# Informational Potentials of Dynamic Speech Rate in Dialogue

**Hanae Koiso*** and **Atsushi Shimojima** and **Yasuhiro Katagiri**
ATR Media Integration & Communications Research Laboratories
2-2 Hikari-dai Seika Soraku Kyoto, 619-02 Japan
{koiso, ashimoji, katagiri}@mic.atr.co.jp

## Abstract

We examine five spontaneous dialogues conducted in Japanese and analyze the potential of speech rate change to signal the structure of information being exchanged in dialogue. We found (1) a bi-directional correlation between speech decelerations and the openings of new information, and (2) another bi-directional correlation between speech accelerations and the absence of information openings. Our data show that the correlations hold not only in the case of a single speaker's speech, but also in the case of multiple speakers' sequential utterances, with or without turn shifts. We also study possible disturbances to these default correlations and identify the limitation on speakers' cognitive resources as one major constraint that interferes with the accurate signaling of information opening by decelerated speech.

## Introduction

The research reported in this paper is an attempt to elucidate the function of dynamic speech rate as contextualization cues in conversational Japanese. We claim that changes in speech rate in conversational Japanese have definite potential in cuing the structure of information collaboratively constructed by the participants of a conversation.

When people engage in conversation, they not only express, by means of uttering words and sentences, what they intend to get across to other conversants about the topics of discourse, but also express, both intentionally and unintentionally, messages concerned with the situation in which the conversation is taking place. This second type of expression in many ways directs how the first type of expression is to be interpreted by conversants. The second type of expression is usually carried out through non-referential and non-lexical means such as prosody, gesture/posture, gaze and backchannels. Gumperz and other sociolinguists (Gumperz, 1989; Gumperz, 1991; Erickson and Schultz, 1982; Auer and Luzio, 1992) introduced the notion of contextualization to capture these processes in everyday language use, and called the signals involved in the second type of expression "contextualization cues."

There has not been much research devoted to speech rate phenomena from the perspective of their potential as contextualization cues. This is partly because primary interest in speech rate phenomena has been directed toward apparent rhythmicity in speech and its relationship with underlying physiological mechanisms of speech production and perception. Speech researchers have been studying temporal organization of human speech and have found that an isochronous unit can be hypothesized behind speech production and perception (Lehiste, 1975), even though the actual speech materials do not manifest exact isochrony due to intervening physiological, cognitive, and linguistic factors. It has also been observed among speech engineers (Sagisaka and Tohkura, 1984; Fujisaki and Higuchi, 1980) that control of the temporal organization of synthetic speech greatly contributes to its naturalness. Their interests have also been focused on physiological and cognitive mechanisms of human speech production and perception, and their results are mostly based on read speech data of at most sentence length.

In contrast, our research focuses on the informational potential of speech rate changes observed in human spontaneous dialogues. In actual conversations, we notice that people change their rates of speech from time to time, and these local modulations in speech rate, together with insertion of pauses, provide hearers with various clues about the state of the speakers and the state of ongoing conversations, which will subsequently be exploited by hearers to steer their reasoning in interpreting what they hear. Thus, deviations from an isochronous speech rate in conversation have potential as contextualization cues. Several researchers (Crystal and House, 1990; Koopmans-van Beinum and van Donzel, 1996) have recently started looking at their informational potential to indicate discourse related information, such as the given/new distinction and the importance of the information being expressed. In this paper, we focus our attention to local changes of speech rate, namely, accelerations and decelerations of speech over consecutive pairs of utterances. Speech rate is particularly interesting for conversations since it can be, and we've found that it is, collaboratively managed through synchronization and sharing of speech rate across speakers.

The problems, then, are (a) exactly what types of information are conveyed by local dynamic speech rate, and (b) how that information is exploited by conversational participants. The main focus of this paper is on the first problem of informational potential of speech rate, and we examine corpus data of spontaneous interactive dialogues to validate our hypothesis relating local dynamic speech rate to information structures.

