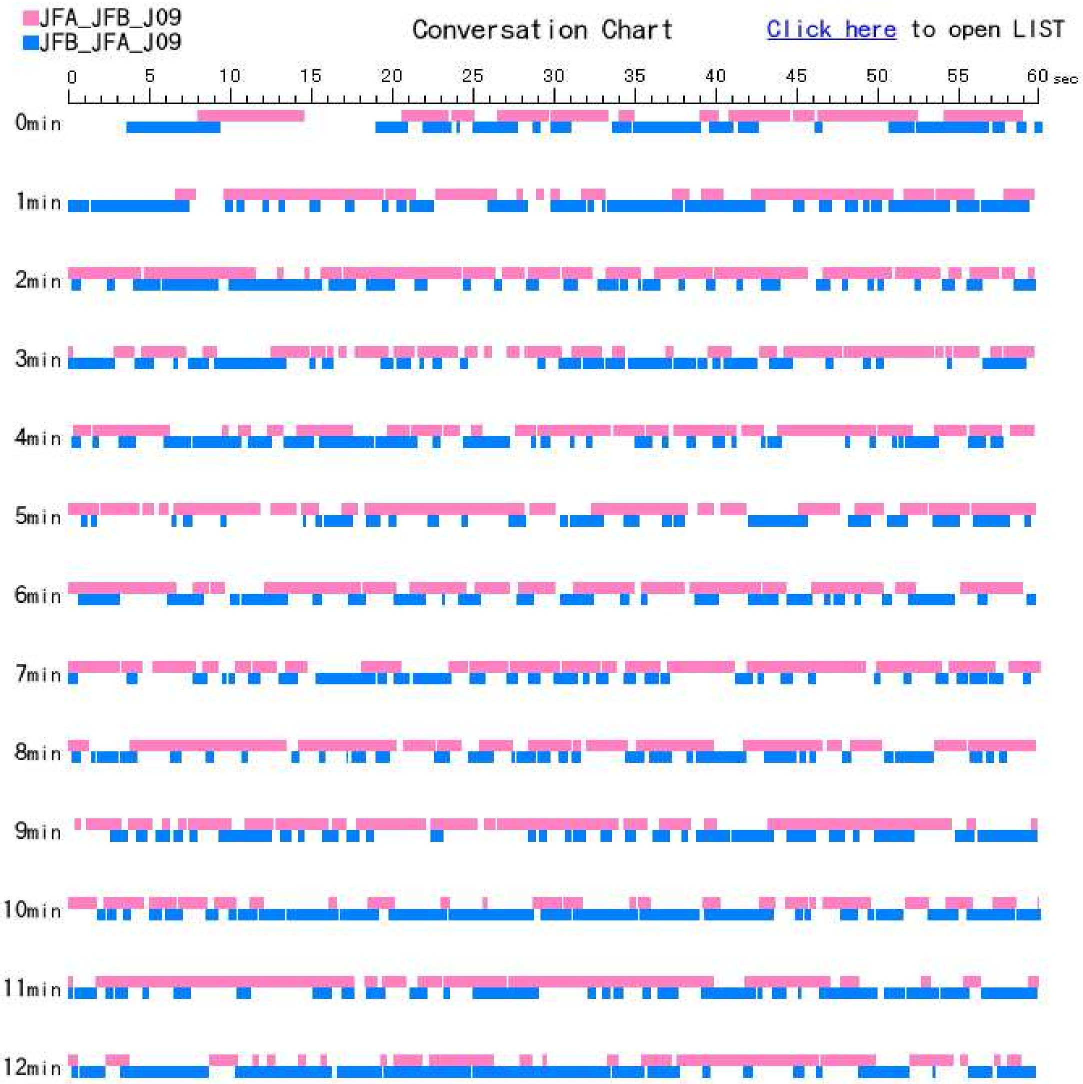

Figure 1: A speech activity plot for the first thirteen minutes of the ninth conversation between Japanese female partners JFA and JFB. Note how much overlap takes place, and how the relative dominance gradually shifts from one speaker to the other throughout this typical conversation fragment.

