Analysis on Prosodic Features of Japanese Reactive Tokens in Poster Conversations

Tatsuya Kawahara, Zhi-Qiang Chang, Katsuya Takanashi

School of Informatics, Kyoto University,
Sakyo-ku, Kyoto 606-8501, Japan

Abstract

For effective indexing of presentation speech such as lectures and seminars, we explore a novel approach based on detection of the audience's interest level. In this work, we deal with poster presentations and focus on the backchannel responses or reactive tokens, which are frequently observed in poster conversations and presumably used for expressing the audience's interest level. First, we note that the most common reactive token "*hai* (yes)" is mainly used for acknowledging the speech segments, and that there are specific kinds of reactive tokens which can be used for expressing non-verbal information. Then, we made a prosodic analysis and identified effective combinations of the syllabic and prosodic patterns which express interest and surprise.

Index Terms: prosody, backchannel, reactive token, audio indexing

1. Introduction

As digital archiving of lectures and meetings has become pervasive, automatic indexing and annotation is one of the important technical issues so that we can efficiently access these kinds of archives. A number of projects have been conducted to address automatic summarization and retrieval of speech archives.

We compiled the Corpus of Spontaneous Japanese (CSJ) [1], which contains a thousand academic presentations at technical conferences. Using this corpus, we investigated automatic indexing of key sentences based on discourse markers or cue phrases combined with keywords statistics [2]. The underlying idea of summarization including other methods [3] relies on the features, such as lexical and prosodic features, of the presenter's speech, which are presumably related to the core or emphasized portion of the speech. This approach is typically called "content-based" indexing, because it requires processing, such as automatic speech recognition (ASR) and lexical analysis of the audio content to be indexed. Studies on ASR and summarization of meeting archives have been intensively conducted by AMI/AMIDA [4] and CHIL projects. Moreover in ICSI, high-level annotations, such as dialogue act tagging [5] and action item identification [6], are also being investigated.

We have started a new project on multi-modal recording and analysis of poster presentations [7]. Poster sessions have become a norm in many technical conferences, exhibitions, and open laboratories, since they provide more "interactive" characteristics in presentations. Typically, a presenter explains his work to a small audience using a poster, and the audience gives feedback in real time by nodding and/or acoustic backchannels, and occasionally makes questions and comments.

We are investigating automatic indexing of poster conversations based on the interactive characteristics. As opposed to the conventional content-based approach which focuses on the presenter's speech, we focus on the audience's reaction. Specifically, we focus on the audience's reactive tokens, rather than investigating overall prosodic patterns as adopted in former studies on "hot-spot" detection [8][9].

By reactive tokens (*Aizuchi* in Japanese), we mean the listener's verbal short response, which expresses his state of the mind during the conversation. Its prototypical lexical entries include "*hai*" in Japanese and "yeah" or "okay" in English. Note that many of them are non-lexical and used only for reactive tokens, such as "*hu:n*" in Japanese and "uh-huh" in English. It is well-known that the backchannel response using this kind of reactive tokens suggests that the listener is understanding what is being said, and also that the current speaker can continue to utter by keeping the dialogue turn. Moreover, we hypothesize that the audience signals their interest level with the syllabic and prosodic pattern of the reactive tokens. We expect that detection of the audience's interest level is useful for indexing the speech archives, because people would be interested in listening to the points other people were interested in.

In this paper, we first describe the corpus of poster sessions, and then syllabic and prosodic features of reactive tokens in Section 2. In Section 3, we present an analysis of reactive tokens in relation with the conversation mode. In Section 4, we extract characteristic reactive tokens and investigate their relationship with the interest level based on the prosodic analysis.

2. Setup for Analysis

2.1. Corpus of Poster Sessions

We have recorded a number of poster sessions specifically designed for multi-modal data collection [7]. In this study, we use four poster sessions, in which the presenters and audiences are different from each other. In each session, one presenter had prepared a poster on his own academic research. The poster had one main theme and was divided into four sub-topics, which were arrayed in quarters on its surface. In each session, there was an audience of two persons, standing in front of the poster and listening to the presentation. They had not heard the presentation before. The duration of each session was around 20 minutes.

All speech data were segmented into IPUs (Inter-Pausal Unit) with time and speaker labels, and transcribed according to the guideline of the CSJ. Annotation of clause boundaries were also manually done.

