Inter-Speaker Interaction in Speech Rhythm

Kuniko KAKITA
Department of Liberal Arts and Sciences
Faculty of Engineering, Toyama Prefectural University
Kosugi-machi, Imizu-gun, Toyama 939-03, JAPAN
e-mail: kakita@pu-toyama.ac.jp

0. Introduction

Inter-speaker interaction has been studied extensively in the field of psychology and sociology, with main focus on the personalities and social attitudes of the speakers.[1,2] However, little is known about the actual manifestation of inter-speaker interaction in terms of speech production and its acoustic consequences. The ultimate goal of the present study is to clarify the manner of interaction between two speaker's speech characteristics, e.g. speech rate, pause duration, and fundamental frequency of the voice, and to formulate it's underlying rule(s).

For this purpose, two lines of studies have been conducted. One is the 'take-over study' which aims to examine the effect of the preceding speaker's speech on the following speaker's speech in the reading of a text.[3,4] The other is the 'dialog study' which aims to examine the interaction of the two speakers' speech characteristics in dialogs, where two people speak alternately.[5] The following is the status report of these two lines of studies.

1. 'Take-over study'

The objective of this specific study is to examine the way in which a person's speech is influenced by the preceding speaker's speech.

1.1 Method

Utterance texts consisting of five short sentences were read by five subjects in two different ways. In one condition, a series of five sentences (S1-Int1-S2-Int2-S3-Int3-S4-Int4-S5; S=Sentence, Int=Intersentence interval) was read through by a single subject; in the other condition, the subject took over from the preceding speaker after the third sentence and read the remaining sentences (Int3-S4-Int4-S5). The durations of the sentences and intersentence intervals following the takeover were compared with those at corresponding locations in single speaker readings.

1.2. Results and Discussion

Fig. 1 compares the five subjects' 'post-takeover' durations (filled data points) with their 'inherent' durations (unfilled data points). The data points for both the 'inherent' and 'post-takeover' durations represent the median duration values derived from 20 sentences (S4 x 10 and S5 x 10) and 20 intervals (Int3 x 10 and Int4 x 10). The cross in the figure indicates the 'inherent' sentence/interval duration of the preceding speaker, and is derived from 30 sentence samples (S1 x 10, S2 x 10 and S3 x 10) and 20 interval samples (Int1 x 10 and Int2 x 10).

The figure shows that both the sentence duration and the interval duration following the takeover (the 'post-takeover' duration) deviated from one's 'preferred' duration obtained from single speaker readings (the 'inherent' duration). The figure also reveals that, for all subjects, the deviation was mostly assimilative, and that for both sentences and intervals,

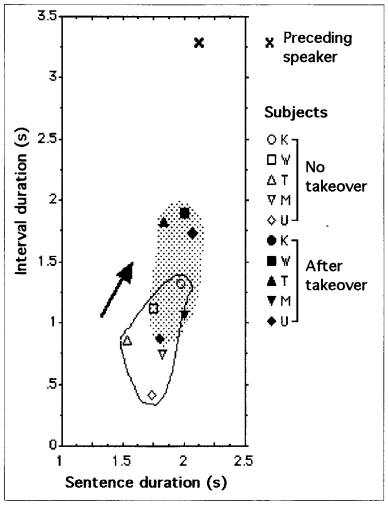


Fig. 1 The duration of sentences and intervals in the two experimental sessions. [Session A (no takeover): unfilled data Points; Session B (takeover after S3): filled data points]

the 'post-takeover' duration settled somewhere between the subject's 'inherent' duration and the preceding speaker's 'inherent' duration.

The subjects' 'post-takeover' durations became closer to the preceding speaker's values but not quite to the extent of completely assimilating to the preceding speaker. This observation leads one to hypothesize that, when the subjects are exposed to the preceding speaker's speech, they reorganize their speech according to two competing criteria, 'their own temporal criterion' and 'the preceding speaker's temporal criterion'. Inter-speaker interaction, then, may tentatively be defined as the reorganization of one's speech with reference to one's own speech production framework, on one hand, and to the speech production framework of another speaker, on the other.

