INTERSPEECH 2009 BRIGHTON

Acoustic Event Detection for Spotting "Hot Spots" in Podcasts

Kouhei Sumi[†], *Tatsuya Kawahara*[†], *Jun Ogata*[‡], *and Masataka Goto*[‡]

[†]Graduate School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan,
[‡]National Institute of Advanced Industrial Science and Technology (AIST)
Tsukuba, Ibaraki 305-8568, Japan

Abstract

This paper presents a method to detect acoustic events that can be used to find "hot spots" in podcast programs. We focus on meaningful non-verbal audible reactions which suggest hot spots such as laughter and reactive tokens. In order to detect this kind of short events and segment the counterpart utterances, we need accurate audio segmentation and classification, dealing with various recording environments and background music. Thus, we propose a method for automatically estimating and switching penalty weights for the BIC-based segmentation depending on background environments. Experimental results show significant improvement in detection accuracy by proposed method compared to when using a constant penalty weight.

Index Terms: acoustic event detection, laughter detection, podcast, Bayesian Information Criterion

1. Introduction

In recent years, there has been an increase of audio media (MP3 audio files, etc.) on the internet, including podcasts, internet radio, and audio blogs. Unlike text-based content, audio content cannot be fully understood if we do not listen to the entirety of the content. Thus, it is difficult to search and browse audio content because speech, music, and sound effects are not visible.

As a solution to this problem, some applications such as Podscope [1], Google Audio Indexing [2], and PodCastle [3] have been developed to search and browse through transcriptions of audio data by applying automatic speech recognition (ASR). In Podcastle, language model is adaptated using RSS metadata, and acoustic model training is enhanced by user-corrected transcripts [1, 2]. Google Audio Indexing transcribes election video material and makes the content searchable [3]. However, even the state-of-the-art ASR systems cannot accurately transcribe all audio data, because their content and recording environments are of a wide variety. They contain not only speech but also music, sound effects, and environmental sound. Another problem for ASR is that most of speech is spontaneous, conversational, and often in dialogue-style.

To circumvent these difficulties in ASR, we focus on audible reactions, present in the content, which have non-verbal information and were ignored in the traditional ASR. For example, laughter occurs after talking about interesting things. Applause often comes after utterances which impress the audience favorably. In addition, a certain kind of reactive tokens in back-channel responses during dialogue suggest a level of interest in the topic of conversation [4]. By detecting these meaningful audible reactions, we expect we can detect "hot spots" without recognizing the speech. Here, we define hot spots as segments in which the listeners would be interested in the audio data [4], and hot spot candidates exist right before these events as shown in Figure 1. Our goal is to detect not only the audible reactions but also segment the preceding utterances, which potentially make up hot spots.

Figure 1: Audible reaction and hot spot

The rest of the paper is organized as follows: section 2 describes the conventional segmentation method based on BIC and the issues involved in detecting acoustic events in podcasts. In section 3, we present a method for optimally deciding penalty weights for the BIC-based method. Our proposed system is described in section 4, and the experimental results are reported in section 5. Finally, conclusions and future work are discussed in section 6.

2. Acoustic event detection in podcasts

In this work, we focus on laughter and reactive tokens which are frequently observed acoustic events in podcasts and presumed to be closely related to hot spots. We also address the problem of segmentation of speech by different speakers and cassification of speech, music, and speech mixed with music. We regard all these categories as acoustic events in this paper.

2.1. Audio segmentation and classification in podcasts

The conventional framework for acoustic event detection (AED) or audio segmentation and classification is based on either explicit models or metrics. These methods are mainly studied targeting at broadcast news and recordings of meetings. Whereas broadcast news rarely contain dialogue, there are many dialogue-style programs in podcasts, thus containing a number of short utterances. Moreover, background music is often used in podcasts while it never appears in meetings. These issues combined make it very difficult to conduct audio segmentation and classification of podcasts.

[1] http://www.podscope.com/
[2] http://labs.google.com/gaudi
[3] http://podcastle.jp/

[4] Wrede *et al.* [5] used the term "hot spots" for regions in which two or more participants are highly involved in a meeting discussion. Our definition is different from this.

