Detection of Hot Spots in Poster Conversations based on Reactive Tokens of Audience

Tatsuya Kawahara, Kouhei Sumi, Zhi-Qiang Chang, Katsuya Takanashi

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

Abstract

We present a novel scheme for indexing "hot spots" in conversations, such as poster sessions, based on the reaction of the audience. Specifically, we focus on laughters and non-lexical reactive tokens, which are presumably related with funny spots and interesting spots, respectively. A robust detection method of these acoustic events is realized by combining BIC-based segmentation and GMM-based classification, with additional verifiers for reactive tokens. Subjective evaluations suggest that hot spots associated with reactive tokens are consistently useful while those with laughters are not so reliable. Furthermore, we investigate prosodic patterns of those reactive tokens which are closely related with the interest level.

Index Terms: audio indexing, acoustic event detection, hot spots, reactive token, prosody

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. These works try to provide structural understanding of multi-party conversations.

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 "interactive" characteristics in presentations. Typically, a presenter explains his/her 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 studying 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 and laughters. By reactive tokens (*Aizuchi* in Japanese), we mean the listener's verbal short response, which expresses his/her state of the mind during the conversation. Its prototypical lexical entries include "*hai*" in Japanese and "yeah" or "okay" in English, but many of them are non-lexical and used only for reactive tokens, such as "*hu:n*", "*he:*" in Japanese and "wow", "uh-huh" in English. In this study, we focus on the latter kind of reactive tokens, which are not used for simple acknowledgment.

We hypothesize that the audience signals their interest level with this kinds of non-lexical reactive tokens. And 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. We also presume that people would be interested in the funny spots where laughters were made. In this work, we define those spots which induced (or elicited) laughters and non-lexical reactive tokens, as hot spots, [1] and investigate their automatic detection and effectiveness for audio indexing.

In this paper, we first describe the corpus of poster sessions and its annotations in Section 2. Then, we

[1] Wrede et al.[8][9] defined "hot spots" as the regions where two or more participants are highly involved in a meeting. Our definition is different from it.

present our approach to detect laughters and reactive tokens for extracting hot spots in Section 3. Subjective evaluations of the detected hot spots are reported in Section 4. Further detailed analyses on reactive tokens are made in Section 5.

2. 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 eight poster sessions, in which the presenters and audiences are different from each other. In each session, one presenter had prepared a poster on his/her 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 20-30 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. We also annotated laughters and reactive tokens manually. Fillers were also separately annotated. They are usually followed by utterances by the same speaker, while reactive tokens are uttered by themselves. Reactive tokens used in backchannels, typically "*hai*" in Japanese and "yeah" or "okay" in English, suggest that the listener is understanding what is being said, and also that the current speaker can continue to utter by keeping the dialogue turn.

In this study, we focus on non-lexical reactive tokens, for example, "*hu:n*", "*he:*" in Japanese. They cannot be used for simple acknowledgement and presumably related with the state of the mind of the listener. These can be articulated with a variety of prosodic patterns; they can be prolonged to an arbitrary length.

3. Automatic Detection of Laughters and Reactive Tokens

3.1. Detection Method

Detection of laughters has been addressed by several studies [10][11][12]. Typically, a dedicated classifier such as GMM and SVM is prepared for discriminating laughters against speech. On the other hand, studies on detecting reactive tokens is limited. Ward [13] investigated prosodic patterns of reactive tokens, but did not conduct automatic detection. Other works [6][14] focused on distinction of affirmative answers "yes" and tokens used in backchannels. In Japanese, there are a variety of syllabic patterns in reactive tokens, including both lexical and non-lexical tokens.

We have designed a scheme for acoustic event detection in audio recordings of conversations, and applied it to podcast content [15]. In this work, we revise it for detection of laughters and reactive tokens in poster conversations. The scheme is based on a combination of BIC-based segmentation and GMM-based classification.

Every 10-msec speech input frame is parameterized into 26-dimensional acoustic features of MFCCs, ΔMFCCs, power and Δpower. Segmentation based on BIC (Bayesian Information Criterion) [16] is first applied to detect change points in speakers or acoustic events such as laughter and noise. In the BIC, the only parameter, called λ, which balances the likelihood and data complexity, virtually controls the number of generated segments. We proposed a method to automatically determine this value based on Gaussian distributions of the corresponding audio category (speech, music, speech+music) [15].

Then, for each segment, classification based on GMM (Gaussian Mixture Model) is applied. We prepared GMMs for five classes of male speech, female speech, noise, laughter and reactive tokens. A newspaper reading corpus (JNAS) was used for training the speech and noise models, and podcast data for the other two classes. Note that there was not matched large-scale training data for the poster conversations in this experiment. Laughters are detected with this GMM-based classification.

Reactive tokens are more difficult to detect, because they are much similar to normal speech in terms of acoustic characteristics, especially MFCC features. Thus, we incorporate two additional processes to verify the candidates of reactive tokens hypothesized by GMM-based classification. One is the filled pause detector which considers monotonousness of spectral and pitch patterns [17]. The other is an ASR system trained with the CSJ. It is used to filter out filled pauses, which are lexically included in the CSJ. In summary, we detect reactive tokens only when supported by all the following three classifiers.

- dedicated GMM
- filled pause detector (to reject normal speech)
- speech recognizer (to reject fillers)

3.2. Evaluation of Detection Accuracy

We evaluated detection accuracy of laughters and reactive tokens using the eight poster sessions. The results are shown in Table 1 with evaluation measures of recall, precision and F-measure. Here, the F-measure is defined with a double weight on precision, because there are a number of indistinct laughters and reactive tokens, which are hard to recall and not useful for indexing.