The duration of sentences and intervals was reexamined with respect to two different ways of looking at inter-speaker interaction. One was to see 'how close the subjects' durational values came to the preceding speaker's value'. This was called 'approximation', and was defined as the difference between the preceding speaker's 'inherent' duration and the subjects' 'post-takeover' duration. The other way of looking at the result of interaction was to see how much one's 'post-takeover' value deviated from one's 'inherent' value. This was called 'deviation', and was defined as the difference between the subject's 'inherent' duration and the subjects' 'post-takeover' duration. It was argued that if reorganization of sentence and interval duration (under speaker interaction) was governed primarily by 'one's own temporal criterion', one would more likely find invariance in 'deviation', whereas if it was governed primarily by 'the other speaker's criterion', one would more likely find invariance in 'approximation'.

In Figures. 2(a) and 2(b), 'approximation' and 'deviation' are plotted against 'inherent difference', the difference between the 'inherent' durations of the preceding speaker and the subject. The filled data points pertain to sentences, while the unfilled data points pertain

to intervals. The dotted line in Figure 2(a) denotes 'full approximation', while the dotted line in Figure 2(b) denotes 'no deviation'.

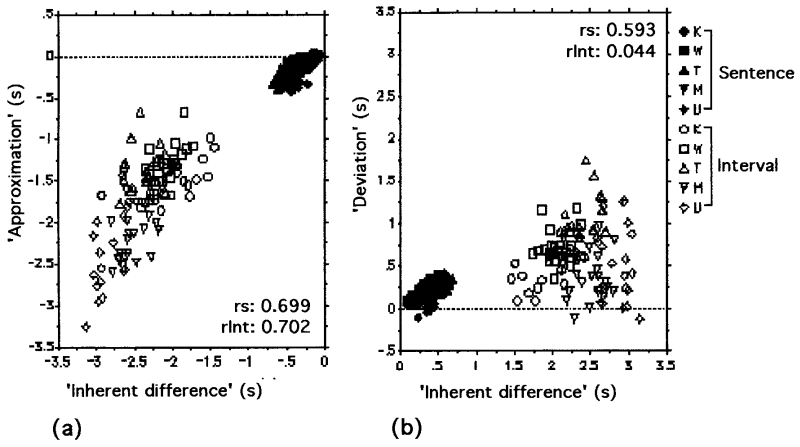


Fig. 2. (a) 'Approximation' (the difference between the preceding speaker's 'inherent' duration and the subject's 'post-takeover' duration (s)) Plotted against 'inherent difference' (the difference between the 'inherent' durations of the preceding speak (b) 'Deviation' (the difference between the subject's 'POSt-takeover' duration and the 'inherent' duration (s)) Plotted against 'inherent difference' (the difference between the 'inherent' durations of the preceding speaker and the subject (s)) [The 'inherent' duration denotes the duration of Int3, S4, Int4, or S5 in Session A, while the 'post-takeover' duration denotes the duration of Int3, S4, Int4, or S5 in Session B. See Fig. 1]

The results showed that sentences and intervals differed characteristically in the way 'approximation' and 'deviation' were correlated with 'inherent difference'. For sentences, high positive correlation was observed both between 'approximation' and 'inherent difference' (r: 0.699*) and between 'deviation' and 'inherent difference' (r: 0.593*). For intervals, however, the correlation between 'approximation' and 'inherent difference' was very high (r: 0.702*) but the correlation between 'deviation' and 'inherent difference' was extremely low (r: 0.044). [*: Significant at the 0.01 level].

The fact that 'approximation' and 'deviation', for sentence duration, manifested equally high correlation with 'inherent difference' was interpreted to suggest that the reorganization of sentence duration was governed both by 'one's own temporal criterion' and by 'the preceding speaker's criterion'. The fact that, for interval duration, the correlation between 'approximation' and 'inherent difference' was high whereas that between 'deviation' and 'inherent difference' was low, was interpreted to suggest that the reorganization of interval duration was governed primarily by 'the preceding speaker's criterion' and that 'one's own criterion' did not play a significant role in the reorganization of interval duration.