2.1.1. Model-based methods

In the model-based methods, classifiers are trained for a given set of pre-determined acoustic classes, using Gaussian Mixture Models (GMMs), Hidden Markov Models (HMMs), or Support Vector Machines (SVMs) [6, 7]. For example, Knox *et al.* presented a method for detection of laughter in meetings using neural networks [8].

In podcasts, we observe frequent speaker changes as well as acoustic events of short duration. Thus, it is not easy to apply explicit models because the decision made with the features of short duration is not reliable, influenced by local variations.

2.1.2. Metrics-based segmentation method

Segmentation of the audio input is accomplished by measuring the "uniformity" or distance with different segmentation models. The methods do not require training data though they cannot classify the segments into some category. One of the most widely-used method is based on the Bayesian Information Criterion (BIC) [9, 10]. BIC is used to determine which of parametric models $M = M_1, M_2, \cdots, M_m$ best represents N data samples $D = D_1, D_2, \cdots, D_N$. According to the BIC theory, the best model maximizes $BIC(M_i)$ as follows:

$$BIC(M_i) = \log P(D_1, D_2, \cdots, D_N | M_i) - \frac{1}{2}\lambda d_i \log N \quad (1)$$

where d_i is the number of parameters of model M_i, and P is the maximum likelihood of the data under model M_i.

BIC-based segmentation is realized by computing the BIC difference between two models: model M_0 where $X = x_k, k = 1, 2, \cdots, N$ is identically distributed to a single Gaussian $N(\mu_0, \Sigma_0)$, and model M_{12} where $x_k, k = 1, 2, \cdots, j$ is drawn from one Gaussian $N(\mu_1, \Sigma_1)$ and $x_k, k = j+1, j+2, \cdots, N$ is drawn from another Gaussian $N(\mu_2, \Sigma_2)$. Then, $\Delta BIC(j) = BIC(M_0) - BIC(M_{12})$ is derived as follows:

$$\begin{aligned}\Delta BIC(j) &= \frac{1}{2}(N \log|\Sigma_0| - j \log|\Sigma_1| - (N-j)\log|\Sigma_2|) \\ &\quad -\frac{1}{2}\lambda(d + \frac{1}{2}d(d+1))\log N \end{aligned} \quad (2)$$

where N is the sample number of the merged segment, d is the dimension of the feature space, and λ is called the penalty weight. If j is chosen such that $j = \arg\max_{1<j<N} \Delta BIC(j)$ and $\Delta BIC(j) > 0$, the frame j is identified as a segment boundary. One problem with this method is that the penalty weight λ is usually task-dependent and must be tuned for every new task [10].

In this work, we adopt a scheme that first applies the BIC-based segmentation and then classifies each segment using the GMMs. For segmentation of podcasts with the BIC-based method, it is hard to fix the appropriate penalty weight because acoustic characteristics are different when background music exists and when it does not. Therefore, we propose to switch the penalty weight λ depending on the audio characteristics.

3. Estimating penalty weight of Bayesian Information Criterion

3.1. Statistical characteristics of speech, music, and their mix

The variance of acoustic features is different for speech, music, and also their mix. Because of a wide variety of instruments and tones, apparently acoustic features in music segments have more variations than in speech segments, and music segments tend to be split too much if we use the same penalty weight λ. On the other hand, when several persons speak in the presence of background music (mix segments), it is hard to find speaker changes because acoustic features are smeared by the background music, usually monotone music, while there can be a variety of music in music-only segments.

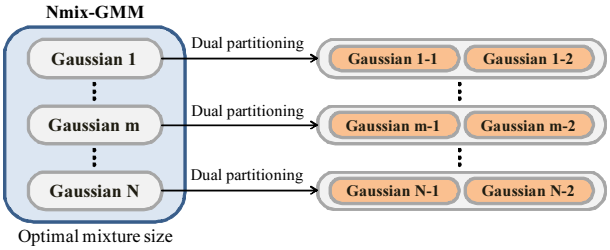

Figure 2: Dual partitioning of each Gaussian

Based on the observation, we set up these global three classes (speech, music and mix) and corresponding classifiers based on GMM as a preprocessing step before the BIC-based segmentation. For each classified segment, the BIC-based segmentation is applied by selecting the penalty weights $\lambda_{spe}, \lambda_{mix}, \lambda_{mus}$ accordingly.