We also classified segments of the sessions into explanation (EX) mode and question-answer (QA) mode. In the EX mode, the presenter keeps an initiative and mainly gives an explanation on his work, accompanied by some feedback from the audience, such as backchannels and short comments. In the QA mode, one person of the audience takes an initiative and raises questions, which are replied by the presenter. The annotation was done manually by considering who takes an initiative in the conversation segment. Although it was done by a single annotator, there was not much difficulty in the judgement because the data is not a free conversation.

The statistics of utterance duration of the conversation modes for each poster session is given in Table 1. For the QA mode, the breakdown for two persons of the audience is also given.

Table 1: Statistics of utterance duration (sec.) by conversation modes

	EX	QA	total
session 1	461	697 (189+508)	1158
session 2	457	820 (649+171)	1277
session 3	470	975 (269+706)	1445
session 4	609	910 (647+263)	1519

2.2. Reactive Tokens used in Backchannels

We define the backchannel responses with IPUs consisting of only reactive tokens (*Aizuchi*), including non-lexical entries, made by the audience. Explicit affirmative answers are excluded although there are few of them since the presenters rarely asked questions to their audience. Fillers are also separately annotated and excluded in the analysis of this study.

The syllabic (phonological) representation of reactive tokens addressed in this work is listed in Table 2. Here,

Table 2: List of reactive tokens (*Aizuchi*)

aa (ああ)	*a: (あー)*
uN (うん)	*u:N (うーん)*
ee (ええ)	
haa (はあ)	*ha: (はー)*
hai (はい)	
huN (ふん)	*hu:N (ふーん)*
	he: (へー)

":" denotes a prolonged vowel and is often confused with double vowels, for example, "a:" vs. "aa". The double-vowel entry should have two morae with a distinct accent in the first vowel and can be repeated, for example, "*ee, ee*", while the prolonged-vowel entry can last for an arbitrary duration but never be repeated. This set is chosen from the observation of the corpus we collected, but it covers most of the typical patterns in general Japanese. Please be advised that "*e:*" is mainly used as a filler in Japanese, not as a reactive token.

2.3. Prosodic Features

It is reported that prosodic features are useful in identifying backchannels in many former works [6][10]. Our objective in this work is not identification, but classification of reactive tokens. Still, the prosodic features plays an important role in conveying para-linguistic and non-verbal information. Ward [11] made an analysis of pragmatic functions conveyed by the prosodic features in English non-lexical tokens. We conduct a systematic analysis on both syllabic and prosodic features, considering that Japanese has a more variety of syllabic patterns of reactive tokens than English.

We extracted the following prosodic features for each reactive token: duration, F0 (maximum and range) and power (maximum). The prosodic features are normalized for every person as follow; for each feature, we compute the mean and this mean is subtracted from the feature values.

3. Analysis on Occurrence Statistics and Conversation Mode

First, we investigate occurrence statistics of reactive tokens and their relationship with the conversation mode. We hypothesize that more reactive tokens are observed in the QA mode by the initiative person of the audience because he should be more interested in the answers to his questions.

We define the frequency of reactive tokens as their occurrence count divided by the duration (min.) and compute it for each conversation mode (EX and QA). We further classified the QA mode into QA_self (initiated by the person of the audience himself) and QA_other (initiated

Table 3: Relationship of frequency of reactive tokens (count/min.) with conversation mode

session	audience	EX	QA	QA_self	QA_other
1	1	5.2	10.7	14.0	9.4
	2	10.7	12.2	13.2	10.1
2	1	9.1	11.9	12.4	9.8
	2	9.5	11.4	15.8	10.2
3	1	7.3	9.5	13.7	8.0
	2	7.6	10.4	10.7	9.6
4	1	10.6	19.3	19.6	18.5
	2	6.4	6.7	6.8	6.5
average		8.3	11.5	**13.3**	**10.3**

Table 4: Statistics of reactive tokens in conversation modes

	total count	QA_self (count/min.)	QA_other (count/min.)	EX (count/min.)
hai	188	**2.5**	**0.5**	**1.0**
u:N	544	2.9	3.1	2.8
uN	356	1.7	2.0	2.1
huN	166	1.3	0.4	1.0
hu:N	114	0.7	1.1	0.4
he:	78	0.6	0.7	0.2
a:	59	0.5	0.4	0.1
haa	55	0.5	0.5	0.2
ee	38	0.5	0.3	0.1
aa	23	0.4	0.1	0.1
ha:	21	0.1	0.2	0.1

by the other person of the audience). Table 3 shows the statistics for these modes. It is apparent that the reactive tokens are more frequent in the QA mode, especially in the QA_self mode.