## Hypothesis

Interactive dialogues manifest information structures at various levels: local coordinative structures (e.g., adjacency pairs), dialogue topic structures, and higher level intentional structures on both discourse and task related plans. These structures form a hierarchy of information structures for dia-

*Also with Nara Institute of Science and Technology.

logue. Within a hierarchical structure, the simplest piece of information to be signaled by contextualization cues is the boundary/non-boundary between information units. With regard to sentence and phrase-final prosodic rise/fall patterns, Pierrehumbert and Hirschberg (1990) and Hirschberg and Grosz (1992) found that English prosody can signal the end boundaries of information units at the end of prosodic segments. Similarly, we consider that local changes in speech rate can mark the start boundaries of information units at the beginning of local speech segments. The hypothesis we examine in this paper can be stated as follows:

> Local decelerations and accelerations in speech rate can work as contextualization cues for the openings of new information units and for the continuations of existing information units, respectively.

## Methods

### Dialogue Data

We examined five task-oriented dialogues conducted by different pairs of speakers to test our hypothesis. All of the dialogues were taken from the Chiba Map Task Corpus (Aono et al. 1994), a Japanese-language version of the Edinburgh Map Task Corpus (Anderson et al. 1991). In each dialogue, one speaker (the "follower") tries to draw a route on a simple map of landmarks, following the instructions of the other speaker (the "giver"), who holds a similar map with the complete route on it. The giver's map and the follower's are hidden from each other. Also, landmarks on one map may be labeled differently, or even absent, from the other map. This makes the task reasonably difficult, and the average length of the dialogues in the entire corpus is about 11 minutes. We used the first 3 minutes of each dialogue for our analysis.

### Speech Rate Measurement

The dialogue corpus data consist of speech data with their transcriptions into Japanese. The speech data are sampled at 20KHz and with 16-bit accuracy. Our measurement of speech rate was based on the average mora durations of sub-utterance units. A *sub-utterance unit* (SU) of a dialogue is a stretch of a single speaker's speech bounded by silence, and the *average mora duration* (AMD) of an SU is the duration of the SU divided by the number of morae appearing in the SU[1]. Thus, a smaller AMD of an SU means a greater speech rate over the SU. Table 1 provides basic facts about our data, including the mean and the standard deviation of the AMDs in each dialogue.

We compare the speech rates of a pair of temporally consecutive SUs and call a pair *accelerating* if the AMD of the succeeding SU is smaller than that of the preceding SU; a pair is *decelerating* if the AMD of the succeeding SU is greater than that of the preceding SU.

### Information Units

We intend to study whether the changes in speech rate can function as a contextualization cue to the openings of information units in dialogue. For this purpose, we classify the pairs of SUs in dialogue into *opening pairs* whose second SUs start the expression of (typically new) pieces of information and *non-opening pairs* whose second SUs do not start such an expression. Our task is to study possible correlations between the acceleration/deceleration properties of the pairs of SUs and their opening/non-opening properties.

[1] Crystal and House (1990) used the average syllable duration (ASD) as the measure of speech rate in English. Our measure replaces syllables with morae, which are the segmental units of Japanese speech.

Table 1: Characteristics of sub-utterance units(SUs) and average mora durations (AMDs) in dialogues from the Chiba Map Task Corpus.

| | **SU** | | **AMD** | |
|---|---|---|---|---|
| dialogue | number | mean duration (ms) | mean (ms) | std. dev. (ms) |
| 1 | 109 | 836.11 | 125.33 | 64.64 |
| 2 | 93 | 873.03 | 122.22 | 61.39 |
| 3 | 134 | 781.58 | 121.34 | 46.26 |
| 4 | 132 | 612.02 | 128.50 | 76.18 |
| 5 | 137 | 777.04 | 122.98 | 59.43 |

It should be noted that task-oriented dialogues are particularly suitable for our analysis because the structure of the task itself makes it clear what information is to be exchanged in what order, and thus it is relatively easy to reach an agreement about when a piece of information is introduced during the dialogue. This is important because we need an independent characterization of information structures in studying the relationship between dynamic speech rate and information structures to avoid circular arguments. The task for the Map Task dialogues consists of subtasks for the identification of landmarks and the following of paths by connecting landmarks. We determined openings of information units by referring to the start of these subtasks. Exact points are located by identifying syntactic markings of sentence and phrase beginnings.[2]

### Types of Sub-utterance Unit Pairs

(a)

Speaker A: $\boxed{u_1}\boxed{u_2}$

(b)

Speaker A: $\boxed{u_1}$
B: $u_2$ (backchannel)
A: $\boxed{u_3}$

(c)