It is interesting to note, too, that the correlation between 'deviation' and 'inherent difference' was high for sentences but low for intervals. This can be explained in terms of the characteristic difference in the production of sentences and intervals. The production of

sentences (utterances) involves active movements of the articulators and is also accompanied by aural and tactile feedback. It would be relatively easy to establish and maintain 'one's own temporal criterion' in the presence of inter-speaker interaction. Hence the high correlation. In contrast, during intervals, there are no such 'active' movements or feedback. It would be more difficult to maintain 'one's own temporal criterion', or to put it differently, it would be easier for interval duration to 'deviate' from the 'inherent' value. Hence the low correlation.

2. 'Dialog Study'

The results of the 'take-over study' suggested that, when two speakers spoke alternately in a dialog, speech characteristics of each speaker would assimilate to those of the dialog partner and that two speakers' speech characteristics would become increasingly similar as the dialog progressed. To test this, controlled dialogs of simple question-answer type were recorded and analyzed. The results pertaining to the fundamental frequency (F0) of voice will be reported here. The results of the temporal characteristics of speech, e.g. speech rate, will be discussed in the final report (now in preparation).

2.1. Method

Each speaker was given a sheet of paper with a list of 20 identification numbers (1A-1 to 1A-20) and corresponding 20 five digit numbers. Sp(eaker)1 would choose from the list an ID# at random and ask, for example, 'What is 1A-3?' Sp(eaker)2 would then search for the number corresponding to that ID# and answer, e.g. 'It's eleven thousand three hundred fourteen.' Thus, a dialog proceeded in the following manner: Sp1 questions – Sp2 answers – Sp2 questions – Sp1 answers – Sp1 questions . . . Twelve dialogs, involving 11 speakers, were recorded. Most of the speakers took part in more than two dialogs – with different partners each time. For each dialog recorded, there were 25 utterances for each of the two speakers.

The peak F0 value corresponding to /...ie:.../ portion of /ichie:/['1A...'] in the question part of the utterance was measured. For each speaker in each dialog, 25 F0 measurements were obtained.

2.2. Results and Discussion

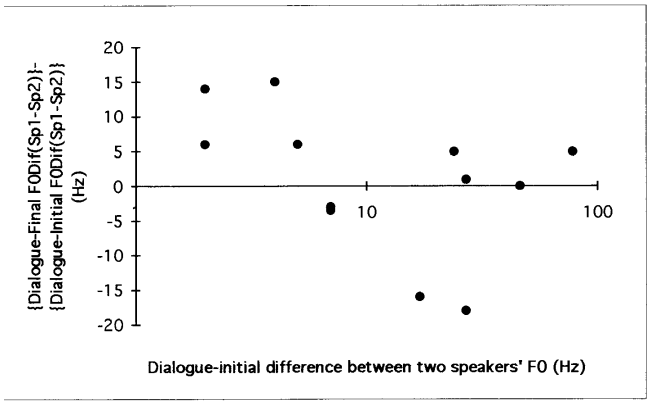
The F0 values of the two speakers came closer together as the dialog progressed only in 4 out of 12 dialogs. In the rest of the cases, the two speakers' F0 values either became farther apart, or stayed more or less the same throughout the course of the dialog. The interaction in F0, then, did not manifest itself dominantly as assimilation, as expected; on the contrary, the manner of F0 interaction between dialog partners was found to be realized in many varying ways.

The question to be asked then was if there was a rule behind this varying interaction patterns.

Since the speakers varied in their natural (inherent) voice pitch, some of the dialogs were characterized by a large difference in the two speakers' F0 at the start of the dialog, while others were characterized by a small difference, the rest falling somewhere in between. It was asked if dialog-initial difference in the two speakers' F0 determined the consequent pattern of F0 interaction during the dialog – assimilation, dissimilation, etc.

In Fig. 3, the difference between the dialog-initial F0 difference between two speakers and the dialog-final F0 difference between two speakers is plotted against the dialog-initial difference between the two speakers' F0. The negtive values on the ordinate indicate that two speakers' F0 values came closer to each other, whereas the positive values indicate that

they became farther apart. The values near zero indicate that the two speakers' F0 difference did not change much during the dialog.



