Analysis of prosodic and linguistic cues of phrase finals for turn-taking and dialog acts

Carlos Toshinori Ishi, Hiroshi Ishiguro and Norihiro Hagita

Intelligent Robotics and Communication Laboratories
ATR, Kyoto, Japan
carlos@atr.jp, ishiguro@ams.eng.osaka-u.ac.jp, hagita@atr.jp

ABSTRACT

This paper presents an analysis on the functions carried by phrase final tones in turn-taking and dialog acts, taking into account linguistic information about the part of speech (particles and auxiliary verbs) attributed to the morphemes at phrase finals. Natural conversational speech data are segmented in inter-pause units, and each utterance unit is arranged according to the phrase final morphemes. Turn-taking functions are annotated, and tones of each phrase final are described by acoustic-prosodic features. Analysis results show a relationship between tones and turn-taking functions in most of the morphemes, while no clear relationship is found in some classes of morphemes which are final particles.
Index Terms: turn-taking, phrase final tones, morphemes, dialog acts.

1. INTRODUCTION

A suitable prosodic processing is fundamental for getting an appropriate speech understanding or a natural speech synthesis in applications dealing with daily conversational speech.

In Japanese, the tones of phrase finals (phrase final tones) have important roles of modality (such as question, statement, unexpectedness, and urging), focus, indication of semantic boundaries, and turn-taking. The term "**phrase final tone**" is used in this paper to indicate pitch movements in the last syllable of the last morpheme of a phrase.

The functions carried by the phrase final tones depend on the presence/absence of a **final particle** (particles that occur in sentence-final position, such as "ka", "yo", "ne" and "wa") [1].

In sentences not ending with final particles, question rise tones are used to request a response or a reaction, or to ask again, while emphatic rise tones are used to express a strong claim or persistence. Flat tones don't carry any specific function, but they can express several modalities (such as statement, order, volition, and wish) according to the particle or the auxiliary verb put at the phrase final [1].

Emphatic rise tones or rise-fall tones are also often used at the boundary of phrases in the middle of a sentence, as indicative of a semantic punctuation, indicating that the utterance didn't finish yet [2]. This strategy not only increases the discourse understanding, by making the semantic boundaries clear, but also enables the speaker to insert a pause during his utterance, still keeping the discourse turn. It often appears at the breathing breaks, also leading the listener to produce a back-channel [2]. Thus, it keeps the discourse turn, and requests attention or agreement from the listener.

Final particles have the role of bridging a gap of understanding between the speaker and the listener. A detailed analysis on the functions of different tones is reported in [3] for the final particle "ne". However, the results for "ne" can not be straightly applied for other particles. The functions carried by the phrase final tones may differ according to the **part of speech** (e.g., particles and auxiliary verbs) put at the phrase final. Further, the final particles "ne" and "sa" have the function of call the listener's attention, when inserted in the boundary of phrases in the middle of a sentence [4], e.g. "Saikin-**ne**, konna hyougenga-**ne**, hayatteiruyo." ("Recently (!), this kind of expression (!) is getting popular.") This is called "interjectional usage" of final particles [4].

Therefore, the use of tone information in speech synthesis or recognition applications would require knowledge about the relationship between the tones and linguistic information, such as part of speech. In [5], syntactic and prosodic features were investigated at the points where turn-taking and backchannels occur. The main findings were that some instances of syntactic features make extremely strong contributions, but prosodic features contributes as strongly as, or even more strongly than syntax. A more recent work reports the use of prosodic features of the whole accentual phrase for turn-taking [6], but disregarding linguistic information. In the present work, we focus on the prosodic features of the phrase finals, considering morphologic and syntactic features.

Considering the importance of the phrase final tones in dialog speech communication, in our past work [7], we proposed and evaluated an automatic detection of phrase final tones, by using perceptually-related acoustic parameters representing pitch movement and duration of phrase finals.

In the present work, we make use of this automatic detection tool, and investigate the functional roles of phrase final tones (prosodic cues) on turn-taking and dialog acts, taking into account information about the part of speech attributed to the morphemes appearing in phrase finals (linguistic cues).

