PROSODIC CHARACTERISTICS OF JAPANESE NEWSCASTER SPEECH FOR DIFFERENT SPEAKING SITUATIONS

Shizuka Nakamura[1], Carlos Toshinori Ishi[2] & Tatsuya Kawahara[1]

[1]Graduate School of Informatics, Kyoto University, Japan
[2]Hiroshi Ishiguro Laboratories, ATR, Japan
shizuka@sap.ist.i.kyoto-u.ac.jp, carlos@atr.jp, kawahara@i.kyoto-u.ac.jp

ABSTRACT

Aiming at realizing adaptation to various speaking situations by non-human newscasters in the future, this study was conducted to clarify prosodic characteristics in representative speaking situations. Among various speaking situations, we focused on (i) neutral manuscript readings, (ii) program advertisements, and (iii) narrations. We analyzed Japanese speech utterances by nine newscasters belonging to a TV station. Analysis showed the following results on prosodic characteristics for different speaking situations; differences from neutral manuscript reading were found in pause duration (within-sentence: -123 msec, between-sentence: -353 msec), F0 (+4.4 semitones), and intensity (+5.9 dB) for program advertisements, and in pause duration (within-sentence: +206 msec, between-sentence: +486 msec) and F0 (-2.0 semitones) for narrations.

Keywords: Prosodic characteristics, newscaster, speaking situation, Japanese.

1. INTRODUCTION

In recent years, non-human newscasters, such as a virtual newscaster "Ananova" in England and an AI anchor "Xinhua" in China, using speech synthesis technology have been appeared. In Japan, an AI newsreader "Yomiko" is given as an example of animation characters. An android "ERICA" has also started to be active as a newscaster.

Human TV newscasters in Japan acquire skills at announcement school, TV station, and so forth. Training of vocalization and pronunciation required for newscasters is conducted at this type of school. In addition, they also acquire skills specific to various speaking situations including news manuscript readings, program advertisements, narrations, MCs, and on-the-spot reports of sports.

At the moment, the speech by non-human newscasters are monotonous. That is, they can not deal with various speaking situations like human newscasters. However, it is strongly desired by the broadcasting industry including TV stations that non-human newscasters can speak naturally and expressively like human newscasters.

Several studies on speaking styles have been carried out so far [1-7]. For example, some studies have reported on the differences of prosodic characteristics between spontaneous and read speech [3-4]. Phonetic variations among several speaking styles including news readings were analyzed in [5]. In the speech synthesis area, there are studies on prosody of various speaking styles using newscaster speech. For example, "interview", "news", and "live sports" were handled as different speaking styles, by using F0 and phone duration as prosodic features [6]. In [7], F0 features are controlled for synthesizing different styles in "neutral", "good" and "bad" news. However, in such previous studies, the speaking style categorization of newscaster speech and the prosodic characteristics required for those categorizations are not systematically clarified. Further, we consider that the changes in prosodic characteristics for different speaking styles differ depending on the language.

Therefore, aiming at realizing adaptation to various speaking situations by non-human newscasters, this study focuses on representative speaking situations and intends to clarify prosodic characteristics in those speaking situations. The duration, F0 and intensity are treated as prosodic characteristics. The global characteristics are measured for the entire utterances except pauses, and the local characteristics are calculated in each Inter-Pausal Unit (hereafter, IPU) [8]. In addition, pause duration is also treated.

2. MATERIALS

To reveal the characteristics for different speaking situations quantitatively, the speech uttered by newscasters was used as materials.

2.1. Speakers

The speakers are professional newscasters (20-50s, female, nine people) belonging to "the Nippon Television Network Corporation" which is one of the major TV stations in Japan. Every speaker read the same text contents.

2.2. Speaking situations

Among various speaking situations, we focused on (i) neutral news manuscript readings (hereafter, NEU), (ii) program advertisements (hereafter, ADV), and (iii) narrations (hereafter, NAR). NEU is the default speaking situation as a newscaster. These are representative speaking situations which occur with high frequency, so that the adaptation by non-human newscasters to these speaking situations is strongly expected by TV stations. The manuscripts of NEU, ADV and NAR were indicated to be read in normal, uplifting, and subdued styles, respectively.

2.3. Reading manuscripts and corresponding speech properties

Reading manuscripts, which were judged by a TV station as being representative for each speaking situation, were used for analysis. The NEU manuscript was about a daily recommendation on health management, the ADV manuscript was about a special broadcasting of the 50th anniversary variety show, and the NAR manuscript was about a visit of the US President to Hiroshima, Japan. These reading manuscripts have been used in actual TV broadcasting. Table 1 shows the manuscript composition and the speech properties for each speaking situation.