The next question is whether the frequent reactive tokens are really made by the interest in the corresponding segments of the speech. In order to get some clue, we investigate the frequency of each syllabic pattern, as shown in Table 4. We can see that majority of the increase in the QA_self mode is "*hai* (yes)", and there is not a significant difference between the QA_self mode and the EX mode for other kinds of reactive tokens.

Since the major function of "*hai*" is presumably acknowledgment, suggesting "I hear/understand you", it is not appropriate to conclude that the increase of the token means a higher interest level. Instead, it is possible to attribute the phenomena to the role in the conversation, that is the person who raised a question should have the courtesy to acknowledge the answer.

4. Analysis on Prosodic Features and Interest Level

Next, we incorporate prosodic features and focus on the relationship with the interest level of the audience. Each prosodic feature was normalized by subtracting its mean for every person. Table 5 lists the mean and standard deviation (SD) after mean normalization for all reactive tokens concerned. We can see three tokens of "*hu:N*", "*he:*" and "*a:*" have two or more prosodic features with a significantly larger variation (SD marked with bold font). They should have a larger capacity to convey non-verbal information such as the interest level with prosodic features [1]. On the other hand, entries in the lower half of Table 5 such as "*hai*" and "*huN*" do not have such prominent prosodic features since they are mainly used for acknowledgment.

Thus, we select the three tokens "*hu:N*", "*he:*" and "*a:*" for investigation of the relationship with the interest level of the audience. We hypothesize that the audience express their interest with specific kinds of reactive tokens and specific prosodic patterns, thus we designed an experiment to identify the effective combinations. For each reactive token (syllabic pattern) and for each prosodic feature, we picked up top-ten and bottom-ten samples, i.e. samples that have largest/smallest values of the prosodic feature. In theory, we had to prepare 240 samples (= 3 kinds $\times$ 4 features $\times$ 2 (top/bottom) $\times$ 10), but many samples were shared by different features, so 148 samples were actually selected in total. For each of them, an audio file is segmented from the corpus to cover the reactive token and its preceding clause unit.

Then, we had five subjects to listen to the audio samples and evaluate the audience's state of the mind. We prepared twelve items to be evaluated in a scale of four ("strongly feel" to "do not feel"), among which two items are related to the interest level and other two items are related to the surprise level [2]. Table 6 lists the results (marked by "*") that have a statistically significant ($p < 0.05$) difference between top-ten and bottom-ten samples.

It is observed that prolonged "*hu:N*" means interest and surprise while "*a:*" with higher pitch or larger power means interest. On the other hand, "*he:*" can be emphasized in all prosodic features to express interest and surprise. Although these findings of selective usage of reactive tokens are understandable for native Japanese speakers, this is the first report of a systematic experiment to our knowledge.

5. Conclusions

We have investigated the role of Japanese reactive tokens (*Aizuchi*) in terms of the response to the presentation, in order to explore the feasibility of detecting the interest

[1] "*haa*", "*aa*", and "*ha:*" also have prominent prosodic features, but we expect that they are similar to "*a:*" and we use "*a:*" for the following analysis because it has the largest number of samples among them.

[2] We used different Japanese wording for interest (「興味」,「関心」) and for surprise (「驚き」, 「意外」) to enhance the reliability of the evaluation; we adopt the result if the two matches.

Table 5: Statistics of prosodic features of reactive tokens (values are normalized by subtracting the mean for every speaker)

	count	duration (sec.)		F0 max (Hz)		F0 range (Hz)		power (db)	
		mean	SD	mean	SD	mean	SD	mean	SD
hu:N	114	0.32	**0.44**	7	22	4	**38**	-1.2	4.3
he:	78	0.65	**0.54**	4	**34**	4	**41**	2.7	5.4
a:	59	0.28	0.37	8	**35**	13	**39**	6.5	**6.4**
haa	55	-0.20	0.24	10	**35**	-1	36	4.4	**6.3**
aa	23	-0.19	0.17	7	30	-5	**38**	7.9	**6.3**
ha:	21	0.58	**0.65**	1	32	-4	30	3.9	4.8
ee	38	-0.20	0.10	-12	31	-19	37	2.9	5.5
huN	166	-0.16	0.31	3	25	-9	21	-2.7	4.1
hai	188	-0.29	0.19	-2	28	-8	24	6.3	5.8
uN	356	0.22	0.15	-1	25	-4	30	-2.9	4.9
u:N	544	0.13	0.27	5	27	9	35	-1.5	4.6