2. CLASSIFICATION OF PHRASE FINALS

For analysis, we used utterances extracted from the CREST/ESP speech database of natural conversations uttered by one adult female speaker with Tokyo dialect (FSM). The utterances are segmented in inter-pause units composed by one or more accentual phrases, when pauses longer than 300 ms are present at the phrase boundaries. These utterance units are simply called "**phrases**", henceforth. 3422 phrases were extracted from the speech database.

2.1 Classification of turn-taking functions

Here, we classify the turn-taking functions based on the classification proposed in [8]. According to [8], in conversation we use turn-yielding cues, back-channel cues, and turn-maintaining cues.

Turn-yielding cues are used by speakers to let the listener know that they have finished what they want to say and that someone else may speak. The display of a turn-yielding cue does not require the listener to take the floor; he may remain silent or reinforce the speaker with a back-channel cue. If the turn-taking mechanism is operating properly, the listener will take his turn in response to a turn-yielding cue emitted by the speaker, and the speaker will immediately yield his turn [9]. In the present work, we created a sub-category for turn-yielding cues, when the speaker requests a response or an agreement from the listener, as shown in Table 1 (*Yd* and *Rq*).

Back-channel cues are used by listeners to indicate that they do not wish to talk even though the speaker is displaying turn-yielding cues. So, the listener stays in his or her position when there is an opportunity to become the speaker [9].

Turn-maintaining cues, in which speaking-turn claims are suppressed, are used by speakers to keep their speaking turn. Although hand gestures may constitute the most important nonverbal behavior for this purpose, some vocal cues may be used alone or may accompany hand gestures [9]. Using fillers (e.g., "ee...", "ano…") instead of silent pauses is a useful method of turn-maintaining (*Kp* and *Fi* in Table 1).

In our data, we observed that the speaker sometimes raises the intonation in the phrase final as if the speaker was asking to oneself, connoting some uncertainty. We decided to treat this as an especial case of turn-maintaining cue, as shown in Table 1 (*Kp?*).

All phrases of the speech database are labeled according to the turn-taking labels shown in Table 1. The turn-taking labels are annotated by 3 native subjects, and later corrected by another subject.

Table 1: Description of turn-taking and dialog act functions.

Turn-maintaining	*Kp*	Keep the turn of the discourse
	Kp?	Keep the turn, but as asking to himself; uncertain
	Fi	Fillers
Neutral	*Nt*	Applicable to both keeping or yielding the turn
Turn-yielding	*Yd*	Yield the turn of the discourse for the listener
	Rq	Request a response or an agreement
	Bc	Back-channels

2.2 Classification of phrase finals based on morpheme and part of speech

Linguistic information about the part of speech attributed to the morphemes appearing at phrase finals is taken into account when verifying the influence of tones on turn-taking cues. For example, phrases ending with final particles are expected to have turn-yielding functions, while phrases ending with conjunctive particles are expected to have turn-maintaining functions.

All phrases were first arranged according to the morpheme appearing at their phrase finals. The identification of the morpheme was conducted by one subject, based on the text information available in the database.

From the analysis results relating tones and turn-taking functions, we grouped the morphemes according to the categories shown in Table 2. The numbers within the parenthesis indicate the number of occurrences for each morpheme. Descriptions of the categories and functions carried by several morphemes (or parts of speech) can be found in [4] and [10].

In 661 phrases, the morphemes could not be categorized by the ones shown in Table 2. The analysis of these phrases is left for future work.

Table 2: List of morphemes grouped according to syntactic functions, and corresponding number of utterances

Final particles/auxiliary verbs	*1a*	Final particles having multiple parts of speech	"ne" (303), "sa" (142), "yo" (122), "ka" (114), "no" (98), "wa" (10), "ga" (39), "kedo" (88).
	1b	Final particles and final auxiliary verbs having request functions	"ke" (14), "deshoo" (57), "darou" (6), "wake" (11), "jan" (37), "janai" (25), "kana" (31), "kai" (2).
	1c	Remaining final particles and final auxiliary verbs	"desu" (14), "masu" (5), "mon" (7), "da" (32), "ta" (63), "nai" (45), "kamo" (3), "mashou" (6), "kudasai" (8).
2		Non-final particles and non-final auxiliary verbs	"wa" (61), "wo" (12), "mo" (32), "ni" (42), "to" (29), "kara" (56), "dakara" (16), "shi" (21), "tara" (38), "ba" (18), "te" (170), "de" (96).
Fillers	*3a*	Possibly accented fillers	"kou" (6), "maa" (16), "jaa" (12), "demo" (13), "nanka" (63), "hora" (9).
	3b	Unaccented fillers	"ano" (23), "sono" (18), "eeto" (9).
4		Backchannels	"un" (308), "ee" (35), "aa" (37), "sou" (41), "sousou" (22), "hai" (35), "fun" (35).