The definition of “an utterance” in conversational speech is difficult. A common practice is to use e.g., any pause in the speech of greater than 200 milliseconds as an objective delimiting boundary, but we noticed that even many single words contained pauses of more than 300 milliseconds in these conversational data. We therefore proposed to our transcribers a “one-yen-per-line” principle, whereby they would increase their payment for more lines produced by cutting the speech into shorter utterances, but would be penalised for breaking up a single utterance into too small or “unnatural” units. This resulted in most of the segmentation being performed at the level of the ‘phrase’ or minor intonation unit, i.e., a word or group of words demarcated by a single intonation contour, but in many cases the transcribers actually produced longer units, including comma punctuation, because of uncertainty about whether a clearly distinguishing intonational break could be heard or not. The hundred conversations provided 98,698 utterances of between one and fifty syllables in length. 25% of these utterances were less than 500 milliseconds and another 25% longer than 1.5 seconds, with the longest being 11.5 seconds. Median duration of all utterances was 0.9 seconds. Figure 1 shows an example sequence.

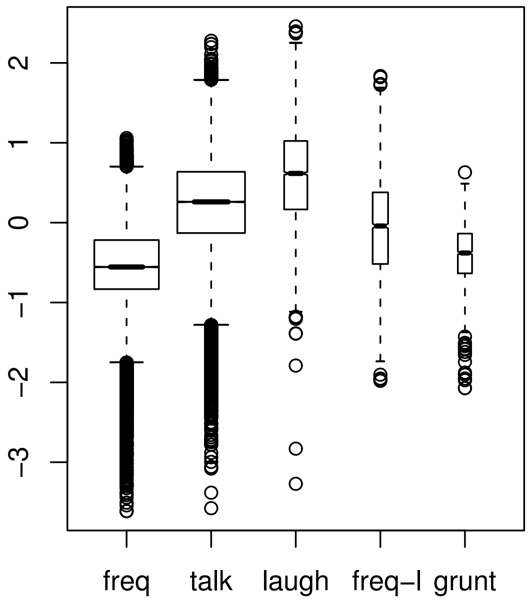

Figure 2: Showing the log duration distributions for five classes of utterance. frequent (n>25), talk, long laugh, frequent (short) laugh, and grunt,

The definition of a "turn" in conversational speech can be even more problematic. As Figure 1 shows, there is no clear on/off switching between talk and silence as might be found if both speakers were sharing a half-duplex channel using a 'walkie-talkie' for example, nor is it clear exactly when the dominance has shifted from one speaker to the other. There are clear periods when one partner dominates, but the listener during these periods is far from passive.

With the definition of an utterance given above, several utterances can be combined to form a speaker turn. In this implementation, the program counted each, incrementing if another utterance from the same speaker followed, but resetting the counters whenever the conversation partner started speaking, storing the number of uninterrupted utterances as a parameter in the data table, independently of the duration of any gap between them. Also stored was a variable indicating whether or not the partner was speaking at the time of onset of the speaker's new utterance. Table 1 gives details of (a) the number of utterances, and (b) the number of utterances per turn thus derived for the six native-speakers of Japanese in the corpus. By the above criteria, it is clear that by far the majority of turns consist of a single utterance.

3.1. Patterns of Talk and Silence

Table 1: Showing the number of utterances (top) and the number of utterances per turn (bottom) for the six Japanese speakers of the corpus. JFB and JMB spoke only to Japanese partners and so took part in more conversations.