Speaker A: $\boxed{u_1}$
B: $\boxed{u_2}$

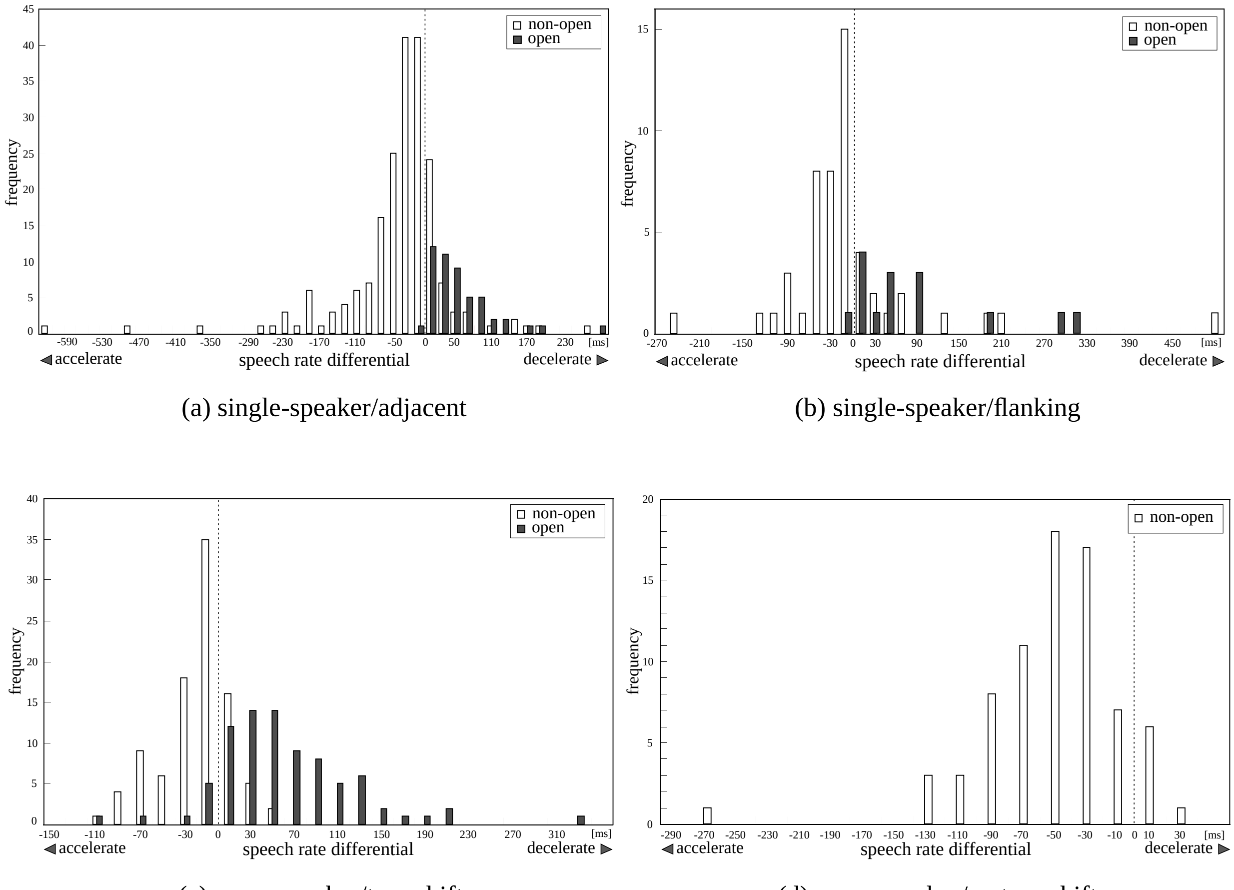
(d)

Speaker A: $\boxed{u_1}$
B: $\boxed{u_2}$ (backchannel)

Figure 1: Types of SU pairs considered.

We classify pairs of SUs into four categories in terms of conversational turn structures. The categories are:

[2] Our demarcations of information openings roughly coincide with what Traum (1994) calls "initiations of discourse units." It is, however, yet to be seen whether we can use Traum's notion of "discourse unit" as a precise characterization of information units in conversation.




(a) single-speaker/adjacent

(b) single-speaker/flanking

(c) cross-speaker/turn shift

(d) cross-speaker/no turn shift

Figure 2: Speech rate and information openings.

(a) the pairs of directly adjacent SUs made by same speakers,
(b) the pairs of SUs made by same speakers flanking a voiced backchannel,
(c) the pairs of SUs involving turn shifts across speakers, and
(d) the pairs of SUs involving voiced back channels from hearers but no turn shifts.

