A PHONETIC INVESTIGATION OF TURN-TAKING CUES AT MULTIPLE UNIT-LEVELS IN JAPANESE CONVERSATION

Hanae Koiso[a] *& Yasuharu Den*[b]

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

ABSTRACT

In this paper, we investigate acoustic, prosodic, and syntactic cues at multiple unit-levels for turn-taking in Japanese conversation, proposing an incremental and hierarchical model of turn-projection, which is applicable to both overlapping and non-overlapping speech. Based on a quantitative analysis of Japanese three-party conversations, we identify several turn-taking cues that are located earlier in the utterance than utterance-final position. We discuss similarities and differences between overlapping and non-overlapping turn-transitions.

Keywords: turn-taking cues, multiple unit-levels, incremental and hierarchical model, speech overlap

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 in various fields such as conversation analysis [11], social psychology [3], interactional linguistics [4, 12], speech science [8] and computational linguistics [6]. Many studies have identified various cues, located at the end of utterances, for speaker-change, including rising or falling intonation, rapid drop in loudness, sound-stretches, and utterance-final elements.

Although these studies have established prolegomena to precise modeling of turn-taking phenomena, they lacked perspective on on-line processing. Turn-transitions are usually very rapid; the mode of the frequency distribution of transition times between consecutive turns in our conversation data is between 0 ms and 200 ms. Considering that reaction to a verbal stimulus would need roughly 200 ms [5], this fact implies that large part of the turn-taking cues so far identified at the end of utterances may not be available to participants acting in real time.

To conquer this problem, we proposed a new model of turn-taking, in which completion point of an utterance is projected incrementally and hierarchically using acoustic, prosodic, and syntactic cues at multiple levels including intonation unit (short utterance-unit, or SUU, in our terminology), accentual phrase, and word [7]. We focused specifically on turn-transitions involving overlapping speech, where the next turn is initiated prior to the completion of the current turn and, thus, utterance-final cues are evidently unavailable to the next speaker. We found that an SUU bearing a fast speaking rate marks the following SUU as the final one, within which an accentual phrase with certain characteristics, such as decreased maximum power, predicts the final accentual phrase in its subsequent position, within which a word included in a class of utterance-final elements [12] finally invites start of the (overlapping) next turn immediately afterward.

The current study extends our previous study in that smooth turn-transitions, involving no speech overlap, are also targeted by slightly modifying our model. We show that several cues at multiple unit-levels incrementally project turn's completion in both overlapped and non-overlapped utterances. We also show differences between the two cases.

2. AN INCREMENTAL AND HIERARCHICAL MODEL

The key idea of Koiso & Den's [7] incremental and hierarchical model (the KD model) comes from the notion of the *projectability* of turn-constructional units [10, 11]. In their influential work on turn-taking for conversation, Sacks, Schegloff, & Jefferson [10] proposed *turn-constructional units* as basic units of interaction, to which turns are allotted with reference to a set of turn-taking rules. Various syntactic units, i.e., sentence, clause, phrase, and word, may constitute a turn-constructional unit depending on a context. Sacks et al. 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. Hearers can utilize this property of turn-constructional unit in predicting its completion point.

Figure1: An incremental and hierarchical model of turn-projection. The following glosses are used: COP: copula, FP: final particle, N: nominalizer, NOM: nominative case marker, PAST: past tense, TOP: topic marker.

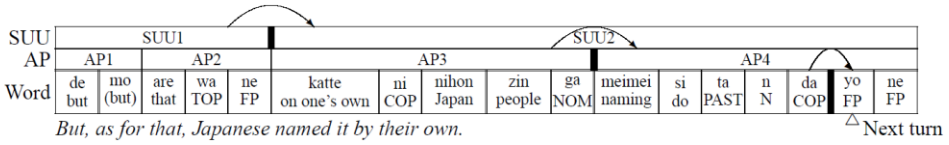


Figure 1 illustrates the basic idea of the KD model. The triangle mark shows the starting point of the next turn, and, at each unit-level, the boundary indicated by a thick line corresponds to the ending point of the *penultimate* unit at that level. The arcs represent projection from the penultimate units to the final units at three different levels. Note that penultimate units are defined relative to the actual start of a new turn. Therefore, the penultimate word, for example, may not be the second to the last word in the utterance but may appear earlier, as *da* in Figure 1.