3. EXTRACTION OF ACOUSTIC FEATURES

The duration, F0, and intensity in phoneme units were extracted as acoustic features for analysis. In this section, the overall statistics of the acoustic features are presented for the NEU situation, which will be used as reference to evaluate the differences to the other speaking situations in Section 4.

The duration of each phoneme and pause was automatically measured by "the speech segmentation toolkit using Julius." In this toolkit, speech is segmented with 10 msec resolution. Table 2 shows the distributions of the phoneme durations for each speaker in the NEU situation.

For the pitch-related parameters, the F0 values were estimated each 10 msec by a conventional autocorrelation-based method [9]. All the estimated F0 values were then converted to a musical (log) scale before any subsequent processing. The following equation was used to produce F0 values in semitone intervals.

$$F0[semitone] = 12*log_2(F0[Hz])$$

Table 3 shows the distributions of F0 for each speaker in the NEU situation.

Table 1: The manuscript composition and the speech properties for each speaking situation.

Manuscript composition & Speech property / Speaking situation	Total number		Overall duration including pauses [sec]			
	Mora	IPU	AVG	SD	MIN	MAX
NEU: neutral manuscript reading	218	11	20.8	3.7	13.9	14.8
ADV: program advertisement	146	10	13.6	1.0	11.7	23.3
NAR: narration	143	8	15.0	1.1	13.5	16.9

Table 2: Distributions of the phoneme durations for each speaker in the NEU situation.

[msec]

Speaker	A	B	C	D	E	F	G	H	I	AVG
AVG	80	80	73	75	77	79	73	80	73	77
SD	64	53	36	38	37	41	47	44	37	
MIN	30	30	30	30	30	30	30	30	30	30
MAX	730	530	170	280	270	260	450	300	220	357

Table 3: Distributions of F0 for each speaker in the NEU situation.

[semitone]
([Hz] in parentheses)

Speak-er	A	B	C	D	E	F	G	H	I	AVG
AVG	95 (246)	94 (232)	95 (247)	95 (243)	94 (221)	96 (256)	93 (217)	96 (254)	95 (246)	95 (240)
SD	3.2	3.8	4.5	3.9	4.2	3.2	3.4	3.6	4.5	
MIN	88 (156)	87 (149)	84 (127)	86 (145)	80 (101)	88 (163)	86 (145)	88 (161)	83 (122)	86 (140)
MAX	101 (341)	100 (322)	104 (397)	102 (353)	101 (345)	102 (353)	100 (326)	102 (364)	102 (370)	102 (352)

Table 4: Distributions of intensity for each speaker in the NEU situation.

[dB]

Speaker	A	B	C	D	E	F	G	H	I	AVG
AVG	42.4	40.9	40.2	40.5	38.2	41.6	39.4	48.9	41.2	41.5
SD	9.5	8.8	9.2	8.7	7.5	9.5	8.2	9.0	10.3	
MIN	7.2	9.1	10.8	9.0	5.6	10.8	15.1	14.1	6.4	9.8
MAX	60.7	58.9	61.0	60.7	52.8	59.8	54.0	66.6	61.8	59.6

The intensity of the speech signal was calculated every 10 msec in dB. Table 4 shows the distributions of intensity for each speaker in the NEU situation.

4. ANALYSIS OF PROSODIC CHARACTERISTICS

Global and local characteristics on duration, F0, and intensity were analysed for the ADV and NAR situations, in comparison to the NEU situation.

4.1. Pause duration

In a previous study, it is reported that the difference in the average pause durations located within-sentence and between-sentence is significantly large regardless of the speech rate [10]. Therefore, we categorized pauses into these two groups in this study.

Table 5 shows the distributions of pause durations for different speaking situations and pause locations. T-tests confirmed that the pauses located between-sentence were significantly longer than those located within-sentence, for all speaking situations.

For the same pause location, the pause duration of ADV was found to be significantly shorter than NEU, and that of NAR was significantly longer than NEU. This result is consistent with the general impression that ADV has a faster speech rate than NEU, and NAR is slower than NEU [11-12].

4.2. Phoneme duration

The average phoneme durations in a speech segment reflect the speech rate of that segment. We examined if there is difference in phoneme duration between different speaking situations. Table 6 shows the distributions of the ratios between the average phoneme durations of different speaking situations relative to the NEU situation, for each speaker.

No significant differences were found by t-tests. This result suggests that the general impressions that ADV is faster than NEU and NAR is slower than NEU may have been affected by the pause durations rather than the phoneme durations (i.e., the speech rate).

