Ten Prosodic Patterns of Turn-Taking in Japanese Conversation

Nigel G. Ward
University of Texas at El Paso
nigelward@acm.org

Abstract
In spoken dialog, proper management of turn taking is a cornerstone of effective communication. In many languages, including Japanese, prosody plays an important role, but previous work has described only a few aspects of this. Based on a systematic analysis of corpus data using automatic tools and close listening, this paper presents a more comprehensive account, including descriptions of the prosody and functions of ten patterns of turn-taking in Japanese, four of which have not been previously described, including some not observed in other languages.

Index Terms: dialog, prosodic constructions, interaction, joint behaviors, backchanneling, fillers, cross-language comparisons, non-universals

1. Motivation
In spoken dialog, proper management of turn taking is a cornerstone of effective communication. Elucidating how this is done has both practical value and scientific interest [1]. In Japanese in particular, turn-taking has been the topic of much work. Prosody is important, as noted even in work which focuses on the social and speech act aspects of turn-taking [2, 3, 4]. Studies of the prosody of turn taking often start with the Brady model or the Sacks, Schegeloff and Jefferson model [5, 6], and apply quantitative methods to identify correlates of the transitions, such as selected prosodic properties of turn holding versus turn yielding [7, 8, 9, 10], but also those involved in backchanneling, filler production, overlaps and other phenomena [11, 12, 13, 14, 15, 16, 17]. In recent years an alternative approach has been to eschew description, in favor of black-box models learned from data, with or without annotations [18, 19, 20, 21, 22, 23, 24]. While such models can cover all surface-level turn-taking phenomena, they have contributed little to our understanding of how turn-taking is managed. Thus, despite substantial work, our understanding of of turn-taking in Japanese has been very limited.

Thus the aim of this paper is to provide a more comprehensive description of the prosodic patterns that speakers of Japanese use as they manage who speaks when.

2. Model and Discovery Method
In this paper I take a prosodic constructions approach [25, 26]. Specifically I assume that much of the prosody of interaction can be represented by prosodic constructions, each of which a) is a temporal configuration of prosodic features, b) has a meaning or function, c) is not necessarily closely aligned with words, d) can be present to a greater or lesser degree, and e) can appear superimposed with other constructions [27, 28]. An advantage of modeling prosody in this way is that it enables the automatic discovery, from data, of candidate prosodic configurations. This is done by applying Principal Component Analysis (PCA) to hundreds of thousands of timepoints sampled from dialogs. The prosody in the vicinity of each timepoint is characterized by the values of multiple prosodic features, computed over windows at various temporal offsets, together spanning about 3 seconds of context, for both speakers. Here I used 10 base prosodic features, namely intensity, pitch lowness, pitch highness, creakiness, pitch narrowness, pitch wideness, lengthening, enunciation, reduction, and peak disalignment [29]. Across all windows and both speakers, there were a total of 212 features. The output of the PCA is a set of dimensions, each specifying a temporal pattern of feature values. Each dimension yields two candidate constructions, one where the values on the dimension are positive and one where they are negative, or only one candidate in the degenerate case where the behaviors of the two speakers in the dimension are symmetric. By a simple application of the dimension loadings to data, it is possible to identify timepoints where each candidate was strongly present. (Full details on the features, the processing, and the procedure appear elsewhere [28].)

These timepoints were then used to help understand the nature and role of the construction, in two ways. First, I listened to the contexts of many such timepoints to inductively infer each construction’s meaning and function. Second, I examined statistics on lexical items characteristic of each time range of each construction. As the data lacked word-level timestamps, these statistics were not precise, but nevertheless interesting tendencies emerged. Final inference of the meaning of each construction was thus based on three sources of information: loadings, examples, and lexical statistics. In each case these were mutually confirming, and thus I am confident in the validity of each of the descriptions below.

Among the many differences with previous approaches, two are worth highlighting. First, datapoints are sampled everywhere, to avoid the limiting assumption that turn-taking considerations and actions are relevant only at a few points and to avoid the practical difficulty of reliably identifying such points, and also to potentially support development of dialog systems with continuous turn-taking [30, 31]. Second, the large feature set enables discovery of patterns that span several seconds and those that involve both participants.