(Hz)

Fig. 3 Difference between the dialogue-initial F0 difference between two speakers and the dialogue-final F0 difference between two speakers Plotted against the dialogue-initial difference between two speakers' F0.

The plot shows that the two speakers' F0 values tend to assimilate when dialog-initial difference between the two speakers' F0 is within a specific range – roughly between 5Hz and 25Hz. Outside this range, the two speakers' F0 tend to dissimilate. It also shows that the degree of dissimilation is smaller when dialog-initial difference (F0di) between the two speakers is great (F0di¿25Hz) than when it is small(F0di¡5Hz).

In Fig. 4, the difference between (a) the dialog-initial F0 difference between two speakers and (b) the dialog final F0 difference between two speakers is plotted against the difference between two speakers' median F0. The median F0 value was derived from 25 tokens/dialog for each speaker. The (absolute) values along the ordinate shows how much difference there was between the two speakers' initial F0 difference and final F0 difference, irrespective of the fact that the difference was a decrease (assimilation) or an increase (dissimilation).

The plot shows that the absolute amount of change in the two speakers' F0 differences (F0d) is great (F0d¿=14Hz) when the difference between the two speakers' median F0 (F0dm) is within a certain range (8Hz=¡F0dm=¡18Hz) but that it is small (F0d=¡6Hz) when the median F0 difference falls outside this range.

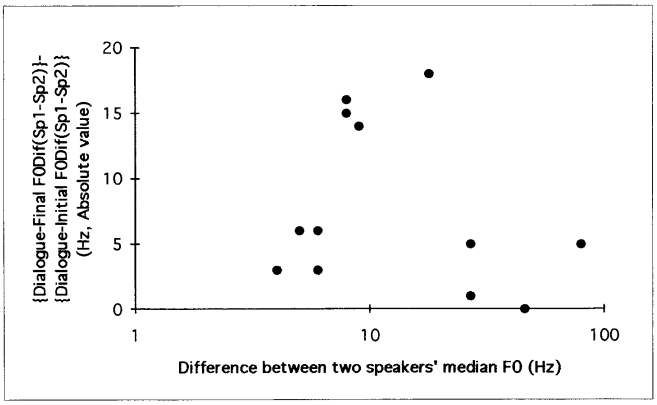
Based on the results of this 'dialog study', some tentative rules of inter-speaker interaction may be expressed as follows:

(1) The two speakers' speech characteristics assimilate when they are within certain optimal range.

(2) Outside this range, they dissimilate.

(3) The amount of dissimilation, however, is small when two speakers' speech characteristics (either at the start of the dialog, or throughout the dialog) are too 'far apart'.

References



(Hz)

Fig. 4 Difference between the dialogue-initial F0 difference between two speakers and the dialogue-final F0 difference between two speakers Plotted against the difference between two speakers' median F0. The median F0 value was derived from 25 tokens/dialogue for each speaker.

[1] Matarazzo, J. D. et. al., (1963) Interviewer Influence on Durations of Interviewee Speech, Journal of Verbal Learning and Verbal Behavior, 1, 451-458.

[2] Jaffe, J & Feldstein, S (1970) Rhythms of dialog , (Academic Press: New York).

[3] Kakita, K. (1994) Inter-Speaker Interaction of the Duration of Sentences and Intersentence Intervals, Proceedings of the Fifth Australian International Conference on Speech Science and Technology, Vol. 1, 34-38

[4] Kakita, K. (1994) Hatsuwa no washakan sougosayou: hatsuwa sokudo to pohzuchou ni kannsuru kekka. Proc. of the 94-Fall Meeting of the Acoustical Society of Japan 383-389 [in Japanese]

[5] Kakita, K. (1995) Kan'tan na shitsumon-kotae keishiki no taiwaniokeru F0 no washakan sougosayou, Proc. of the 95-Fall Meeting of the Acoustical Society of Japan 305-306 [in Japanese]

Keywords: Inter-speaker interaction, speech rhythm, speech rate, pause duration, fundamental frequency