4.3. F0

The global characteristics of F0 across different speaking situations were firstly evaluated. Table 7 shows the average F0 differences (in semitones) between the different speaking situations and the reference NEU situation, for each speaker. F0s in ADV were higher than in NEU for all speakers; the differences were on average +4.4 semitones. Significant differences were observed in all speakers, by t-tests. Conversely, F0s in NAR were lower than in NEU for all speakers; the differences were on average -2.0 semitones. Significant differences were observed in all speakers except one, by t-tests. These differences seem to be perceptually salient.

Table 5: Distributions of pause durations for different speaking situations and pause locations.

[msec]

Speaking situation	ADV		NEU		NAR	
Location	w-sent	b-sent	w-sent	b-sent	w-sent	b-sent
Average	109	669	232	1,022	438	1,508
	|_____|***		|_____|***		|_____|***	
	|_____|***					
		|_____|***				
			|_____|***			
				|_____|***		

Significant difference
***: $p<0.001$, **: $p<0.01$, *: $p<0.05$, ns: no significance.

Symbol w-sent: within-sentence, b-sent: between-sentence.

Table 6: Distributions of the ratios between the average phoneme durations of different speaking situations relative to the NEU situation, for each speaker.

[%]

Speaker	A	B	C	D	E	F	G	H	I	AVG
ADV	101	99	91	99	98	104	108	105	92	100
	ns	ns	ns	ns	ns	ns	ns	ns	ns	
NAR	103	98	95	107	97	95	104	95	97	99
	ns	ns	ns	ns	ns	ns	ns	ns	ns	

Significant difference
***: $p<0.001$, **: $p<0.01$, *: $p<0.05$, ns: no significance.

Table 7: Distributions of the average F0 differences between the different speaking situations and the reference NEU situation, for each speaker.

[semitone]

Speaker	A	B	C	D	E	F	G	H	I	AVG
ADV	3.3	4.7	4.4	3.6	4.1	3.2	5.9	5.4	5.0	4.4
	***	***	***	***	***	***	***	***	***	
NAR	-2.7	-0.7	-3.3	-1.5	-1.4	-1.8	-3.2	-2.2	-1.3	-2.0
	***	ns	***	***	***	***	***	***	**	

Significant difference
***: $p<0.001$, **: $p<0.01$, *: $p<0.05$, ns: no significance.

The results in Table 8 indicate that F0s at the sentence end (hereafter, s-end) were significantly higher than those at the non-sentence end (hereafter, non-s-end), for ADV. This reason is considered that ADV is a bright program promotion aimed at gathering attentions of the audience, and the sentences with high F0s at the end of a sentence are considered to be heavily used. Conversely, F0s at s-end was significantly lower than at non-s-end, for NEU and NAR situations. This tendency is consistent with the characteristics of typical Japanese declarative sentences.

4.4. Intensity

Table 9 shows the average intensity differences (in dB) between the different speaking situations and the reference NEU situation, for each speaker. The intensities in ADV were bigger than in NEU for all speakers; the difference was on average 5.9 dB. Significant differences were observed in all speakers between ADV and NEU, by t-tests. No significant differences were found between NAR and NEU, in almost all speakers.

To examine the local characteristics, the intensities in IPU level were computed. Table 10 shows the distributions of normalized intensity values computed at the IPU level, for different speaking situations and sentence locations, as in Table 8 for F0 values. On one hand, there was no significant difference in ADV. On the other hand, intensity values were significantly smaller at s-end in comparison to non-s-end, in NEU and NAR situations.

Speaker B was the only newscaster showing a significant difference between the intensities in NAR and NEU. Conversely, speaker B was also the only one showing no significant differences between the F0s in NAR and NEU, as shown in Table 7. In other words, it is possible to interpret that only this speaker tends to express the difference of NAR from NEU by controlling intensity instead of F0. This kind of difference might be related to their acquired skills. This speaker has one-year experience as a newscaster, which is the least among the nine.

5. CONCLUSIONS

Prosodic characteristics of Japanese newscaster speech for different speaking situations were analyzed.

Analysis showed the following results on the global prosodic characteristics for different speaking situations; differences from neutral manuscript reading were found in pause duration (within-sentence: -123 msec, between-sentence: -353 msec), F0 (+4.4 semitones), and intensity (+5.9 dB) for program advertisements, and in pause duration (within-sentence: +206 msec, between-sentence: +486 msec) and F0 (-2.0 semitones) for narrations.

Furthermore, analysis on the local prosodic characteristics showed the following results. Differences depending on the IPU locations were observed in the following acoustic features; F0 and intensity for neutral manuscript readings, phoneme duration and F0 for program advertisements, and phoneme duration, F0, and intensity for narrations. In addition, differences in pause duration depending on the pause locations were observed in all speaking situations.

