10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

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 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.

10.21437/SpeechProsody.2020-156

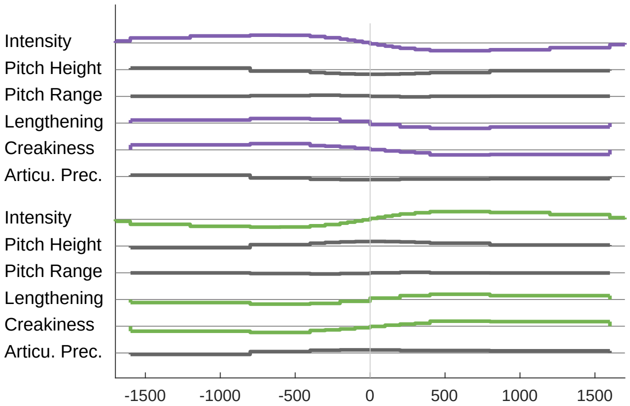

Figure 1: *Basic Turn Hand-Off: The loadings of PCA Dimension 2, with A-side speaker feature loadings above, and B-side features below. Each segment of each thick line represents the loadings of the specified feature for the specified temporal window. On the y axes the thin lines mark loading values of zero. Times are in milliseconds relative to 0 at the pattern center.*

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 refered 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 the final particles 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 Backchaneling 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:

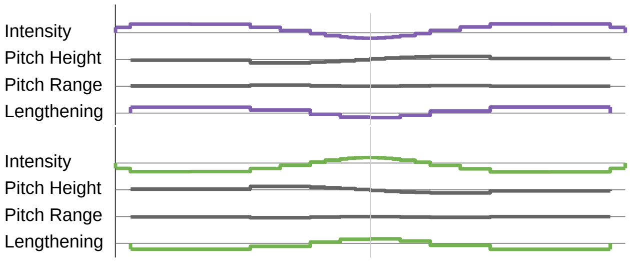

Figure 2: *Backchanneling: Some loadings of Dimension 3.*

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

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

そそそ、女の子はね　　だけど代わりに自分の体の一部を
うん

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, もう, and が; in the backchannel position (B speaker around 0) unsurprisingly うん; and around +500 ms for the A speaker again それで and 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 semi-formal or near read-speech style, as in the middle phrase of

どうしよう、　日本の食べ物で恋しいもの、とかいいすかね
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、張り(笑)、あっ、張り紙が書いてある
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 occured 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、晴れるんだけど...　　　　そそそ、女の子はね
　　　　ん,　　　祈らないと晴れない

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？　　　　　へ　　　　豚骨か
　　　　　　　え？　豚骨、豚骨
the ramen is? tonkotsu?　　　　　*ah,*　　　*tonkotsu*
　　　　　what?　*tonkotsu, tonkotsu*

A&B	–1600 ms ∼ 1600 ms	low pitch
A&B	–200 ms ∼ 200 ms	a half-second of silence

4.6. The Particle-Assisted Turn-Taking Construction

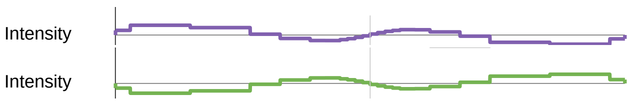

Figure 3: *The intensity loadings of Dimension 5.*

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

感動だな、どっちかと、、、　　　　うん
　　　　　　　　感動？　　泣くの？泣けるの？

it's moving, more than amusing　　　*yeah*
　　　　　　　　moving　　*makes you cry?*

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
A	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 are followed by speech after about a second of delay, as in:

自分の体の一部失っていく
　　　　　　　　へーー　　　なにそれ

part of her body liquifies
　　　　　　　　whaaat　　　*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
　　　　　　　　　　　(笑)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.

じゃ俺は,俺はない　ない　　　　泣いたかな　泣いた気がする
た？

well I, I　　　　*I don't remember.*　*I think I cried*
cried?

In terms of lexical tendencies, after such pauses the word やはり (*after all*) is more common.

4.10. The Rapid Turn Interleaving Construction

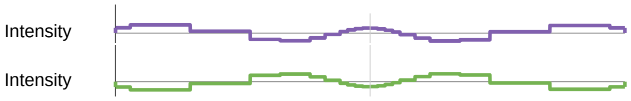

Figure 4: *Interleaved Short Contributions: The intensity loadings of Dimension 8.*

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.