The temporal structure of pairs for each of these categories is shown in Figure 1.

## Results

Figures 2(a) to 2(d) respectively show the distributions of SU pairs of categories (a) to (d), relative to their properties of acceleration/deceleration and opening/non-opening of information units. The abscissa of each figure marks the ranges of local speech rate changes over the SU pairs. The black and the white column over a range represent the number of opening pairs and the number of non-opening pairs, respectively, falling within the range. Since backchannels, by definition, do not start new information units, all the data in Figure 2(d) are classified as non-openings. Table 2 summarizes the results.

These figures clearly show that there are distributional differences between opening and non-opening SU pairs in local dynamic speech rate changes. Openings of new information units tend to co-occur with decelerations of speech, and non-openings tend to co-occur with accelerations of speech. There is a significant effect of information opening/non-opening on dynamic speech rate: $F(1, 571) = 158.144, p < .001$. The same effect is found in each case of (a), (b), and (c): $F(1, 251) = 42.466, p < .001$; $F(1, 177) = 136.367, p < .001$; $F(1, 64) = 9.016, p < .01$, respectively. Thus co-ocurrences are universal across SU pairs in different turn structures.

There are two points that need to be noted. First, from Figures 2(a) and 2(b), it is evident that insertion of backchannels by other parties doesn't change the speech rate characteristics within single speaker utterances. Secondly, a comparison of Figures 2(a)(b) with Figures 2(c)(d) shows that local speech rate characteristics are universal across single-speaker utterances and cross-speaker utterances.

Table 2: Speech rate change in dialogues.

| | single-speaker | | | | cross-speaker | | |
|---|---|---|---|---|---|---|---|
| | (a) adjacent | | (b) flanking | | (c) turn-shift | | (d) backchannel |
| | open | non-open | open | non-open | open | non-open | non-open |
| num | 50 | 203 | 15 | 51 | 82 | 97 | 75 |
| mean | 55.63 | -43.15 | 83.99 | -5.57 | 62.47 | -21.35 | -52.75 |
| std.dev. | 59.78 | 93.78 | 105.81 | 124.29 | 65.61 | 33.56 | 45.45 |

Table 3: Accuracy and comprehensiveness.

| | **Ia** non-open /accel. | **Ib** accel. /non-open | **IIa** open /decel. | **IIb** decel. /open |
|---|---|---|---|---|
| (a) | 99.4% | 78.8% | 66.1% | 74.0% |
| (b) | 97.4% | 74.5% | 52.6% | 66.7% |
| (c) | 91.3% | 76.0% | 90.0% | 76.8% |
| (d) | – | 90.7% | – | – |
| total | 96.8% | 79.8% | 75.9% | 74.8% |

## Informational Potentials

Observation of these clear correlations between local speech rate dynamics and openings of information leads us to examine the potential of local speech acceleration/deceleration to work as contextualization cues for signaling information structures. For a possible signal to work as a contextualization cue, we need to establish its signal accuracy and comprehensiveness. The former is a measure of the correctness of information signaled, whereas the latter is a measure of the coverage of the signaling. Table 3 summarizes our analysis of speech rate dynamics from the standpoint of these two factors.

### Acceleration

In order for accelerations in local speech to be *accurate* signals for non-openings of information, there must be a regularity of the form:

**Ia: Accuracy of acceleration signals**

If a pair of SUs is accelerating, it is a non-opening pair.

Correspondingly, in order for accelerations in local speech to be *comprehensive* signals for non-openings of information, there must be a regularity of the form:

**Ib: Comprehensiveness of acceleration signals**

If a pair of SUs is non-opening, it is accelerating.

The regularity **Ia** means that, at least potentially, speech accelerations function as accurate markers of non-openings of information. On the other hand, the regularity **Ib** means that all non-openings of information are potentially marked by speech accelerations.

In each of the Figures 2(a) to 2(d), the area to the left of the zero line has much more (and longer) white columns than black columns. That is, a great majority of the accelerating pairs of SUs are in fact non-opening pairs. More specifically, Table 3 shows that 99.4%, 97.4%, and 91.3% of the accelerating pairs of types (a), (b), and (c), respectively, are non-opening pairs. Overall, 96.8% of the accelerating pairs are non-opening pairs. These numbers support the regularity **Ia**, and indicate that accelerated speech is a fairly *accurate* signal for non-openings of information.