2.3 Acoustic parameters for the phrase finals: *F0move* and *duration*

Here, we use a set of parameters proposed in [7], for describing the intonation of phrase finals (last syllable of the phrase), based on F0 and duration information.

For the pitch-related parameters, F0 is first estimated based on the normalized autocorrelation function of the LPC inverse-filtered residue of the pre-emphasized speech signal. Details about the F0 estimation procedure can be found in [7]. All F0 values are converted to a musical (log) scale before subsequent processing. The expression (1) shows a formula to produce F0 in semitone intervals.

$$F0[semitone] = 12 * \log2\,(F0[Hz]) \quad (1)$$

In [7], each syllable is broken in two segments of equal length, and representative F0 values are extracted for each segment. Several candidates for the representative F0 values have been tested in [7]. Here, we use the ones that best matched with perceptual scores of the F0 movements. For the first segment, an average value is estimated using F0 values within the segment (*F0avg2a*). And for the second segment, a target value is estimated as the F0 value at the end of the segment of a first order regression line of F0 values within the segment (*F0tgt2b*). A variable called *F0move* is then defined as the difference between *F0tgt2b* and *F0avg2a*, quantifying the amount and direction of F0 movement within the syllable. *F0move* is positive for rising F0 movements, and negative for falling movements. For analysis of the tone functions in the next section, phrase finals are categorized as rise pitch movements (*Rs*) when *F0move* > 1 semitone, fall pitch movements (*Fa*) when *F0move* < -2 semitones, and flat pitch movements (*Ft*) otherwise. These thresholds are based on pitch movement perception experiments. Details about the evaluation of these parameters can be found in [7]. Flat pitch movements (*Ft*) are also attributed to the phrase finals where *F0move* values could not be obtained, mainly due to aperiodicities caused by creaky and whisper phonations, which frequently occurs in low pitched utterance finals.

For phrase final duration, an automatic procedure was first realized, by using power and spectral change constraints. As the objective of the present work is the analysis of tones and turn-taking cues, the errors in the automatic segmentation were manually corrected. The newly segmented boundary intervals are used as segmental duration of the phrase finals. As for pitch movement categories, duration categories are also defined as short (*S*) when *duration* < 200 ms, and long (*L*) otherwise.

A tone is then described by the combination of the pitch movement and the duration categories.

3. RESULTS AND DISCUSSIONS

Table 3 summarizes the results showing the distributions of phrase final tones for each turn-taking or dialog act function, arranged by each group of morphemes. The tones shown in Table 3 are the representative ones (i.e., the ones that showed higher percentage of occurrence) for each turn-taking category and each morpheme group. The numbers within the parenthesis indicate the percentages of occurrence of the representative tones.

For the non-final particles/auxiliary verbs of group *2*, and the particles "ga" and "kedo" of group *1*, long fall tones (*LFa*) and long flat tones (*LFt*) appeared mostly in turn-maintaining cues (*Kp* and *Fi*), while short flat tones (*SFt*) appeared mostly in turn-yielding cues (*Yd*) and back-channel cues (*Bc*). Short and long rise tones (*SRs* and *LRs*) appeared in *Rq*. These results are in accordance with the general functions of phrase final tones reported in past researches, as described in the introduction.

For the final particles of the groups *1b* and *1c*, and for the particle "wa", all phrases have turn-yielding cues. The phrases labeled as *Yd* were mostly short flat tones (*SFt*), while most rising tones appeared in the phrases labeled as *Rq*. However, in group *1b*, flat tones (*Ft*) also appeared in approximately half of the *Rq* phrases. A reason for that is thought to be that the final particle itself has a function of requesting a response from the listener.

The phrases ending with final particles are theoretically expected to appear only as turn-yielding cues. However, multiple parts of speech can be attributed to some of the morphemes. Analysis on the current data showed the particles "no", "sa", "ne", and "ka" (group *1a*) appeared also as turn-maintaining cues, i.e., as a part of speech other than final particles. Each of these particles was then analyzed separately.