JFA	JFB	JFC	JMA	JMB	JMC
15,543	21,624	13,038	13,122	20,841	11,530

	turns	1	2	3	4+
JFA	9,386	5,349	2,384	932	721
JFB	12,534	6,724	3,068	1,254	1,488
JFC	8,509	5,394	2,013	694	408
JMA	9,492	6,868	1,807	531	286
JMB	13,408	8,428	3,049	1,117	814
JMC	7,440	4,711	1,718	595	416

Table 2: Showing distributions of utterance types factored according to whether the partner is silent at the onset of speech (top) or still talking (bottom). A clear tendency can be seen for shorter turns (fewer utterances) when the partner is talking. The numbers in the first column show number of utterances in each turn.

	freq	talk	laugh	freq-l	grunt
1	9,988	11,550	1,065	378	285
2	6,773	9,698	900	373	304
3	3,120	3,908	413	179	158
4	1,246	1,514	150	93	81
5	1,063	865	123	77	69
	freq	talk	laugh	freq-l	grunt
1	16,590	15,588	2,298	1082	954
2	1,966	1,731	307	133	119
3	226	184	49	14	14
4	27	25	4	1	1
5	6	4	2	0	0

It is beyond the scope of the present paper to enter into a detailed analysis of the linguistic content of each utterance, but for simplicity the transcribed utterances were classified into 5 types: (i) frequent utterances, i.e., speech patterns which appeared more that 25 times each in the transcriptions, (ii) infrequent utterances ('talk') appearing less than 25 times, assumed to be more propositional than phatic in content, (iii) laughs, which were subdivided into longer more expressive variants and (iv) shorter more common simple laughs of up to three syllables, and (v) other non-speech noises (grunts) such as sniffs, sharp intake of breath, or coughs which might be used for discoursal purpose.

Figure 2 plots the durations of these utterance types in log(seconds), and Table 2 shows the distributions of these according to number of utterances in the turn. The table shows a clear difference between distributions for solo speech (in the top part) as against overlapping speech (in the lower part).

3.2. Patterns of Overlapping Speech

In this section we examine in more detail the patterns of overlapping speech as illustrated in Figure 1 for JFA, one female speaker from the corpus. The figure illustrates the on/off organisation of her ninth conversation with a female partner JFB, when they have already become quite familiar, and it is clear from the figure that considerable overlap is taking place throughout the conversation.

Table 3 shows summary durations in minutes for overlapping speech, solo speech, silence, and total talking time for speaker JFA and her various partners averaged across thirty conversations. The table differentiates between solo talking, when only one partner is active, overlapping speech, when both are simultaneously active, and silence. Talk time shown is the sum of solo and overlapping speech times. Silence is similarly split between times when one partner is talking (and the other is presumably listening) and those when neither is active.

Table 5 provides exact details of this speaker's speech activity timing patterns per conversation with a range of different interlocutors, both male and female, and foreign and native. The table details 5 conversations each with two non-Japanese partners and 10 each with two Japanese partners, and provides counts in minutes showing how much time was spent in each state by the various partners throughout the series of conversations. Similar data has been produced for all speakers of the corpus and is summarised in Table 4.

Table 5 shows for example that JFA is a more prolific speaker than her partners and that she speaks less with foreigners than with the native-speaker partners, though she warms to the Chinese female towards the end of their series. On average she spends 22.7 minutes (sd=2.11) talking during each 30-minute session. Her partners spend on average 17.5 minutes talking with her (sd=2.6). These times sum to more than the total time of each conversation. There is on average 8.6 minutes (sd=2.3) of overlapping speech, and an average of 14.2 minutes of solo speech for JFA with an average of 8.9 minutes of solo speech per partner. Rounding to whole minutes, we find not only that her partners spend the same amount of time in overlapping speech as they do in solo speech but also that she spends more than 60% of her talking time in overlapped speech. Table 4 confirms that she is not exceptional. Across the whole range of quartiles for similar data for all speakers, comparing solo talking time to overlapping talking time reveals that all partners spend more than half of their talking time speaking while the other is also speaking.

Table 3: Showing mean durations in minutes for overlapping speech ('ovlp'), solo speech, silence, and total talking time for speaker JFA (A) and her various partners (B) for the 30 conversations whose timing details are presented in Table 4.

