Proceedings of DiSS-LPSS Joint Workshop 2010,
25-26 September 2010, Tokyo, Japan, pp. 55-58.

Towards a precise model of turn-taking for conversation: A quantitative analysis of overlapped utterances

Hanae Koiso[1], *Yasuharu Den*[2]

[1]National Institute for Japanese Language and Linguistics, Japan
[2]Faculty of Letters, Chiba University, Japan
koiso@ninjal.ac.jp, den@cogsci.l.chiba-u.ac.jp

Abstract

In this paper, we present the outline of a new model of turn-taking that is applicable not only to smooth transitions but also to transitions involving overlapping speech. We identify acoustic, prosodic, and syntactic cues in overlapped utterances that elicit early initiation of a next turn, based on a quantitative analysis of Japanese three-party conversations, proposing a model for predicting a turn's completion in an incremental fashion using sources from units at multiple levels.

Index Terms: turn-taking, overlapped utterances, incremental processing

1. Introduction

How the context of turn-taking is characterized by syntactic, intonational, pragmatic, and non-verbal features is one of the central questions discussed so far in various fields such as conversation analysis [1], social psychology [2], interactional linguistics [3], cognitive science [4, 5] and computational linguistics [6]. Duncan & Fiske [2] proposed a turn-taking model in which participants regulate their turns by exchanging various cues that indicate their turn-taking states. They found that there was a strong correlation between these cues and the probability of speaker-shifts. Koiso & Den [5] and Gravano & Hirschberg [6], using Japanese and English task-oriented dialogs, respectively, analyzed the relationship of several acoustic, prosodic, and syntactic features to turn-taking, showing that some features of speaker's speech tended to occur distinctively at turn transitions, and that the probability of the occurrence of speaker-shifts increased with the number of these features.

Although these studies have established prolegomena to precise modeling of turn-taking phenomena, they seem to have the following problems:

1. They assumed particular units of analysis and used features involved in them, usually ones from the final portion of these units. Few discussions have been made on the adequacy of using such units in turn-taking study.
2. They formulated the task as discrimination between two types of units, i.e., those followed by speaker-shift and those followed by continuation of the same speaker. They confused two tasks that should be distinct from each other: i) completion/non-completion of a unit for exchange of speakership and ii) transfer/non-transfer of speakership upon completion of such a unit.
3. They lacked perspective on on-line processing. Turn transitions are usually very rapid, and large part of the turn-taking cues so far identified at the end of units may not be available to participants acting in real time.

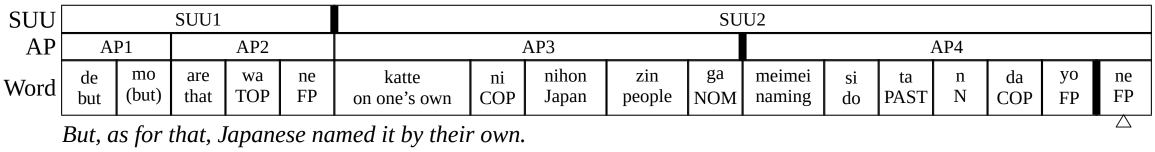
In this paper, we present the outline of a new model of turn-taking that resolves these problems. In particular, we take the third problem seriously. In order to develop a model that is applicable not only to smooth transitions but also to transitions involving overlapping speech, we investigate acoustic, prosodic, and syntactic cues that elicit overlapping utterances. In the end, we propose a model for predicting a turn's completion in an incremental fashion using sources from units at multiple levels.

2. A precise model

A solution to the above problems may be found in an influential work by Sacks, Schegloff, & Jefferson [7] on turn-taking for conversation. They described a turn-taking system that consists of two sub-components: the turn-constructional component and the turn-allocation component. The turn-constructional component concerns with the construction of basic units of interaction to which turns are allotted. They called such units *turn-constructional units*, and emphasized that turn-constructional units are *projectable* in the sense that the unit under way can project what it will take for such type of unit to be completed. The turn-allocation component describes two ways of allocating a new turn to one party: (a) current speaker's selecting next speaker and (b) self-selection by next speaker. A set of rules are said to operate at every possible completion point of a turn-constructional unit, or *transition relevance place* (TRP). There are three options in the rule-set: (a) the turn is transferred to next speaker by the use of a 'current speaker selects next' technique; (b) the turn is transferred to next speaker by self-selection; and (c) the turn is continued by the current speaker.