Table 3. Distributions of phrase final tones for each turn-taking function, arranged by the morphemes of the phrase finals. The numbers within the parenthesis indicate the percentage of occurrence for each turn-taking function and each morpheme group.

			Turn-taking function: Phrase final tones (% of occurrence); Total number of occurrences							
			Rq		*Yd, Bc*		*Nt*		*Kp, Fi*	
Group *2* + ga + kedo			*SRs* (71%) *LRs* (14%)	14	*SFt* (65%) *LFt* (15%)	121	*SFt* (61%)	72	*LFa* (51%) *LFt* (27%)	469
Group *1*		Group *1c*	*Rs* (72%)	24	*SFt* (65%) *LFt* (20%)	133				
		Group *1b*	*Rs* (52%) *Ft* (46%)	67	*SFt* (48%) *LFt* (22%)	95				
		wa	-	0	*SFt* (80%) *SRs* (10%)	10				
		no	*SRs* (48%) *SFt* (32%)	40	*SFt* (75%)	20	*SFt SRs*	8	*LFa* (50%) *LFt* (30%)	24
		sa	-	0	*SFt* (60%) *LFt* (40%)	15	*SFt LFa*	7	*LFa* (56%) *LFt* (19%)	119
	ne	ne	*SFt* (64%) *SRs* (35%)	14	*Ft* (64%) *Rs* (22%)	99	*Rs*	8	*Ft* (43%) *Rs* (31%)	16
		nine,kedone,wane	-	0	*Rs* (53%) *Ft* (47%)	17	*Ft*	3	*Rs* (56%) *Ft* (43%)	16
		tene, none	-	0	*Rs* (50%) *Ft* (50%)	20				
		dane, yone	*SFt* (50%) *Rs* (31%)	16	*Ft* (62%) *Rs* (22%)	58				
	yo	yo	*SFt*	1	*Ft* (71%) *SRs* (14%)	119				
	ka	ka	*Ft* (68%)	21	-	0	-	0	*SFt*	3
		toka, toyuuka	-	0	*SFt*	6	*SRs*	3	*LFa* (57%) *LFt* (19%)	37
		desuka, masuka	*Ft* (76%) *Rs* (16%)	30						
Group *3*		Group *3a*							*Ft* (58%) *Fa* (28%)	128
		Group *3b*							*LFt* (56%) *SFt* (20%)	50
Group *4*			*Rs* (58%)	40	*LFa* (54%) *LFt* (27%)	450				

"no" appeared in phrases with turn-maintaining functions as case particles, e.g. "koosokudooro**no**…" ("… of the expressway."). The tones appearing in *Kp* and *Yd* show the same trends for group *2*. However, some confusion is found between *Rq* and *Yd*, which share short flat tones (*SFt*).

In the case of "ka", besides the question marker functions as final particles, some phrases appeared as adverbial particles or conjunctive particles. Table 3 shows that short flat tones appear in all turn-taking categories. The knowledge about the previous morpheme could help the decision of turn-taking, since it was always *Rq* when "ka" is preceded by final auxiliary verbs (e.g. "masuka", "desuka"), while it was *Kp* or *Yd*, in "toka" and "toyuuka".

"ne" and "sa" frequently appeared as interjectional particles, as described in the introduction. Almost all phrases ending with "sa" were used as interjectional particles in the present speech data, following the same rules of group *2*.

In the case of "ne", no clear relationship could be found between tones and turn-taking functions. However, it was observed that the phrases ending with "ne" always have turn-yielding functions, when the particle preceding "ne" is a final particle (e.g., "yone", "dane", "dayone"). Further, if "ne" is preceded by non-final particles (e.g. "nine", "kedone", "wane"), *Rq* never appears.

The final particle "yo", which has functions of directing attention, warning, or notice, appeared almost always in *Yd.* Only 1 sample appeared in *Rq*. However, the tone information could not be used for discriminating them.

Regarding the fillers (group *3*), they can be sub-categorized in words that can be accented (group *3a*), and words that are unaccented (group *3b*). The tones of group *3a* can be flat or fall, while in group *3b*, long flat was predominant. Although it is less common, some samples also appeared with rising tones.