	JMA	JFB	CFA	EMA
ovlp	8.641	10.932	7.158	5.12
soloA	14.949	12.304	15.6	15.006
soloB	8.247	8.968	8.638	10.308
silA	10.967	10.803	11.06	13.962
silB	17.675	14.138	18.002	18.652
talkA	23.59	23.236	22.758	20.13
talkB	16.888	19.901	15.796	15.428
silent	2.728	1.84	2.422	3.66

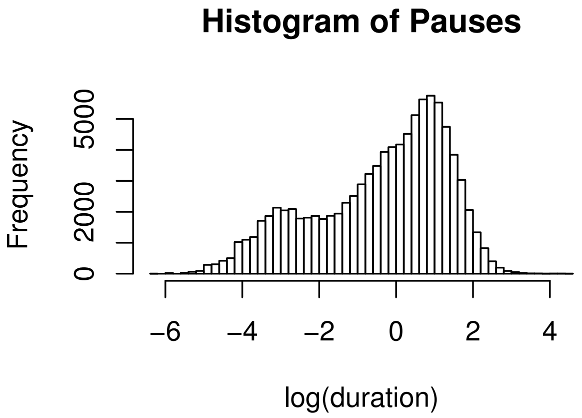

Figure 3: Showing a clear bimodal tendency in the durations of pauses preceding each utterance.

3.3. Patterns of Pauses

This section briefly examines the nature of the silent portions of the conversations. Figure 3 shows a histogram of pause durations preceding each utterance as averaged across the whole of the corpus. Durations are as usual plotted in the log domain because that transform better represents their distributional characteristics, with typically many short tokens and fewer tokens of ever-increasing length.

The figure clearly shows there to be two classes of pause duration, long and short, with peaks at 2.7 seconds (exp(1)) and 0.05 seconds (exp(-3)) respectively Whereas the shorter peak may be an artifact of the segmentation criteria for utterances, examination of Figure 1 (and other similar plots) supports the differentiation between long and short pauses.

Figure 4 plots pause durations by utterance type, showing a small but significant effect with lengthening perhaps being due to the backchannel nature of the more frequent utterances.

Table 4: Showing quantiles summarising speech activity durations for all one-hundred conversations in the corpus. Silence ('sil') is recorded when neither partner is speaking, overlap ('ovlp') when both are speaking at the same time. 'Sil' shows the time each speaker individually (A or B) was quiet, presumably listening. 'Solo' shows the total duration of non-overlapping speech per speaker (A or B), and 'talk' the total overall speech time including overlaps. 'Total' shows timing statistics for the entire conversation (assumed to be 30 minutes by default). All times are shown in minutes. Data are calculated from the time-aligned transcriptions of 100 30-minute conversations

	min	25%	50%	75%	max
sil	0.99	2.08	2.85	3.81	7.03
silA	6.73	10.68	14.02	16.91	22.46
silB	5.72	13.09	14.68	17.68	21.58
soloA	4.14	9.51	11.66	14.68	18.17
soloB	4.55	8.39	10.64	13.32	18.90
ovlp	2.66	5.53	7.01	9.04	12.80
talkA	10.80	16.04	18.75	22.44	28.52
talkB	12.20	15.66	17.93	20.15	27.15
total	28.57	32.00	32.93	33.96	37.98

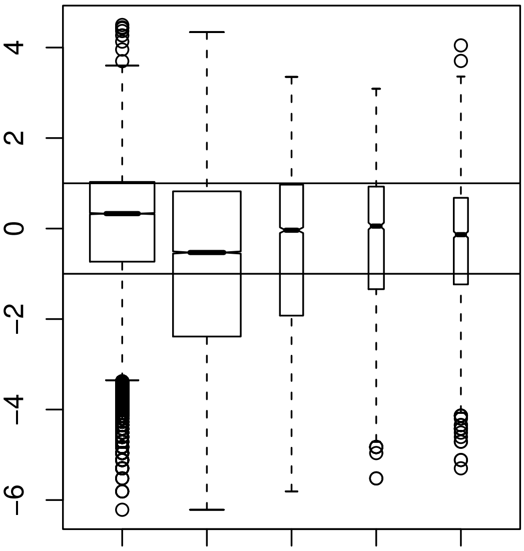

Figure 4: Durations of the pauses before each type of utterance. They are shorter before infrequent utterances and longer before the most frequent.