Note also that in each of the figures, most of the white columns are to the left of the zero line. That is, a great majority of the non-opening pairs of SUs are accelerating pairs. In fact, Table 3 shows that 78.8% of the non-opening pairs of type (a), 74.5% of the non-opening pairs of type (b), and 76.0% of the non-opening pairs of type (c) are accelerating pairs. A pair of type (d) has a voiced backchannel as the second SU, and by definition a backchannel opens no piece of information in dialogue. So, if it is a general tendency that non-opening pairs are accelerating pairs, then a great majority of the pairs of SUs of type (d) must be accelerating. In fact, our data shows that as much as 90.7% of the pairs of type (d) are accelerating. Overall, 79.8% of the non-opening pairs are accelerating. These numbers support the regularity **Ib**, and indicate that accelerated speech is also a fairly *comprehensive* signal for non-openings of information.

### Deceleration

Parallel to the regularities in accelerations, the following forms of regularities must hold in order for speech decelerations to be accurate and comprehensive signals for information openings:

**IIa: Accuracy of deceleration signals**

If a pair of SUs is decelerating, it is an opening pair.

**IIb: Comprehensiveness of deceleration signals**

If a pair of SUs is opening, it is decelerating.

The regularity **IIa** means that, at least potentially, speech decelerations function as accurate markers of openings of information. On the other hand, the regularity **IIb** means that all openings of information are potentially marked by speech decelerations.

Unlike the case of accelerations, however, our data suggest that, as far as signaling regularities are concerned, a sizable amount of slowdown between SU pairs is necessary to be counted as a deceleration, especially for the accuracy regularity **IIa**. From inspection of the data displayed on the right-hand sides of Figures 2(a) to 2(c), we choose +20ms for our examination of the deceleration regularities.

From Figure 2 we see more (and longer) black columns than white columns in the area to the right of the +20ms line, except for Figure 2(b). This means that for SU pairs of types (a) and (c), a great majority of the pairs decelerating by more than 20ms are opening pairs. More specifically, 66.1% and 90.0% of the decelerating pairs of types (a) and (c), respectively, are opening pairs. These numbers support the regularity **IIa**, and indicate that decelerated speech is a fairly *accurate* signal for openings of information, at least for pairs not involving backchannels.

In each of the Figures 2(a) to 2(c), most of the black columns are to the right of the +20ms line. Table 3 shows that 74.0% of the opening pairs of type (a) decelerate by more than 20ms, and the same holds for 66.7% and 76.8% of opening pairs of types (b) and (c), respectively. These numbers support the regularity **IIa**, and indicate that decelerated speech is also a fairly *comprehensive* signal for openings of information.

Thus, we can conclude that speech rate changes in dialogue have definite potential for signaling the structure of information being expressed. Speech decelerations of more than 20ms cue openings of information, while speech accelerations mark the absence of information openings. Both types of signals are highly accurate and comprehensive.[3]

## An Interfering Factor

It would be unrealistic to suppose that these cues are infallible and totally comprehensive. As Table 3 shows, the correlations between dynamic speech rate and openings/non-openings of information are only statistical, not logical. The regularities **Ia** and **IIa** allow overall exceptions of 3.2% and 24.1%, respectively. Consequently, there are cases in which speech rate changes carry *mis*information: speech acceleration may indicate the absence of an information opening where there is one, and deceleration may indicate an information opening where there is none. Also, the regularities **Ib** and **IIb** allow overall exceptions of 20.2% and 25.2%, respectively. That is, there are cases in which due signals of dynamic speech rate are *absent*: there may be an information opening where there is no speech deceleration that signals it, and some non-openings may be unmarked by speech accelerations.

Table 3 shows that the greatest irregularities are to the accuracy of deceleration signals **IIa** in single speaker speech: of all type-(b) pairs that decelerate by more than 20ms, 47.4% are non-opening; of all type-(a) pairs with the same deceleration property, 33.9% are non-opening. Let us look at these irregular pairs more closely to find what kinds of extra factors can interfere with the accurate signaling of openings/non-openings of information by dynamic speech rate.