3. Data
From the Callhome Japanese corpus, a collection of long distance telephone conversations between family members, I selected portions of 14 conversations, chosen to avoid regional dialects and poor recording quality, for a total of 128 minutes. Sampling both sides every 20 milliseconds gave 768 000 data points for the PCA. Interpretation of candidate constructions was done using both this data, a smaller corpus [12], and a tiny corpus, with the illustrations below taken from the latter to avoid permission issues. The lexical statistics were computed over 22 Callhome dialogs, totalling about 10 hours and 44,000 total words.
4. The Ten Patterns

This section describes the patterns observed, starting with the most familiar.

4.1. The Basic Turn Hand-off Construction

Figure 1 shows the loadings of Dimension 2 from the PCA. (The numeric values for this and the other dimensions are available in the online appendix, at http://www.cs.utep.edu/nigel/jtt/.) From the intensity loadings we see that first one participant (A) is speaking, and then goes silent while the B speaker takes the turn. (The A and B roles are of course not fixed, but are used only for convenience of reference.) The other lines in the figure indicates typical prosodic feature values in the vicinity of such turn hand-offs. These include

- A –1000 ms high pitch
- A –600 ms creaky, slow, reduced, slightly louder
- A –400 ms low pitch, intensity drop, stops speaking
- B 0 ms starts speaking, high in pitch
- B 600 ms lengthened

To mention only the areas in which the features take on the most extreme values. Some of these properties correspond to well-documented correlates of “typical” turn-taking in Japanese, confirming the utility of this method. An example occurs in a discussion of restaurants

...丸亀たまにいくけど
うまいよね。丸亀製麺
... we eat at Marugame
yeah, it’s good, Marugame Noodles

This example was chosen as one centered around a time-point where the computed value for this dimension was highly positive. (While here the prosodic pattern is clearly present, the observed prosody does not exactly match the pattern, since, as always, the observed prosody also reflects the superimposed contributions of other constructions, often some of them fairly strongly.) This example was also selected as one hopefully able to evoke for the reader the prosody present. The English translation is intended to indicate the topic, content, and timing of the utterances, although of course the syntax and nuances are off, and it is unlikely to help evoke the prosody of the original. Readers are referred to the audio itself, available at http://www.cs.utep.edu/nigel/jtt/.

In terms of lexical tendencies, words more common around –500 ms for the yielding speaker (A) include よ and じゃ, and around 0 ms for the incoming speaker (B) the connective でも.

4.2. The Backchanneling Construction

Figure 2 shows the Backchanneling Construction, in which prototypically a contribution by B neatly fits in a pause in the ongoing speech of A. The well-known low pitch cue [12] is clearly visible, around –600 ms, and some additional properties:

- A –1000 ms lengthened, creaky
- A –500 ms low pitch
- B 0 ms short, lengthened, falling pitch, slightly creaky speech (the backchannel)
- A 500 ms resumes speaking, initially high in pitch, then with lengthening

This is seen, for example, when explaining a movie:

そこで、女の子はね だけど代わりに自分の体の一部を
right, but just for the girl but the downside is that she
uh-huh

where the main speaker cues a backchannel, receives it, and then goes on to deliver more information.

In terms of lexical tendencies, around –500 the A speaker’s more common words include the connectives それでで、でも、う, and か; in the backchannel position (B speaker around 0) unsurprisingly うん; and around +500 ms for the A speaker again それでで で, but also the resumptive particle なん か and とか.

4.3. The Has-Floor Construction

Dimension 1 is about which speaker has the floor. This can only loosely be considered a construction, but it does have prosodic correlates:

- A –1600 ~ +1600 ms pitch slowly falling (declination), slightly slow speaking rate

These are common properties of “turn-holding” prosody across languages, and can sound, perceptually, like a semiformal or near read-speech style, as in the middle phrase of とうしよう。日本の食べ物で美味しいもの、とかいいですかね
well... Japanese foods that you miss, or something like that

where the speaker is proposing a topic for discussion. In terms of lexical tendencies, words more common around 0 ms include が.
4.4. The Topic Enthusiasm Construction

The previous patterns involved who speaks when, but there is also a weak pattern from Dimension 4 relating to when speech will occur, without specifying who will do the speaking. In this one or both participants contribute a region of high pitch and clearly articulated speech, with a shorter region of increased intensity, and this weakly predicts continuing speech by one or the other. Without any precise temporal configuration, these properties appear broadly over a span of a few seconds:

- A and/or B -1600 ms ~ 1600 ms high pitch, articulated
- A and/or B -400 ms ~ 400 ms increased intensity
- A and/or B 1600 ms ~ 3200 ms speaking

For example, the speaker suddenly noticed a note on the wall with instructions and mentioned this excitedly:

おお、なんか、張り詰め、あっ、張り詰める書いてある

*oh, there’s a, a sticky note, oh, with something written on it*

which naturally led to reading it and discussing of what it meant. Another example occurred just before the “downside” example above. In this case both speakers are showing engagement with the topic, the scenario in a movie, and indeed the discussion then continued:

祈ると晴れる、晴れるんだけと、おそらく、女の子はね

*if she prays the weather clears, clears up right, but really? otherwise not?*

Incidentally, while this construction is agnostic as to which participant(s) will be speaking in future, it generally occurs with other constructions superimposed that do impose such expectations.