In this model, the central task is to identify features that distinguish the penultimate unit from the prior units at each level. This presupposes that turn-taking cues could be found earlier in the utterance than utterance-final position. The presence of such *early* cues has been suggested by several studies in conversation analysis [11, 12] as well as by a recent study in computational linguistics [6].

One drawback of the KD model is that it is not directly applicable to smooth turn-transitions, although it aimed at developing a general model that covers both overlapping and non-overlapping speech. In smooth transitions, the start of a next turn comes after the completion of the current utterance, and, thus, the ending points of the penultimate units at all levels coincide with the utterance end.

To solve this problem, we define the penultimate units relative to the *reaction onset*, instead of the starting point of a next turn. The reaction onset is the point at which the next speaker has recognized the cues for turn's completion and initiates the execution of her speech plan. According to the experimental result on the simple reaction time to a verbal stimulus [5], the reaction onset would be located about 200 ms prior to the actual start of a next turn. Therefore, for cases where the transition times fall within a range between 0 ms and 200 ms, which are frequent in our conversation data, the same schema as in Figure 1 can be applied, now regarding the point indicated by the triangle mark as the reaction onset.

In this paper, we compare several acoustic, prosodic, and syntactic features of the penultimate units with those of the prior units at three levels, i.e., short utterance-unit, accentual phrase, and word, in both overlapped and non-overlapped utterances.

3. METHOD

3.1. Dialog data

Twelve dialogs, produced by 36 native speakers of Japanese, were selected from the *Chiba three-party conversation corpus* [1]. The Chiba corpus is a collection of casual conversations among 3 participants. 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) [2] were employed. LUUs were identified mainly by syntactic and pragmatic completion of an utterance. Fragments and response tokens were separately labeled.

In order to investigate features at multiple levels, three other units were annotated: short utterance-unit (SUU) [2], accentual phrase (AP), and word. SUUs are sub-components of LUUs, and defined as a stretch of speech followed by a pause longer than 100 ms and/or by a strong intonational disjuncture. They roughly correspond to intonation units. APs were annotated based on the X-JToBI scheme [9], which gives us information about final boundary tones and break indices. A boundary with break index = 2 or greater represents the boundary of an AP. Word boundaries and parts of speech were labeled by hand, and every word boundary was manually time-aligned to the speech sound.

3.3. Acoustic, prosodic, and syntactic features

Seven acoustic features were extracted from regions of SUUs, APs, and words: the mean/minimum/maximum (log) F0 values, the mean/maximum power values, the duration, and the average mora duration. The F0 and power values were converted into z-scores on per-speaker basis. 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 features of APs and words, respectively. Furthermore, to roughly represent the complexity of units, the number of constituent APs/words contained in an SUU/AP was employed.

3.4. Annotation of turn-transitions

For each dialog, only dialog 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 utterance, which immediately preceded that LUU, was identified, by making reference to the time information and the linguistic content. Only transitions from completed utterances to utterances constituting a genuine turn were considered in the current analysis, excluding cases where antecedent utterances were followed by reactive tokens.

3.5. Annotation of penultimate units

Penultimate units were annotated automatically in the following way. First, for each transition, the reaction onset was determined as the point 200 ms prior to the start of the next turn. Then, the penultimate unit at each unit-level was identified as the last completed unit at that level appearing before the reaction onset. All the units, at that level, prior to the penultimate unit, except for those occurring before the penultimate unit at a higher level, were treated as the prior units.

3.6. Data selection

In three-party conversations, two hearers may sometimes start next utterances simultaneously. To avoid a possible discrepancy between turn-taking cues for respective next-speakers, those cases involving simultaneous starts of next turn were excluded.