On a closer examination of the immediate contexts of the irregular pairs in question, we find:

A. 48.1% have their second SUs preceded by silence of more than 700ms.[4] For example:[5]

| | | |
|---|---|---|
| G: | `ima iru tokoro kara` | (From where you are now) |
| F: | `hai` <4256> | (Yes) |
| G: | `nanameue ni` | (Obliquely above) |
| F: | `hai` | (Yes) |
| G: | `iki masu` | (Proceed) |

B. 33.3% have their second SUs followed by silence more than 500ms, which in turn is followed by the same speaker's SU.

| | | |
|---|---|---|
| G: | `sekizō no aida wo` | (Between the stone statues) |
| G: | `tōtte` <1408> | (Go through) |
| G: | `sekizō wo gurutto mawatte kudasai` | (Go around the stone statue) |

C. 29.6% have second SUs that are exceptionally slow, namely, their AMDs are 300ms or more.[6]

| | | |
|---|---|---|
| G: | `sosite soko kara::` | (And from there) |
| G: | `ettone` (AMD: 320ms) | (Well) |
| G: | `hidari:::` | (The left) |
| G: | `hidari naname ni`<576> | (Obliquely to the left) |

D. 11.1% have filled pauses, or "fillers," as their second SUs.

| | | |
|---|---|---|
| G: | `hansen no::` | (Of the sailboat) |
| G: | `n:::` | (Uhm) |
| G: | `hidari no yonsenti kurai` | (About four centimeters to the left) |

There are many irregular pairs that have two or more of the structural properties A–D. The following is a pair of SUs that has *all four*:

| | | |
|---|---|---|
| G: | `syuppatu titen kara::` | (From the starting point) |
| F: | `hai` <864> | (Yes) |
| G: | `eetto:::` (AMD: 331ms) <1536> | (Well) |
| G: | `karuku migini hukureru youna kanji de` | (As though you lightly deviate to the right) |

A plausible account of these structural properties is that since the cognitive load in producing the second SU or subsequent SUs is heavy, the speaker delays and slows down the second SU (case A), or slows down the second SU and delays the third SU (case B), or makes the second SU exceptionally slow (case C), or uses a slowed-down filler before the third SU (case D). Overall, 74.1% of the irregular pairs in question have one or more of these structural properties, and thus can be plausibly attributed to the cognitive load in producing the second SU or its successors.[7]

---

[3] Note that these findings are concerned with the *potential* of signaling. They are the basis for the study of the cognitive processes in which such potential is exploited by conversants.

[4] A silence of 700ms is more significant than it may first appear. The mode of silence duration of our data is 150ms, and silences of 700ms or longer are less than 20% of all silences.

[5] In the following, the pairs of SUs in question are enclosed in boxes. "G" and "F" stand for the giver and the follower in the dialogue. The number in <> placed after an SU represents the length of silence between the SU and the succeeding SU. An English translation of each SU is shown to its right.

[6] The AMDs of 300ms or longer are "exceptional" in the sense that the value of 300ms is more than twice the standard deviation away from the average AMD. See Table 1 for the average AMD and the standard deviation of the AMDs.

[7] There are also other irregular pairs whose immediate contexts are none of the above kinds, but whose global contexts can be considered "stammering speech." Here we simply characterize the relevant cognitive loads in this and other cases as those in producing second or subsequent SUs, but this does not preclude the possibility that other kinds of cognitive loads (say, in comprehending the partner's speech, memorizing or inferring the configuration of her map) are involved as remote causes.

This suggests the following view on the cuing of information openings by speech decelerations. Generally, in the case of a single speaker's sequential speech, decelerations of more than 20ms potentially signal openings of information. This is perhaps also the case of a single speaker's non-sequential speech flanking a voiced backchannel. With this "default" regularity, however, a cognitive constraint on the production of subsequent speech interferes with and slows down the relevant utterance where there is no opening of information. Given the default informational potential of decelerated speech, this slowing down carries *mis*information, indicating an opening of information where there is none.

Thus, the signaling of information opening and non-opening by speech rate change is not an isolated, or "shielded," function of speech. There can be various extra constraints in dialogue situations that interfere with the regularities **Ia** through **IIb**, making the signaling less accurate and comprehensive. Our analysis in this section identifies the limitations on speakers' cognitive resources as one of the major constraints interfering with the accuracy of the signaling.

## Conclusions

We examined corpus data of spontaneous interactive dialogues, and applied statistical methods to find regular correspondences between local changes in speech rate and the structures of information being expressed. This method permitted us to identify the potential of dynamic speech rate as contextualization cues. We found bi-directional correlations between speech accelerations and the absence of information opening, and between speech decelerations (of more than 20ms) and the presence of information opening. Thus, dynamic speech rate in dialogue can potentially function as accurate and comprehensive cues for openings and non-openings of information expressed in dialogue.