The relationship between the previous works on turn-taking cues and Sacks, Schegloff, & Jefferson's model may be depicted as follows. In the turn-taking cue studies [4, 6], target units are discriminated into two categories, i.e., the speaker-shift and the continuation categories. The former category contains cases in which the options (a) and (b) of the turn-taking rules are applied. The latter category, on the other hand, contains not only cases where rule option (c) is applied at TRPs, but also cases where target units are in the midst of turn-constructional units, since the units of analysis employed in these studies do not necessarily coincide with turn-constructional units. The cues of rule option (c) may be substantially different from the cues of mid-unit; instead, it may resemble the cues of rule options (a) and (b). Thus, in order to construct a proper model of turn-taking, we need, in the first place, to find cues to distinguish TRPs from mid-unit positions and, then, to find cues to predict which rule option is applied at a given TRP.

In this study, we focus on turn-taking data involving only rule option (b), as a first step toward a precise model of turn-



SUU	SUU1					SUU2											
AP	AP1		AP2			AP3					AP4						
Word	de but	mo (but)	are that	wa TOP	ne FP	katte on one's own	ni COP	nihon Japan	zin people	ga NOM	meimei naming	si do	ta PAST	n N	da COP	yo FP	ne FP
																	△

But, as for that, Japanese named it by their own.

Figure 1: Example of most recent unit boundaries (thick lines) at the SUU, AP, and word levels. The triangle indicates the position of the start of the overlapping utterance. The following glosses are used: COP: copula, FP: final particle, N: nominalizer, NOM: nominative case marker, PAST: past tense, TOP: topic marker

taking for conversation. We take seriously the third problem above, i.e., lack of perspective on on-line processing. In our conversation data, the distribution of transition times between consecutive utterance-units involving application of rule option (b) peaks at 0 to 200 ms. This means that participants are very likely to start a new turn before they have fully processed the turn-taking cues located at the end of the previous turn. Furthermore, 44% of the data involve overlapping of consecutive utterances. Analyzing these overlapping cases would lead us to a solution to how to model turn-taking phenomena coping with projectability of turn-construction. In what follows, we try to seek such model based on a quantitative analysis of overlapped utterances in Japanese conversations.

3. Method

3.1. Dialog data

Twelve dialogs, produced by different 36 speakers, were selected from the *Chiba three-party conversation corpus* [8]. The Chiba corpus is a collection of casual conversations among three participants. The participants of each dialog were friends on campus. Each dialog is about 10 minutes long, and a total of 2 hours of dialogs were used in this study.

3.2. Annotation of units

As an analog of turn-constructional units, *long utterance-units* (LUUs) were employed, which have been proposed in our separate study [9]. LUUs are recognized mainly by syntactic and pragmatic completion of an utterance. Fragments, due to false starts and self-interrupted speech, and response tokens, such as *un* and *hee*, which are produced by a hearer during a speaker's turn, were separately labeled. Although LUUs are different from turn-constructional units in their extension and concept, their completion points were found to be nicely coordinated with new turn's start [9]. LUUs used in the current study were labeled by one of the authors.

In order to precisely analyze features of overlapped utterances, identifying what acoustic, prosodic, and syntactic cues elicit early initiation of a new turn, three other units were introduced: words, accentual phrases, and short utterance-units. Word boundaries and parts of speech were identified by hand by an expert annotator, and every word boundary was manually time-aligned to the speech sound by another expert annotator. Accentual phrases (APs) were annotated based on the X-JToBI scheme [10]. The labeling scheme gives us information about final boundary tones and break indices. A boundary with break index = 2 or greater represents the boundary of an AP. The annotation of X-JToBI was performed by an expert annotator, and crosschecked by one of the authors. Short utterance-units (SUUs) [9] are sub-components of LUUs, and defined as a stretch of speech followed by a pause longer than 100 ms and/or by a stronger intonation boundary, i.e., a boundary with break index = 3. They correspond approximately to intonational phrases. SUUs were automatically extracted based on the time-stamps in the word annotation and the break index labels in the X-JToBI annotation.

3.3. Acoustic, prosodic, and syntactic features of the units

Six acoustic features were extracted from regions of words, APs, and SUUs: the mean F0, the minimum F0, the maximum F0, the mean power, the maximum power, and the average mora duration. The F0 values were extracted using WaveSurfer and converted, after log-transformation, into z-scores on a per-speaker basis. The power values were also extracted using WaveSurfer and converted into z-scores. The average mora durations were calculated by dividing the duration of the unit by the number of morae in that unit, and converted into deviations from the overall average mora duration of the speaker.

In addition to these numerically-valued features, the final boundary tone and the part-of-speech tag were also used as categorical features of APs and words, respectively.