新海誠監督のやつ？　　「君の名は」とはどっちが…
うんうん　うん

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 analogs 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 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 of the assumptions underlying the prosodic constructions model and of PCA-based analysis. The toolkit for doing such analyses is freely available [36].

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

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

[1] S. C. Levinson, “Turn-taking in human communication: Origins and implications for language processing,” *Trends in Cognitive Sciences*, vol. 20, pp. 6–14, 2016.

[2] H. Tanaka, *Turn-taking in Japanese conversation: A study in grammar and interaction*. John Benjamins Publishing, 2000.

[3] H. Furo, *Turn-taking in English and Japanese: Projectability in grammar, intonation and semantics*. Routledge, 2013.

[4] H. Tanaka, “Turn projection in Japanese talk-in-interaction,” *Research on Language and Social Interaction*, vol. 33, pp. 1–38, 2000.

[5] P. T. Brady, “A model for generating on-off speech patterns in two-way conversation,” *Bell System Technical Journal*, vol. 48, no. 7, pp. 2445–2472, 1969.

[6] 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.

[7] 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.

[8] T. Ohsuga, M. Nishida, Y. Horiuchi, and A. Ichikawa, “Investigation of the relationship between turn-taking and prosodic features in spontaneous dialogue,” in *Interspeech*, 2005, pp. 33–36.

[9] Y. Ishimoto, M. Enomoto, and H. Iida, “Projectability of transition-relevance places using prosodic features in Japanese spontaneous conversation,” in *Interspeech*, 2011, pp. 2061–2064.

[10] H. Koiso and Y. Den, “A phonetic investigation of turn-taking cues at multiple unit-levels in Japanese conversation,” in *ICPhS*, 2011, pp. 1122–1125.

[11] M. Enomoto, “Turn taking and overlaping in the Japanese Map Task dialog,” in *IPSJ SIG-SLP, 27-3*, 1999, pp. 17–24, (in Japanese).

[12] N. G. Ward and W. Tsukahara, “Prosodic features which cue backchannel responses in English and Japanese,” *Journal of Pragmatics*, vol. 32, pp. 1177–1207, 2000.

[13] M. Toyokura, T. Misu, and T. Kawahara, “Analysis of response timing depending on partners and dialogue acts in human-machine dialogue,” in *Special Interest Group on Spoken Language and Dialog 15th Meeting, Proc. (SIG-SLUD-A702)*. Japanese Society for Artificial Intelligence, 2007, pp. 15–20, (in Japanese).

[14] M. Watanabe, K. Hirose, Y. Den, and N. Minematsu, “Filled pauses as cues to the complexity of upcoming phrases for native and non-native listeners,” *Speech Communication*, vol. 50, pp. 81–94, 2008.

[15] S. Iwasaki, “Initiating interactive turn spaces in Japanese conversation: Local projection and collaborative action,” *Discourse Processes*, vol. 46, pp. 226–246, 2009.

[16] M. Hayashi, “An overview of the question–response system in Japanese,” *Journal of Pragmatics*, vol. 42, pp. 2685–2702, 2010.

[17] T. Kawahara, T. Iwatate, and K. Takanashi, “Prediction of turn-taking by combining prosodic and eye-gaze information in poster conversations,” in *Interspeech*, 2012.

[18] R. Masumura, T. Asami, H. Masataki, R. Ishii, and R. Higashinaka, “Online end-of-turn detection from speech based on stacked time-asynchronous sequential networks,” in *Interspeech*, vol. 2017, 2017, pp. 1661–1665.

[19] C. Liu, C. Ishi, and H. Ishiguro, “Turn-taking estimation model based on joint embedding of lexical and prosodic contents,” in *Proc. Interspeech 2017*, 2017, pp. 1686–1690.

[20] K. Hara, K. Onoue, K. Takanashi, and T. Kawahara, “Prediction of turn-taking using multitask learning with prediction of backchannels and fillers,” in *Interspeech*, 2018.