4.5. The No-Topic Construction

This pattern is the inverse of the previous one. Both speakers are speaking quietly, if at all, and in low pitch, typically with long silences. This is weakly predictive of a longer timespan of mostly silence, as seen when the speakers tire of a topic, as in:

なにラーメン？豚骨？

*the ramen is? tonkotsu?*

**A & B**

- -1600 ms ~ 1600 ms low pitch
- -200 ms ~ 200 ms a half-second of silence

4.6. The Particle-Assisted Turn-Taking Construction

Figure 3 depicts a construction in which a turn change happens over a couple of seconds, prototypically with two particles or other small utterances mediating the process. This frequently involves a progression to a new subtopic, for example, from discussing the genre of a movie to discussing how the other person felt about it:

感動たな、とっかと、。、うん、感動？泣くの？泣けるの？

Figure 3: The intensity loadings of Dimension 5.

```
A ~800 ms speaking then ends
B ~300 ms short contribution
A 300 ms short, low pitch contribution
B 800 ms takes the floor
```

4.7. The Filler Construction

The loadings of Dimension 6 on the negative side imply a pattern:

- A around 0 ms low, flat pitch, low articulatory precision, with a brief lengthened region
- B 1200 ms and on speaking

Generally this is produced by a speaker who doesn’t know what to say immediately, and this is thus a filler, prototypically followed by speech after about a second of delay, as in:

自分の体の一部失っていく

ヘーー なにそれ

part of her body liquifies

whaata that’s just weird.

4.8. The Commentable Information Construction

The loadings of Dimension 6 on the positive side, conversely, imply the following pattern:

- A ~800 ms ~ 800 ms high pitch, wide-ish pitch range, high articulatory precision, fast
- B 1200 ms and on speaking

Generally this serves as a way to elicit a reaction: the speaker produces a word or two that are clearly articulated and have increased pitch, with also a tendency to wider pitch range and faster rate, as in the word *Japanese* in:

それは日本人としてはおかしいって

*(laughs)それは言い過ぎだけど*

“are you really Japanese?” they tease

*(laughs) that’s going too far*

This pattern is very often superimposed on other constructions, including for displaying enthusiasm, backchannel cuing, and turn yielding behavior, in which case it adds or intensifies a cue for the other to speak.

4.9. The Turn-Internal Pause Construction

While many pauses occur as part of one or more of the patterns already discussed, some pauses are turn-internal. Prototypically these occur flanked by words that are carefully articulated and somewhat high in pitch:

- A -1600 ~ -400 increasing pitch height, articulatory precision
- A -400 ~ 400 pause
- A 400 ~ 1600 speaking with relatively high pitch and careful articulation