3.4. Annotation of turn-transition types

In order to extract the data for the analysis of overlapping speech, turn-transition types were annotated by hand by one of the authors [9]. First, for each dialog, only segments in which the turn-taking rules were in operation were selected, the remaining portion, such as story-telling and explanation-giving, being discarded. Then, for each LUU in these segments, its antecedent unit was identified, by making reference to the time information and the linguistic content, and classified into three types according to the rule option being employed.

In annotating with turn-transition types, the annotator ignored whether or not the current unit was properly launched at a transition-relevance place of the antecedent unit. By this reason, the data may contain cases where the overlapping utterance interfered with the previous turn, violating the turn-taking rules. We, thus, adopted only instances in which the overlapping utterance started within the last SUU of the antecedent unit. Among 1427 LUU pairs involving rule option (b), 412 instances were overlapping speech. Of them, 354 instances were selected by the above criteria for further analysis.

3.5. Categorization of units in overlapped utterances

Now, we introduce the notion of *most recent unit boundaries* (MRUBs), which characterize the locations at which some cues for overlapping speech may be present. MRUBs were defined relative to each level of unit, SUUs, APs, and words. For unit level u, the MRUB, denoted as MRUB(u), was defined as the boundary of the most recently completed unit at level u that occurred prior to the start of the overlapping utterance. Consider the example in Figure 1. The overlapping utterance starts at the final word *ne* (indicated by the triangle mark). Thus, the end-

Table 1: Summary of the results. Each row represents an acoustic feature. Each column represents a comparison in which units at a certain level and at a certain location were compared. (ns: not significant, $^{+}$: $p < .1$, *: $p < .05$, **: $p < .01$)

	SUU boundary `000` vs `111`, `110`, `100` (n = 178, 55, 36, 67)			AP boundary `00` vs `11`, `10` (n = 33, 23, 44)		Word boundary `0` vs `1` (n = 93, 75)
	SUU	AP	Word	AP	Word	Word
`F0.mean`	ns	ns	ns	ns	ns	ns
`F0.min`	$\texttt{111}^{*} < \texttt{000}$	$\texttt{111}^{**} < \texttt{000}$	ns	ns	ns	ns
`F0.max`	ns	ns	ns	ns	ns	ns
`pwr.mean`	ns	ns	ns	ns	ns	ns
`pwr.max`	ns	ns	ns	ns	$\texttt{10}^{*} < \texttt{00}$	ns
`AMD`	$\texttt{110}^{**}, \texttt{100}^{+} < \texttt{000}$	$\texttt{110}^{**} < \texttt{000}$	$\texttt{110}^{**}, \texttt{100}^{*} < \texttt{000}$	ns	ns	$\texttt{0} < \texttt{1}^{+}$

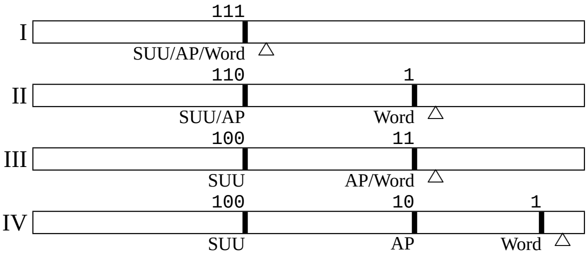


Figure 2: Schematic diagram of the patterns of most recent unit boundaries (thick lines) and their categories. The triangles indicate the positions of the start of overlapping utterances.

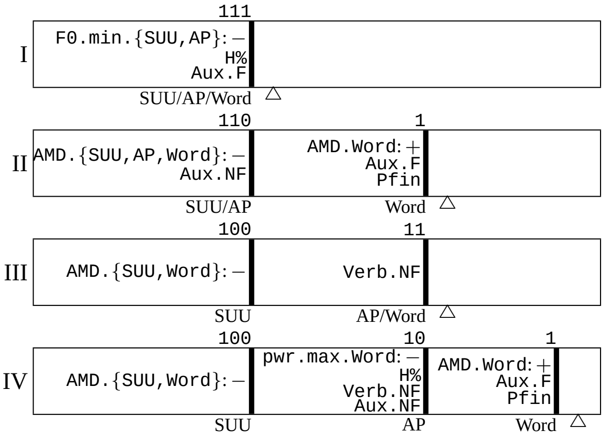


Figure 3: Characteristics of most recent unit boundaries. $+/-$ for an acoustic feature means that the feature exhibits a higher/lower value at that boundary.

ing boundaries of SUU1, AP3, and word *yo* are MRUB(SUU), MRUB(AP), and MRUB(Word), respectively.