Table 6: Significant combinations of syllabic and prosodic patterns

		interest	surprise
hu:N	duration	*	*
	F0 max		
	F0 range		
	power		
he:	duration	*	*
	F0 max	*	*
	F0 range		*
	power	*	*
a:	duration		
	F0 max	*	
	F0 range		
	power	*	

level of the audience. First, we made an occurrence frequency analysis, which suggests that the most typical reactive token "*hai* (yes)" is mainly used for acknowledging utterances, and that there are specific kinds of reactive tokens which can be used for expressing non-verbal information. Then, we made a prosodic analysis and identified effective combinations of the syllabic patterns and prosodic features which express interest and surprise.

Ward [11] argued that pitch height correlates with the degree of interest in English non-lexical tokens. The analysis may be applied to the token "*a:*" which is similar to "uh-" in English. In Japanese, however, there are a dozen of particular syllabic and prosodic patterns to be related with specific functions, part of which were found in this study.

We hope that this information will be useful for identifying "hot-spots" in the presentation and indexing for efficient access to the archives. We are studying automatic detection of these acoustic events [12], which will be integrated with the results of this study. The future work includes integration with visual information such as nodding, which is regarded as another backchannel.

6. References

[1] Sadaoki Furui and Tatsuya Kawahara. Transcription and distillation of spontaneous speech. In J.Benesty, M.M.Sondhi, and Y.Huang, editors, *Springer Handbook on Speech Processing and Speech Communication*, pages 627–651. Springer, 2008.

[2] T.Kawahara, M.Hasegawa, K.Shitaoka, T.Kitade, and H.Nanjo. Automatic indexing of lecture presentations using unsupervised learning of presumed discourse markers. *IEEE Trans. Speech & Audio Process.*, 12(4):409–419, 2004.

[3] S.Furui, T.Kikuchi, Y.Shinnaka, and C.Hori. Speech-to-text and speech-to-speech summarization of spontaneous speech. *IEEE Trans. Speech & Audio Process.*, 12(4):401–408, 2004.

[4] S.Renals, T.Hain, and H.Bourlard. Recognition and understanding of meetings: The AMI and AMIDA projects. In *Proc. IEEE Workshop Automatic Speech Recognition & Understanding*, 2007.

[5] E.Shriberg, R.Dhillon, S.Bhagat, J.Ang, and H.Carvey. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proc. SIGDial*, pages 97–100, 2004.

[6] F.Yang, G.Tur, and E.Shriberg. Exploiting dialog act tagging and prosodic information for action item identification. In *Proc. IEEE-ICASSP*, pages 4941–4944, 2008.

[7] T.Kawahara, H.Setoguchi, K.Takanashi, K.Ishizuka, and S.Araki. Multi-modal recording, analysis and indexing of poster sessions. In *Proc. INTERSPEECH*, pages 1622–1625, 2008.

[8] B.Wrede and E.Shriberg. Spotting "hot spots" in meetings: Human judgments and prosodic cues. In *Proc. EUROSPEECH*, pages 2805–2808, 2003.

[9] D.Gatica-Perez, I.McCowan, D.Zhang, and S.Bengio. Detecting group interest-level in meetings. In *Proc. IEEE-ICASSP*, volume 1, pages 489–492, 2005.

[10] A.Gravano, S.Benus, J.Hirschberg, S.Mitchell, and I.Vovsha. Classification of discourse functions of affirmative words in spoken dialogue. In *Proc. INTERSPEECH*, pages 1613–1616, 2007.

[11] N.Ward. Pragmatic functions of prosodic features in non-lexical utterances. In *Speech Prosody*, pages 325–328, 2004.

[12] K.Sumi, T.Kawahara, J.Ogata, and M.Goto. Acoustic event detection for spotting hot spots in podcasts. In *Proc. INTERSPEECH*, pages 1143–1146, 2009.