For overlapping speech, ‘premature’ start of next turn was also excluded, since such turn may be produced without relying on turn-taking cues. When the next turn was initiated before the start of the predicate of the current utterance, they were considered as ‘premature.’ For non-overlapping speech, only cases where the transition times fall within a range of 0 ms and 200 ms were retained.

Finally, due to the nature of the KD model, analysis can be conducted only when the current utterance is composed of more than one SUU; LUUs consisting of a single SUU were, thus, eliminated. These selection steps left us 161 overlapped LUUs and 116 non-overlapped LUUs for the subsequent analysis.

3.7. Statistical analysis

To see the difference of the feature values between the penultimate and the prior units, we applied linear mixed-effects models with speaker as random effect, and obtained p-values using Markov Chain Monte Carlo (MCMC) sampling.

The comparison was conducted separately for each unit-level. At the SUU level, not only features of the SUU itself but also those of the initial and final APs/words were examined. Similarly, at the AP level, features of the initial/final words were taken into account in addition to those of the AP itself.

4. RESULTS

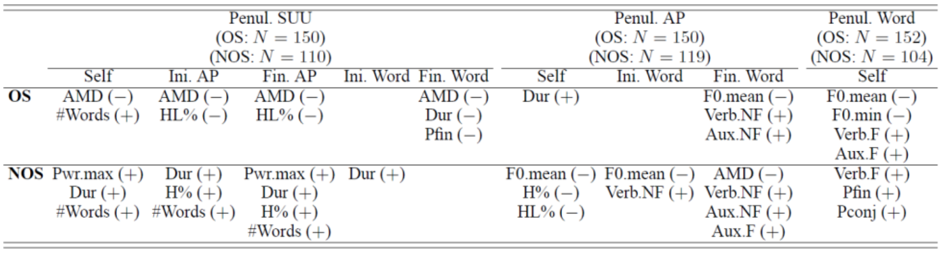
Table 1 summarizes the results. For numerically-valued features, i.e., acoustic and complexity features, those that were found to be significantly greater (+) or smaller (−) (at 5% level) in the penultimate unit than in the prior units are shown. For prosodic and syntactic features, those that were considerably frequent (+) or rare (−) in the penultimate unit compared with the prior units are also shown.

Focusing on similarities and differences between overlapping and non-overlapping speech, the results can be stated as follows. For the penultimate SUUs, the number of constituent words (#Words) tended to be greater than that for the prior SUUs in both the overlapped and non-overlapped utterances. In the overlapped utterances, the average mora durations (AMDs) tended to be shorter in the penultimate position than in the prior positions, while in the non-overlapped utterances, no tendency was observed.

At the AP level, the mean F0 values (F0.mean) tended to be lower in the penultimate APs than in the prior APs in both overlapped and non-overlapped utterances. Also observed in the two cases were frequent occurrences of predicates and auxiliary verbs in non-conclusive form (Verb.NF and Aux.NF). A shorter AMD, on the other hand, was observed only in the non-overlapping case.

At the word level, frequent use of utterance-final elements, i.e., auxiliary verbs in conclusive form (Aux.F) and final particles (Pfin), was characteristics of the penultimate units, which was observed in both overlapped and non-overlapped utterances.

Table 1: Summary of the results (OS: Overlapping speech, NOS: Non-overlapping speech).



	Penul. SUU (OS: $N = 150$) (NOS: $N = 110$)					Penul. AP (OS: $N = 150$) (NOS: $N = 119$)			Penul. Word (OS: $N = 152$) (NOS: $N = 104$)
	Self	Ini. AP	Fin. AP	Ini. Word	Fin. Word	Self	Ini. Word	Fin. Word	Self
OS	AMD (−) #Words (+)	AMD (−) HL% (−)	AMD (−) HL% (−)		AMD (−) Dur (−) Pfin (−)	Dur (+)		F0.mean (−) Verb.NF (+) Aux.NF (+)	F0.mean (−) F0.min (−) Verb.F (+) Aux.F (+)
NOS	Pwr.max (+) Dur (+) #Words (+)	Dur (+) H% (+) #Words (+)	Pwr.max (+) Dur (+) H% (+) #Words (+)	Dur (+)		F0.mean (−) H% (−) HL% (−)	F0.mean (−) Verb.NF (+)	AMD (−) Verb.NF (+) Aux.NF (+) Aux.F (+)	Verb.F (+) Pfin (+) Pconj (+)