The correlations in question hold not only for single speakers' utterances but also for multiple speakers' sequential utterances with or without turn shifts. Thus, even when there is a shift of speakers during dialogue, the subsequent speaker decelerates her speech if and only if her speech opens up a new piece of information; otherwise, the subsequent speaker maintains the acceleration pattern established by the preceding speaker. Intentionally or unintentionally, the speakers collaborate with each other to maintain these regularities governing dynamic speech rate and information structures.

However, such cues are neither infallible nor totally comprehensive. There are various extra constraints interfering with the regularity of these correlations, which partly reduce the accuracy and comprehensiveness of the cuing. On the basis of the structural properties of the "irregular" sequences of utterances, we identified the limitations on speakers' cognitive resources to be one of the major constraints that interfere with the accurate cuing of information openings by decelerated speech.

This paper is thus a development of Gumperz's idea of contextualization cue in the direction of dynamic speech rate. Our approach, however, went beyond that of our predecessors in using strictly statistical methods to identify the relevant cuing potential and to measure its accuracy and comprehensiveness. Our approach is also distinctive in that it gives explicit attention to the extra constraints that interfere with what we have identified as the default cuing.

In this respect, however, we need a more systematic characterization of what those extra constraints are, when they interfere, and how they make the relevant cuing less accurate or comprehensive. It would also be an interesting extension of our work to investigate how other features of speech (e.g. text, pitch, and power) interact with the dynamic speech rate to make redundant, conflicting, or supplementing signals of information opening/non-opening. These investigations will be taken up in future work.

## References


Anderson, A. H. et al. (1991). The HCRC map task corpus. *Language and Speech*, *34*, 351–366.

Aono, M. et al. (1994). The Japanese map task corpus: An interim report (in Japanese). *SIG–SLUD–9402*, 25–30.

Auer, P. and A. Luzio (1992). *The contextualization of language*. Amsterdam: Johan Benjamins Publishing.

Couper-Kuhlen, E. (1993). *English speed rhythm: form and function in everyday verbal interaction*. Amsterdam: John Benjamins Publishing Company.

Crystal T. H. and A. S. House (1990). Articulation rate and the duration of syllables and stress groups connected speech. *The Journal of the Acoustical Society of America*, *88*, 101–112.

Erickson, F. and J. Shultz (1981). When is a context? Some issues and methods in the analysis of social competence. In: J. L. Green and C. Wallat(eds.), *Ethnography and language in educational settings*. Norwood, NJ: Ablex Publishing Corporation.

Fujisaki H. and N. Higuchi (1980). Temporal organization of segmental features in Japanese disyllable (in Japanese). *The Journal of the Acoustical Society of Japan*, *1*, 25–30.

Gumperz, J. (1989). Contextualization cues and metapragmatics: The retrieval of cultural knowledge. MS, University of Berkeley.

Gumperz, J. (1991). Contextualization and understanding. In: Daranti, A and C. Goodwin(eds.), *Rethinking context*, Cambridge: Cambridge University Press.

Hirschberg, J. and B. Grosz (1992). Intonational features of local and global discourse structure. *Proceedings of DARPA Workshop on Spoken Language Systems*, 441–445.

Koopmans-van Beinum, F. J. an M. E. van Donzel (1996). Relationship between discourse structure and dynamic speech rate. *Proceedings of ICSLP 96*, *3*, 1724–1727.

Lehiste, I. (1975). The Phonetic structure of paragraphs. In: A. Cohen and S. G. Nooteboom(eds.), *Structure and process in speech perception*. Berlin: Springer.

Pierrehumbert, J. and J. Hirschberg (1990). The meaning of intonational contours in the interpretation of discourse. In: P. R. Cohen and J. Morgan and M. E. Pollack(eds.), *Intentions in communication*. MIT press.

Sagisaka, Y. and Y. Tohkura (1984). Phoneme duration control for speech synthesis by rule (in Japanese). *The Transactions of the Institute of Electronics, Information and Communication Engineers*, *J67-A*, 629–636.

Traum, D. R. (1994). *A computational theory of grounding in natural language conversation*. Technical Report 545 and Ph.D. Thesis, University of Rochester.