Global results showed that most back-channels (group *4*) indicating agreeable reactions were *LFa*. Rise tones appeared for back-channels indicating a request for repetition. A deeper study about tone functions in backchannels and fillers can be found in [11].

Only a few phrases (about 5 % of the total of phrases) were labeled as *Nt*, in the auxiliary particle/verb groups. The phrase final tones of these phrases were somewhere between *Yd* and *Kp*, but more similar to the *Yd* tones.

Finally, *Kp?* (turn-maintaining with rising tones) is omitted from Table 3, but it also appeared in only a few phrases (about 3 % of the total of phrases). Although the use of rising tones is much less frequent for turn-maintaining (*Kp?*) than for turn-yielding (*Rq*), one should be able to distinguish them. Unfortunately, no clear discrimination could be observed in the tone property of the phrase finals. Other prosodic features accounting the whole phrase could be useful for solving this problem. Further analysis on this problem is left for future work.

The above results indicate that the recognition of the morphemes would be useful prior to the use of tone information for turn-taking or dialog act decisions.

4. CONCLUSIONS

Focusing on discourse turn-taking and dialog act functions, phrase finals were classified according to linguistic information (morphemes), and the functional roles of the phrase final tones were investigated. Analysis results showed relationship between tones and turn-taking functions in utterances ending with non-final particles/auxiliary verbs, and in part of phrases ending with final particles.

For turn-yielding, short flat tones are commonly used for yield functions, while short rise tones are used for request functions. Some final particles don't obey to this general rule, but the request functions could be predicted from the morpheme type itself. For turn-maintaining, the following strategies were found: long fall tones, long flat tones, and the use of fillers or conjunctive particles. Short rise tones were also used, but they were less frequent.

Regarding the final particles "ne", "yo" and "ka", no clear relationship was found between tones and turn-taking functions. The tones of such particles are thought to be more related with other paralinguistic information like manner. This topic concerning the functionality of tones in these final particles is left for future works.

The knowledge about the morphemes was shown to be useful prior to the use of tone information for turn-taking and dialog act decisions.

The next step is to try to use speech recognition results for getting morpheme information, and evaluate automatic predictions of turn-taking and dialog acts.

5. ACKNOWLEDGEMENTS

This work was partly supported by the Ministry of Internal Affairs and Communications. We thank Akiko Nakagawa for discussions about Linguistics.

6. REFERENCES

[1] Kori, S., "Nihongo no intoneeshon – kata to kinou," in *Accent, intonation, rhythm and pause*, Sanseido, 190-196, 1997. (in Japanese)

[2] Inoue, T. "Intoneeshon no shakaisei," in *Accent, intonation, rhythm and pause*, Sanseido, 147-159, 1997. (in Japanese)

[3] Sugito, M., "Joshi to jodoushi," in *Bunpou to onsei III*, Kuroshio, 3-54, 2001. (in Japanese)

[4] Masuoka, T., Takubo, Y. "Kiso nihongo bunpou – kaiteiban," ("Basic Japanese Grammar,") Kuroshio, 49-54, 1992. (in Japanese)

[5] Koiso, H., Horiuchi, Y., Syun, T., Ichikawa, A. "An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs," *Language and Speech*, 41 [3-4], 295-321, 1998.

[6] Ohsuga, T., Nishida, M., Horiuchi, Y., Ichikawa, A. "Investigation of the relationship between turn-taking and prosodic features in spontaneous dialogue," Proc. *Eurospeech 2005*, 33-36, 2005.

[7] Ishi, C.T., Mokhtari, P., Campbell, N., "Perceptually-related acoustic-prosodic features of phrase finals in spontaneous speech," Proc. *Eurospeech 2003*, 405-408, 2003.

[8] Duncan, S. "Some signals and rules for taking speaking turns in conversations," in Journal of Personality and Social Psychology 23(2), 286-288, 1972.

[9] Gibbon, D., "Turn-taking cues,"http://www.spectrum.uni-bielefeld.de/Classes/Winter97/PhonMM/UlrichGruen/cues.htm

[10] Maynard, S.K., *An introduction to Japanese grammar and communication strategies*, The Japan Times, 1990.

[11] Ishi, C.T., Ishiguro, H., Hagita, N. "Using Prosodic and Voice Quality Features for Paralinguistic Information Extraction," Proc. *Speech Prosody 2006*, 2006.