4. DISCUSSION

Like Jefferson [5], we were interested in predicting the pause durations in this corpus as part of a model of conversational speech rhythm. Given such a large body of data where every statistical difference is significant, it should be simple to detect regularities that allow, for example, prediction of pause durations for a computer-based dialogue interface to enable the machine to appear more intelligent to the human interlocutor, but such was not to be the case. Again like Jefferson [6, 7], we encountered several failed hypotheses and were unable to predict the patterns of overlapping segments from the features of the data that have been presented above. "Active listening" is the best term we have come across to describe this phenomenon, and we are currently exploring models of this joint speaker interaction.

A naturally interactive dialogue is not like a tennis match, where there is only one ball that can only be in one half of the court at any given time. Rather it is like a volley of balls being thrown in several directions at once. The speaker does not usually wait silently while the listener parses and reacts to an utterance; there is a constant exchange of speech and gesture, resulting in a gradual process of mutual understanding wherein a 'meeting of the minds' can take place [8].

5. CONCLUSION

This paper has presented some results of an analysis of a large body of conversational speech recordings. It has shown that contrary to naive assumptions of dialogue as a tennis-like exchange of question and answer or topic and comment, it actually presents a complex pattern of simultaneous talking as partners take turns to dominate in the interaction. There appear to be no clear boundaries between one turn and the next, and the shift from backchannel feedback to conversational dominance appears to be more subtle. Future work might employ more sophisticated text analysis as a further contributing factor.

Acknowledgement

This work is partly supported by the National Institute of Information and Communications Technology (NiCT), and includes contributions from the Japan Science & Technology Corporation (JST), and the Ministry of Public Management, Home Affairs, Posts and Telecommunications, Japan (SCOPE). The author is especially grateful to the management of Spoken Language Communication Research Labs at ATR for continuing encouragement and support.

6. REFERENCES

[1] Sacks, H., Schegloff, E., A., and Jefferson, G.,. A simplest systematics for the organization of turn taking for conversation. In Schenkein 1978

[2] Levinson, S., C., *Pragmatics*, Cambridge University Press, 1983.

Table 5: Timing details for different types of speech activity during 4 sets of telephone conversations, 30 in all, each lasting approximately 30-minutes. JFA and JFB are Japanese females, and JMA a Japanese male, EMA is an Australian male, CFA is a Chinese female. All conversations were held in Japanese. Columns show conversation number, duration of overlapping speech ("ovlp"), then for each speaker the duration of silence, total-talk, and solo-talk durations respectively. The final two columns show silence duration and total conversation duration. Times are shown in minutes. It is of interest to compare solo talking time to overlapping speech times; in many cases the amount of overlap is about half of the solo talking time, and for the two Japanese ladies JFA and JFB their overlapping speech appears to be often as long as and sometimes even longer than their solo speech.