There can be four patterns according to whether or not MRUB(u)s are identical for different us (see Figure 2).

Pattern I	MRUB(SUU) = MRUB(AP) = MRUB(Word)
Pattern II	MRUB(SUU) = MRUB(AP) ≠ MRUB(Word)
Pattern III	MRUB(SUU) ≠ MRUB(AP) = MRUB(Word)
Pattern IV	MRUB(SUU) ≠ MRUB(AP) ≠ MRUB(Word)

These patterns reflect the difference in the timing of the next turn's initiation. For instance, in pattern I, where MRUB(SUU) and MRUB(Word) are coincident with each other, the next speaker initiates her overlapping utterance immediately after MRUB(SUU), i.e., within a word immediately following it. In pattern IV, where MRUB(SUU) and MRUB(Word) are different, by contrast, the next turn is initiated within the next SUU after MRUB(SUU) but later in that unit, at least one AP plus one word later.

Next, we assign categories in bit-string forms to all boundaries of SUUs, APs, and words. An SUU boundary may be either MRUB(SUU) or not. There are two cases when an SUU boundary is MRUB(SUU) depending on whether or not it is also MRUB(AP). Moreover, when an SUU boundary is simultaneously MRUB(SUU) and MRUB(AP), it can be further categorized according to whether or not it is MRUB(Word) as well. Thus, an SUU boundary is coded as `111` when it is MRUB(SUU), MRUB(AP), and MRUB(Word), `110` when it is MRUB(SUU) and MRUB(AP) but not MRUB(Word), `100` when it is MRUB(SUU) but not MRUB(AP) nor MRUB(Word), and `000` when it is none of MRUB(SUU), MRUB(AP), or MRUB(Word). In the same way, the categories of AP boundaries and word boundaries are coded by double-digit and single-digit bit-strings, respectively. (see Figure 2; categories `000`, `00`, and `0` are not shown.)

We are interested in looking at cues of units at multiple levels. We, thus, utilized features of all the units at lower levels whose endings coincided with unit boundaries at upper levels. That is, the features of the last AP and the last word of an SUU, whose endings were coincident with that SUU, were also considered in the analysis. Similarly, the features of the last word of an AP were included in the analysis. We compare the acoustic, prosodic, and syntactic features of units at MRUBs to those occurring at non-MRUBs in order to find out the characteristics of most recent unit boundaries.

3.6. Statistical analysis

To see the difference of the features among the boundary categories, we applied liner mixed-effects models with speaker as random effect, and obtained p-values using Markov Chain Monte Carlo (MCMC) sampling implemented in `lme4` and `languageR` packages of the R language. The comparison was conducted separately for each unit level.

4. Results

Table 1 summarizes the results of the statistical tests for acoustic features. Those acoustic features that were found to be significantly different between at MRUBs and at non-MRUBs are shown in Figure 3. The picture also shows the prosodic and syntactic features at MRUBs that were considerably frequent compared with those observed at non-MRUBs.

The results can be stated as follows. In pattern I, the next speaker starts overlapping immediately after the MRUB, when it is ended with a lower minimum F0 value (F0.min:−), a rising tone (H%) and/or an auxiliary verb in conclusive form (Aux.F). In pattern II, the next speaker waits for some words, but not until the completion of the next AP, before initiating the overlapping utterance, when the most recently completed SUU, as well as the last AP and the last word of that SUU, is associated with a faster speaking rate (AMD:−) and an auxiliary verb in non-conclusive form (Aux.NF); the overlapping utterance is started immediately after a word with a slower speaking rate (AMD:+), an auxiliary verb in conclusive form, and/or a sentence-final particle (Pfin). In pattern III, the next speaker waits for some APs, but not until the completion of the next SUU, when the most recently completed SUU, in particular its last word, is spoken more quickly; the next turn is started at the boundary of an AP that is ended typically with a verb/adjective in non-conclusive form (Verb.NF). In pattern IV, the next speaker's response is postponed for some words after the most recently completed AP, which appears in a different configuration as pattern III, with a lower maximum power value (pwr.max:−), a rising tone, and a verb/adjective and an auxiliary verb in non-conclusive form (Aux.NF); the responded word exhibits a slower speaking rate and are typically an auxiliary verb in conclusive form or a sentence-final particle.