Variously they mark the need for time to think or to recall something, or serve as rhetorical pauses that give impact to what comes next. On the page they may resemble fillers, backchannel opportunities, or co-completions, but are prototypically different in prosodic form and in function. The example below is typical of this construction, except for the behavior of the B speaker, who atypically doesn’t remain silent.

```
In terms of lexical tendencies, after such pauses the word じゃない is more common.

4.10. The Rapid Turn Interleaving Construction

Figure 4: Interleaved Short Contributions: The intensity loadings of Dimension 8.

Figure 4 depicts a construction in which the speakers rapidly alternate short turns, both apparently tending to a rhythm of about 1.4 seconds between the loudest parts of each turn. In this construction, unusually, the loadings on the pitch and other features are small, so prosody does not play much of a role here. This pattern is common with co-completions, teasing, and/or joint laughter, and often seems to establish agreement and closure before moving on to a new topic or subtopic.

Makoto Shinkai’s new film? So, comparing it to Your Name… yeah yes

5. Comparisons with English and Mandarin

Previous comparisons of turn-taking in Japanese versus in other languages have found both commonalities and differences [32, 33, 34]. Using this newly developed inventory of patterns, we also find that both English and Mandarin, when analyzed in the same way [28, 35], exhibit analogs of some of the constructions found here, but, crucially, not all, as summarized in the second column of Table 1. We also find, even for patterns with analogous across languages, some salient differences, such as the shorter time from low-pitch cue to response in the Backchanneling Construction in Japanese. Some of these differences may relate to fundamental cultural differences [4].

6. Summary and Discussion

The inventory presented here largely agrees with previous observations about turn-taking (Table 1), but adds new details to several patterns. In addition it includes several patterns not previously reported. At least six of these are truly prosodic constructions, by the strictest definition. (The count is higher if one prefers to treat individual behaviors, such as turn taking and turn yielding, as separate patterns rather than as components of joint constructions.)

Whatever the exact number, these findings show that, even in a well-trodden field, there can still be things to discover, and that data-driven methods can reveal them. It also provides indirect evidence for the utility of the assumptions underlying the prosodic constructions model and of PCA-based analysis. The toolkit for doing such analyses is freely available [36].

<table>
<thead>
<tr>
<th>Function</th>
<th>Languages</th>
<th>Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic Turn Exchange</td>
<td>J E M</td>
<td>2 (6%)</td>
</tr>
<tr>
<td>Backchanneling</td>
<td>J e m</td>
<td>3 (4%)</td>
</tr>
<tr>
<td>Has Floor</td>
<td>J E M</td>
<td>1 (20%)</td>
</tr>
<tr>
<td>Particle-Assisted Turn-Taking*</td>
<td>J E</td>
<td>5 (2%)</td>
</tr>
<tr>
<td>Topic Enthusiasm</td>
<td>J E m</td>
<td>4 (3%)</td>
</tr>
<tr>
<td>No-Topic</td>
<td>J E M</td>
<td>4 (3%)</td>
</tr>
<tr>
<td>Filler</td>
<td>J E</td>
<td>6 (2%)</td>
</tr>
<tr>
<td>Commentable Information*</td>
<td>J --</td>
<td>6 (2%)</td>
</tr>
<tr>
<td>Turn-Internal Pause*</td>
<td>J e m</td>
<td>7 (2%)</td>
</tr>
<tr>
<td>Rapid Turn Interleaving*</td>
<td>J --</td>
<td>8 (2%)</td>
</tr>
</tbody>
</table>

Table 1: Summary listing of turn-taking patterns, with asterisks marking those not previously noted in the literature. The second column indicates whether analogous constructions are known for Japanese, English and Mandarin, with lower case letters for weaker correspondences. The last column names the corresponding PCA dimension and shows the variance explained by that dimension, as a rough indication of the magnitude of its contribution to the observed prosody, across all the data. Some dimensions correspond to two constructions, as noted above.

Two further methodological lessons can be drawn. Some research traditions assume that all the information relevant to turn transitions is localized in the few hundred milliseconds prior to pauses, but here we confirm earlier findings showing the relevance of prosody across wider spans [9]. This implies that empirical investigations should consider wider contexts. Many research traditions further assume that that the structure of turn-taking is the same across languages, with variation found only in cue forms and parameter values, but here we find language-specific patterns, different in kind from those seen in other languages. This implies that research that starts with universal annotation practices may fail to discover important phenomena. Together these imply that previous findings about cross-language tendencies in turn-taking [33, 37], may be less indicative of general universality than is often thought.

There are also implications for builders of dialog systems. Despite the common assumption that turn-taking can be considered an separate module following its own rules, independent of the content being expressed or the dialog activities being enacted, the findings here suggest significant connections, and thus the need for turn-taking knowledge to be deeply incorporated in dialog models. Further, while most dialog systems today are only capable of basic turn-taking, and accordingly implicitly cue the users to restrict their behavior to the simple Basic pattern, designers can consider this wider inventory of patterns as a resource, potentially useful for supporting new genres and styles of interaction.

Many questions remain, about the details of various turn-final intonation patterns, about how the prosody interacts with dialog act options and syntactic and lexical resources [38, 39, 40, 41], and about the cognitive processes involved [42]. Achieving an integrated understanding will doubtless require the use of multiple methods in combination. Another priority for future research is investigation of differences in turn-taking across genres, across subpopulations and among individuals.

7. Acknowledgements

I thank Aaron Alarcon for the code for characteristic word statistics and Yuko Isabel Ward for checking the transcripts.
8. References