5. DISSCUSION

Some features were commonly observed in overlapping and non-overlapping speech. They are:

1. increased number of words contained in the penultimate SUU;
2. a reduced mean F0 in the penultimate AP; and
3. frequent use of utterance-final elements in the penultimate word.

The number of words in an SUU would roughly reflect the amount of information being delivered at a time. The increase in this number may signal the arrival of the most important part of the utterance. If so, it would project the completion of the current utterance within a measurable period of time. A reduced mean F0 in an AP may also be a decisive cue indicating that the completion point is approaching. Utterance-final elements in Japanese clearly mark a pre-possible completion of the current turn [12]. The space after these elements is open for turn-taking.

From these considerations, the above three features together constitute turn-taking cues in an incremental and hierarchical fashion. That is, increase in the number of constituent words in an SUU marks the following SUU as the final one, within which an accentual phrase with a reduced mean F0 predicts the final accentual phrase in its subsequent position, within which a word included in a class of utterance-final elements finally projects the beginning of a transition space immediately afterward. What is striking is that this scheme, which we originally found in a slightly different form for overlapped utterances [7], is also applicable to non-overlapped utterances.

There are, however, some different features between overlapping and non-overlapping speech. One of these differences resides in the average mora durations. Decreased average mora duration is characteristics of the penultimate *SUUs* in the overlapped utterances, but is characteristics of the penultimate *APs* in the non-overlapped utterances. Acceleration of speaking rate at the penultimate SUUs was also reported in our previous study, and the present finding is a replication of these results. In the present case, however, a similar tendency was also observed for non-overlapped utterances, but at later position. Considering that the start of a next turn is earlier in overlapping speech than in non-overlapping speech, it is suggested that acceleration of speaking rate may directly invite the start of a next turn within a certain period of time. This possibility should be further explored in future research.

6. REFERENCES

[1] Den, Y., Enomoto, M. 2007. A scientific approach to conversational informatics: Description, analysis, and modeling of human conversation. In Nishida, T. (ed.), *Conversational Informatics: An Engineering Approach*. Hoboken, NJ: John Wiley & Sons, 307-330.

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

[3] Duncan Jr., S., Fiske, D.W. 1977. *Face-to-face Interaction: Research, Methods, and Theory*. Hillsdale: Lawrence Erlbaum.

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

[5] Fry, D.B. 1975. Simple reaction-times to speech and non-speech stimuli. *Cortex* 11, 355-360.

[6] Gravano, A., Hirschberg, J. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech and Lang.* 25, 601-634.

[7] Koiso, H., Den, Y. 2010. Towards a precise model of turn-taking for conversation: A quantitative analysis of overlapped utterances. *Proc. DiSS-LPSS Jt Workshop 2010* Tokyo, 55-58.

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

[9] Maekawa, K., Kikuchi, H., Igarashi, Y., Venditti, J.J. 2002. X-JToBI: An extended J_ToBI for spontaneous speech. *Proc. 7th ICSLP* Denver, 1545-1548.

[10] Sacks, H., Schegloff, E.A., Jefferson, G. 1974. A simplest systematics for the organization of turntaking for conversation. *Language* 50, 696-735.

[11] Schegloff, E.A. 1996. Turn organization: One intersection of grammar and interaction. In Ochs, E., Schegloff, E.A., Thompson, S.A. (eds.), *Interaction and Grammar*. Cambridge: Cambridge University Press, 52-133.

[12] Tanaka, H. 1999. *Turn-taking in Japanese Conversation: A Study in Grammar and Interaction*. Amsterdam: John Benjamins.