As shown in Table 1, overall recall is not high, but we could detect most of the distinct events such as loud laughters and long reactive tokens. We expect these distinct events are more related with the hot spots than subtle events. The frame-wise classification accuracy among five GMM classes was 82.3%.

Table 1: Detection accuracy of laughters and reactive tokens

	recall	precision	F-measure
laughter	0.419	0.750	0.648
reactive token	0.439	0.707	0.630

4. Subjective Evaluation of Detected Hot Spots

Based on the detected laughters and reactive tokens, we define hot spots corresponding to these two kinds of events. Specifically, hot spots were labeled for utterances which induced (or elicited) the events. The segments are defined by utterance units, i.e. made of a couple of utterances, with a maximum duration determined by a threshold. Thus, the BIC-based segmentation is used for this indexing. We set the maximum duration threshold value to 20 sec.

We made subjective evaluations on the hot spots indexed in this manner. We had four subjects, who had not attended the presentation nor listened the recorded audio content. They were asked to listen to each of the segmented hot spots in the original time sequence, and to make evaluations on the questionnaire, as below.

Q1: Do you understand the reason why the reactive token/laugher occurred?
Q2: Do you find this segment interesting/funny?
Q3: Do you think this segment is necessary or useful for listening to the content?

The result on Question 1 (percentage of "yes"), summarized in Table 2, suggests the ratio of appropriate hot spots or "precision" among the detected hot spots, because the third person verified the spots were naturally inducing laughters or reactive tokens. The figures labeled "(oracle)" in Table 2 show the result when limited to the segments where laughters or reactive tokens were correctly detected. It is confirmed that a large majority of the detected spots are appropriate. There are more "false" detections for the segments accompanying laughters; many laughters were socially made to relax the participants in the poster conversations.

The answers to Questions 2 and 3 are more subjective, but suggest the usefulness of the hot spots. It turned out that only a half of the spots associated with laughters are funny for the subjects (Q2), and they found 35% of the spots not funny. The result suggests that feeling funny largely depends on the person. And we should note that there are not many funny parts in the poster sessions by nature.

On the other hand, more than 90% of the spots associated with reactive tokens are interesting (Q2), and useful or necessary (Q3) for the subjects. The result supports the effectiveness of the hot spots extracted based on the reaction of the audience.

Table 2: Ratio of appropriate hot spots among detected spots ("precision")

	precision (oracle)
spots accompanying laughter	74.7% (89.2%)
spots accompanying reactive token	86.5% (95.2%)

5. Prosodic Analysis of Reactive Tokens

In the system described above, we tried to detect all non-lexical reactive tokens without considering their syllabic and prosodic patterns. In this section, we hypothesize that the audience express their interest with specific syllabic and prosodic patterns. Generally, prosodic features play an important role in conveying para-linguistic and non-verbal information. In previous works [6][14], it was reported that prosodic features are useful in identifying backchannels. Ward [13] made an analysis of pragmatic functions conveyed by the prosodic features in English non-lexical tokens.

In this study, we designed an experiment to identify the syllabic and prosodic patterns closely related with the interest level for detection of hot spots. For this investigation, we select three syllabic patterns of "*hu:N*", "*he:*" and "*a:*", which are presumably related with the interest level and also most frequently observed in the corpus, except lexical tokens.

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

For each syllabic kind of reactive token 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 segment is extracted to cover the reactive token and its preceding utterances. This process is similar to the hot spot detection described in the previous sections, but done manually according to the criteria.

Then, we had five subjects to listen to the audio segments 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 3 lists the results (marked by "*") that have a statistically significant

[2] We used different Japanese wording for interest and for surprise to enhance the reliability of the evaluation; we adopt the result if the two matches.

Table 3: Significant combinations of syllabic and prosodic patterns of reactive tokens

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

($p < 0.05$) 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.

It is expected that using this prosodic information will enhance the precision of the hot spot detection. But the tokens with larger power and/or a longer duration is apparently easier to detect than indistinct tokens, and they are more related with the hot spot. This simple principle is consistent with the proposed scheme.

6. Conclusions

We have presented a novel scheme for indexing hot spots during the poster conversations, based on the reaction of the audience. Specifically, we focus on laughters and non-lexical reactive tokens, and implemented a method to automatically detect them.

Detection of laughters is relatively easier, but we found that the detected spots are not necessarily funny or useful, because the evaluation is largely affected by subjects. On the other hand, the spots associated with reactive tokens were shown to be consistently interesting and meaningful. We have further investigated the specific prosodic patterns closely related with the interest level, which would be useful for enhancing the detection performance.

Although the findings on reactive tokens may be somewhat dependent on Japanese language, we expect that the proposed scheme based on reactions during a conversation will be applicable to other languages and settings. The future work includes integration with visual information such as nodding, which is regarded as another form of reaction.

Acknowledgement: This work was supported by JST CREST and JSPS Grant-in-Aid for Scientific Research.

7. 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] L.S.Kennedy and D.P.W.Ellis. Laughter detection in meetings. In *NIST Meeting Recognition Workshop*, 2004.

[11] K.P.Truong and D.A.van Leeuwen. Automatic detection of laughter. In *Proc. EUROSPEECH*, pages 485–488, 2005.

[12] K.Laskowski. Contrasting emotion-bearing laughter types in multiparticipant vocal activity detection for meetings. In *Proc. IEEE-ICASSP*, pages 4765–4768, 2009.

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

[14] 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.

[15] 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.

[16] S.Chen and P.Gopalakrishnan. Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion. In *DARPA Broadcast News Workshop*, pages 127–132, 1998.

[17] M.Goto, K.Itou, and S.Hayamizu. A real-time filled pause detection system for spontaneous speech recognition research. In *Proc. EUROSPEECH*, pages 227–230, 1999.