Table 8: Distributions of F0 computed in IPU level, for different speaking situations and sentence locations. F0s are normalized (subtracted) by the average F0 for each speaker.

[semitone]

Speaking situation	ADV		NEU		NAR	
Location	non-s-end	s-end	non-s-end	s-end	non-s-end	s-end
Average	3.79	5.29	0.84	-0.99	-1.47	-2.72
	|_____|**		|_____|***		|_____|*	

Significant difference
***: $p<0.001$, **: $p<0.01$, *: $p<0.05$, ns: no significance.

Symbol non-s-end: non-sentence end, s-end: sentence end.

Table 9: Distributions of the average intensity differences between the different speaking situations and the reference NEU situation, for each speaker.

[dB]

Speaker	A	B	C	D	E	F	G	H	I	AVG
ADV	5.0 ***	8.0 ***	7.2 ***	3.2 **	8.5 ***	5.1 ***	6.0 ***	4.6 ***	5.0 ***	5.9
NAR	-0.4 ns	1.9 *	0.2 ns	-0.2 ns	-1.1 ns	1.9 ns	-1.6 ns	-0.6 ns	1.0 ns	0.1

Significant difference
***: $p<0.001$, **: $p<0.01$, *: $p<0.05$, ns: no significance.

Table 10: Distributions of intensity computed in IPU level, for different speaking situations and sentence locations. Intensities are normalized (subtracted) by the average intensity for each speaker.

[dB]

Speaking situation	ADV		NEU		NAR	
Location	non-s-end	s-end	non-s-end	s-end	non-s-end	s-end
Average	6.43	6.18	1.56	-0.95	1.20	-3.76
	|_____|ns		|_____|***		|_____|***	

Significant difference
***: $p<0.001$, **: $p<0.01$, *: $p<0.05$, ns: no significance.

Symbol non-s-end: non-sentence end, s-end: sentence end.

Based on these results, we are planning to design a method of adapting a speech by non-human newscasters to the target speaking situation. We are also planning to clarify the typical characteristics of the other speaking situations.

5. ACKNOWLEDGEMENTS

This study was performed in collaboration with the Nippon Television Network Corporation and supported by JST ERATO Ishiguro Symbiotic Human-Robot Interaction program (#JPMJER1401).

6. REFERENCES

[1] Eskénazi, M. 1993. Trends in speaking styles research. *Proc. EUROSPEECH* Berlin, 501-509.

[2] Granström, B. The use of speech synthesis in exploring different speaking styles. 1992. *J. Speech Communication*, 11, 4-5, 347-355.

[3] Veiga, A., Celorico, D., Proença, J., Candeias, S., Perdigão, F. 2012. Prosodic and phonetic features for speaking styles classification and detection. *Proc. IberSPEECH* Madrid, 260-268.

[4] Barbosa, P. A. 2015. Temporal parameters discriminate better between read from narrated speech in Brazilian Portuguese. *Proc. ICPhS* Glasgow, 1053: 1–5.

[5] Brognaux, S., Picart, B., Drugman, T. 2014. Speech synthesis in various communicative situations: Impact of pronunciation variations. *Proc. Interspeech* Singapore, 1524-1528.

[6] Lorenzo-Trueba, J., Barra-Chicote, R., Yamagishi, J., Watts, O., Montero, J. M. 2013. Towards speaking style transplantation in speech synthesis. *Proc. ISCA Workshop on Speech Synthesis* Barcelona, 159-163.

[7] Sakai, S., Ni, J., Maia, R., Tokuda, K., Tsuzaki, M., Toda, T., Kawai, H., Nakamura, S. 2007. Communicative speech synthesis with XIMERA: a first step. *Proc. ISCA Workshop on Speech Synthesis* Bonn, 28-33.

[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 dialogues. *J. Language and Speech*, 41, 295-321.

[9] Ishi, C. T., Ishiguro, H., Hagita, N. 2008. Automatic extraction of paralinguistic information using prosodic features related to F0, duration and voice quality. *J. Speech Communication*, 50, 6, 531–543.

[10] Fujisaki, H., Ohno, S., Yamada, S. Analysis of occurrence of pauses and their durations in Japanese text reading. 1998. *Proc. ICSLP* Sydney, 1387-1390.

[11] Lass, N. J. 1970. The significance of intra- and inter-sentence pause times in perceptual judgements of oral reading rate. *J. Speech and Hearing Research*, 13, 777-784.

[12] Miron, M. S., Brown, E. 1971. The comprehension of rate-incremented aural coding. *J. Psycholinguistic Research*, 1, 65-71.