[21] D. Lala, K. Inoue, and T. Kawahara, “Evaluation of real-time deep learning turn-taking models for multiple dialogue scenarios,” in *International Conference on Multimodal Interaction*, 2018.

[22] N. Ward, D. Aguirre, G. Cervantes, and O. Fuentes, “Turn-taking predictions across languages and genres using an LSTM recurrent neural network,” in *IEEE SLT*, 2018, pp. 831–827.

[23] K. Hara, K. Inoue, K. Takanashi, and T. Kawahara, “Turn-taking prediction based on detection of transition relevance place,” in *Interspeech*, 2019, pp. 4170–4174.

[24] D. Lala, K. Inoue, and T. Kawahara, “Smooth turn-taking by a robot using an online continuous model to generate turn-taking cues,” in *International Conference on Multimodal Interaction*, 2019, pp. 226–234.

[25] R. Ogden, “Prosodic constructions in making complaints,” in *Prosody in Interaction*, D. Barth-Weingarten, E. Reber, and M. Selting, Eds. Benjamins, 2010, pp. 81–103.

[26] O. Niebuhr, “Stepped intonation contours: A new field of complexity,” in *Tackling the Complexity of Speech*, R. Skarnitzl and O. Niebuhr, Eds. Charles University Press, 2015, pp. 39–74.

[27] N. G. Ward, “Automatic discovery of simply-composable prosodic elements,” in *Speech Prosody*, 2014, pp. 915–919.

[28] ——, *Prosodic Pattterns in English Conversation*. Cambridge University Press, 2019.

[29] ——, “A corpus-based exploration of the functions of disaligned pitch peaks in American English dialog,” in *Speech Prosody*, 2018, pp. 349–353.

[30] G. Skantze, “Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks,” in *Sigdial*, 2017.

[31] V. Tsai, T. Baumann, F. Pecune, and J. Casell, “Faster responses are better responses: Introducing incrementality into sociable virtual personal assistants,” in *Proceedings of the 2018 International Workshop on Spoken Dialog System Technology*, 2018.

[32] H. Yamada, *American and Japanese Business Discourse: A comparison of interactional styles*. Ablex, 1992.

[33] T. Stivers, N. J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, G. Hoymann, F. Rossano, J. P. De Ruiter, K.-E. Yoon *et al.*, “Universals and cultural variation in turn-taking in conversation,” *Proceedings of the National Academy of Sciences*, vol. 106, no. 26, pp. 10 587–10 592, 2009.

[34] F. Roberts, P. Margutti, and S. Takano, “Judgments concerning the valence of inter-turn silence across speakers of American English, Italian, and Japanese,” *Discourse Processes*, vol. 48, pp. 331–354, 2011.

[35] N. G. Ward, Y. Li, T. Zhao, and T. Kawahara, “Interactional and pragmatics-related prosodic patterns in Mandarin dialog,” in *Speech Prosody*, 2016.

[36] N. G. Ward, “Midlevel prosodic features toolkit (2016-2019),” 2019, https://github.com/nigelgward/midlevel.

[37] T. K. Jachmann, “On universality of prosody as a turn-taking cue across languages,” 2015, Universität des Saarlandes, Masters’s Thesis.

[38] N. Ward, “The relationship between sound and meaning in Japanese back-channel grunts,” in *Proceedings of the 4th Annual Meeting of the (Japanese) Association for Natural Language Processing*, 1998, pp. 464–467.

[39] C. T. Ishi, H. Ishiguro, and N. Hagita, “Analysis of prosodic and linguistic cues of phrase finals for turn-taking and dialog acts,” in *Interspeech*, 2006, pp. 2006–2009.

[40] T. Kawahara, Z.-Q. Chang, and K. Takanashi, “Analysis on prosodic features of Japanese reactive tokens in poster conversations,” in *Speech Prosody*, 2010.

[41] Y. Ishimoto, T. Teraoka, and M. Enomoto, “End-of-utterance prediction by prosodic features and phrase-dependency structure in spontaneous Japanese speech.” in *Interspeech*, 2017, pp. 1681–1685.

[42] M. Barthel, “Speech planning in dialogue: Psycholinguistic studies of the timing of turn taking,” Ph.D. dissertation, Radboud University Nijmegen, 2020.