conv	ovlp	A	silA	talkA	soloA	B	silB	talkB	soloB	sil	total
E01	5.44	JFA	13.89	19.51	14.07	EMA	17.26	16.13	10.69	3.20	33.40
E02	5.44	JFA	13.13	20.86	15.42	EMA	18.64	15.39	9.95	3.21	34.03
E03	4.98	JFA	14.49	19.67	14.68	EMA	18.70	15.38	10.40	4.09	34.16
E04	4.70	JFA	14.05	20.43	15.72	EMA	19.36	15.11	10.41	3.64	34.48
E05	5.04	JFA	14.25	20.18	15.14	EMA	19.30	15.13	10.09	4.16	34.43
C01	7.23	JFA	10.34	22.43	15.20	CFA	18.12	14.65	7.42	2.92	32.77
C02	6.69	JFA	13.60	19.26	12.57	CFA	15.12	17.74	11.05	2.55	32.87
C03	7.62	JFA	12.68	20.92	13.30	CFA	15.55	17.99	10.37	2.31	33.60
C04	6.38	JFA	8.38	25.23	18.85	CFA	21.04	12.55	6.17	2.21	33.61
C05	7.87	JFA	10.30	25.95	18.08	CFA	20.18	16.05	8.18	2.12	36.25
J01	8.64	JFA	10.67	23.49	14.84	JFB	17.77	16.42	7.78	2.92	34.19
J02	9.98	JFA	10.77	21.89	11.91	JFB	14.12	18.56	8.58	2.21	32.68
J03	9.49	JFA	12.13	22.40	12.91	JFB	15.72	18.78	9.29	2.84	34.53
J04	10.97	JFA	10.85	23.25	12.28	JFB	13.95	20.14	9.17	1.67	34.10
J05	12.03	JFA	8.98	23.73	11.70	JFB	13.20	19.52	7.48	1.50	32.72
J06	11.06	JFA	9.72	25.21	14.15	JFB	15.74	19.18	8.12	1.60	34.92
J07	11.92	JFA	10.49	24.71	12.79	JFB	14.51	20.69	8.77	1.72	35.20
J08	11.39	JFA	9.70	23.42	12.03	JFB	13.23	19.87	8.48	1.22	33.12
J09	12.80	JFA	10.24	24.21	11.41	JFB	12.40	22.05	9.25	0.99	34.45
J10	11.04	JFA	14.48	20.05	9.02	JFB	10.74	23.80	12.76	1.73	34.54
J01	6.82	JFA	10.66	21.62	14.79	JMA	18.81	13.53	6.71	4.02	32.34
J02	9.02	JFA	12.06	21.02	12.00	JMA	15.18	17.90	8.88	3.18	33.08
J03	9.29	JFA	9.51	24.87	15.58	JMA	17.66	16.72	7.43	2.08	34.38
J04	10.03	JFA	11.56	23.85	13.83	JMA	15.74	19.68	9.66	1.91	35.42
J05	9.49	JFA	10.10	25.48	15.99	JMA	18.53	17.04	7.55	2.55	35.57
J06	7.03	JFA	9.26	24.69	17.66	JMA	20.66	13.30	6.27	2.99	33.96
J07	7.76	JFA	9.51	25.25	17.49	JMA	20.35	14.43	6.67	2.85	34.78
J08	9.62	JFA	11.28	23.20	13.58	JMA	15.69	18.77	9.15	2.13	34.47
J09	9.60	JFA	12.58	25.40	15.80	JMA	18.37	19.60	9.99	2.58	37.98
J10	7.75	JFA	13.15	20.52	12.77	JMA	15.76	17.91	10.16	2.99	33.67

[3] Campbell, N., Databases of Expressive Speech, Journal of Chinese Language and Computing, Vol 14, N.4, pp 295-304, 2004.

[4] Goffman, E., *Behaviour in public places: Notes on the social organisation of gatherings*, Free Press of Glencoe, New York, 1963.

[5] Jefferson, G., Notes on a possible metric which provides for a 'standard maximum' silence of approximately one second in conversation. In D. Roger and P. Bull (Eds. *Conversation; An interdisciplinary perspective*. Clevedon, UK: Multilingual Matters. [Expanded version in *Tilburg Papers in Language and Literature, No. 42*, 1-83 (1983). 1988.

[6] Jefferson, G., On a Failed Hypothesis: 'Conjunctionals' as Overlap-Vulnerable. *Tilburg Papers in Language and Literature, No. 28*, 1-33. Tilburg: Tilburg University. 1983.

[7] Jefferson, G., Another Failed Hypothesis: Pitch/Loudness as Relevant to Overlap Resolution. *Tilburg Papers in Language and Literature, No. 38*, 1-24. Tilburg: Tilburg University. 1983.

[8] McNeill, D., Quek, F., McCullough, K-E., Duncan, S., Furuyama, N., Bryll, R., Ma, X-F., & Ansari, R. Catchments, Prosody, and Discourse. *Gesture* 1, 9-33. 2001.