5. Discussion

Looking at characteristic features at most recent SUU boundaries in patterns II, III, and IV, fast speaking rate is commonly observed. Given that the speech generally tends to accelerate within individual sentences [11], this finding suggests that speaking rate could potentially project an upcoming completion of the current turn. There is, however, difference of the span for which the next speaker waits before initiating a next turn; the next speaker tends to start her turn within the AP immediately following the most recent SUU boundary in pattern II, while she tends to start speaking within a later AP in patterns III and IV. This difference may be attributed to the marked feature that is distinctively observed in pattern II, i.e., auxiliary verbs in non-conclusive form. Considering that this feature is also peculiar to the most recent AP boundary in pattern IV, where the next speaker starts speaking within the AP immediately after the most recent AP boundary, as in pattern II, auxiliary verbs in non-conclusive form may project a turn's completion in a shorter span.

Focusing now on the features marking most recent word boundaries, we see that auxiliary verbs in conclusive form are common in patterns I, II, and IV. In Japanese, one or more auxiliary verbs can follow a predicate and they, as well as sentence-final particles following them, are considered as providing a space for turn-taking [12]. Tanaka [12] stated that the next speaker can start her turn within such *utterance-final elements*. Her claim is consistent with our findings that the next speaker sometimes initiates the response right after an auxiliary verb even when the turn has not yet been fully completed.

In this way, sources from units at multiple levels, i.e., SUUs, APs and words, can be cues for overlapping speech. They project an upcoming completion of the current turn in a shorter or a longer span. This property of turn-taking cues in overlapped utterances leads us to a model of predicting a turn's completion in an incremental fashion. Our pattern IV best illustrates such a view (see Figure 4). An SUU bearing a fast speaking rate at its boundary projects the arrival of the final SUU, within which an AP with certain characteristics such as lower maximum power at its boundary projects the arrival of the final AP, within which a word included in a class of utterance-final elements finally projects the completion of the turn immediately afterward. We will develop such kind of incremental processing model of turn-construction in the future research.

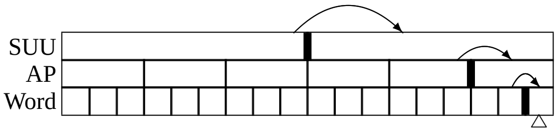


Figure 4: An incremental processing model of turn-construction. The arcs indicate projection at multiple levels.

6. References

[1] E. A. Schegloff, "Turn organization: One intersection of grammar and interaction," in *Interaction and grammar* (E. Ochs, E. A. Schegloff, and S. A. Thompson, eds.), pp. 52–133, Cambridge: Cambridge University Press, 1996.

[2] S. Duncan Jr. and D. W. Fiske, *Face-to-face interaction: Research, methods, and theory*. Hillsdale, NJ: Lawrence Erlbaum, 1977.

[3] C. E. Ford and S. A. Thompson, "Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns," in *Interaction and grammar* (E. Ochs, E. A. Schegloff, and S. A. Thompson, eds.), pp. 134–184, Cambridge: Cambridge University Press, 1996.

[4] H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den, "An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task dialogs," *Language and Speech*, vol. 41, pp. 295–321, 1998.

[5] H. Koiso and Y. Den, "How is the smooth transition between speakers realized?: Consideration based on an analysis of a spoken dialogue corpus (in Japanese)," *Cognitive Studies: Bulletin of the Japanese Cognitive Science Society*, vol. 7, no. 1, pp. 93–106, 2000.

[6] A. Gravano and J. Hirschberg, "Turn-yielding cues in task-oriented dialogue," in *Proc. 10th SIGDIAL*, pp. 253–261, 2009.

[7] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, pp. 696–735, 1974.

[8] Y. Den and M. Enomoto, "A scientific approach to conversational informatics: Description, analysis, and modeling of human conversation," in *Conversational informatics: An engineering approach* (T. Nishida, ed.), pp. 307–330, Hoboken, NJ: John Wiley & Sons, 2007.

[9] Y. Den, H. Koiso, T. Maruyama, K. Maekawa, K. Takanashi, M. Enomoto, and N. Yoshida, "Two-level annotation of utterance-units in Japanese dialogs: An empirically emerged scheme," in *Proc. 7th LREC*, pp. 2103–2110, 2010.

[10] K. Maekawa, H. Kikuchi, Y. Igarashi, and J. J. Venditti, "X-JToBI: An extended J_ToBI for spontaneous speech," in *Proc. 7th ICSLP*, pp. 1545–1548, 2002.

[11] R. S. Brubaker, "Rate and pause characteristics of oral reading," *Journal of Psycholinguistic Research*, vol. 1, pp. 141–147, 1972.

[12] H. Tanaka, *Turn-taking in Japanese conversation: A study in grammar and interaction*. Amsterdam: John Benjamins, 1999.