3.2. Automatic estimation of penalty weight of BIC for segmentation

We estimate the appropriate penalty weights in the GMM training phase. When a GMM for each class is properly set up (with the same mixture size), we expect each Gaussian of the GMM represents a proper cluster size, and the penalty weight λ for the BIC-based segmentation should be decided depending on the dispersion of the Gaussian distributions. We assume that when sufficient training data is available and the number of the mixtures is large enough, each Gaussian of the GMM corresponds to a uniform segment and should not be split any more. On the other hand, clusters of a larger size (smaller mixture size) must have been split. Thus, ΔBIC for one single Gaussian of the final GMM and dual partitioned Gaussians from it, as shown in Figure 2, should satisfy the following formula.

$$\begin{aligned}\Delta BIC &= \frac{1}{2}((n_{G_{m1}} + n_{G_{m2}})\log|\Sigma_{G_m}| \\ &\quad -n_{G_{m1}}\log|\Sigma_{G_{m1}}| - n_{G_{m2}}\log|\Sigma_{G_{m2}}|) \\ &\quad -\frac{1}{2}\lambda_m(d + \frac{1}{2}d(d+1))\log(n_{G_{m1}} + n_{G_{m2}}) \approx 0 \end{aligned} \quad (3)$$

where $m = 1, \cdots, N$ is the index of the Gaussians, and $n_{G_{m1}}$ and $n_{G_{m2}}$ are the number of samples contributing to Gaussian m-1 and Gaussian m-2 in the context of the EM estimation. Actually, the prseudo-equation (3) would stand on the average for all Gaussians, $m = 1, \cdots, N$, thus we calculate ΔBIC for all Gaussians of the GMM, and compute the average of λ derived by ΔBIC being equal to zero, to determine the penalty weight for each global class.

3.3. Estimation result

Using the training data described in section 5, the penalty weights $\lambda_{spe}, \lambda_{mus}, \lambda_{mix}$ for the global three classes were esti-

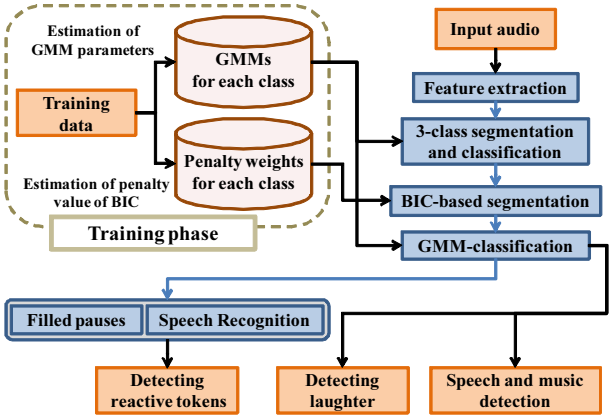

Figure 3: System overview

mated to be 1.68, 3.48, and 1.22, respectively. We set the number of mixtures for each GMM to 256, which is obtained by the optimization by maximizing the BIC (Equation 1) for iterative splits of mixtures from $N = 1, 2, 4,$ to 512. The values of penalty weights are reasonable because music segments, which should not be frequently segmented, obtained a larger penalty weight, and mix segments, which is not easily segmented, received a smaller penalty weight compared to speech segments.

4. Proposed AED system

An overview of our proposed acoustic event detection system is shown in Figure 3. The following sub-sections describe the details of each process.

4.1. Training phase and feature extraction

The parameters of each GMM are trained with feature vectors calculated at frame level. Next, the EM algorithm is used to estimate means, diagonal variance matrices, and weights for Gaussian components. Additionally, we estimated the penalty weights for global classes, as mentioned above.

As acoustic features, we use a 26-dimensional vector which consists of mel-frequency cepstral coefficients (MFCC), delta MFCC, energy, and delta energy. They are calculated with a 25 msec window (10 msec shift), and normalized by the mean and variance for the entire training data.

4.2. Preprocessing and BIC-based segmentation

As a preprocessing step, we coarsely segment and classify the input stream using the GMMs for the global classes (speech-GMM, music-GMM, and mix-GMM). By selecting the different penalty weight for each segment, the BIC-based segmentation is expected to operate appropriately for further segmentation. For accurate segmention, we adopt the variable window scheme described as follows:

1. The window size is initialized by the minimal window size W_{min} (100 frames).
2. If no segment boundary is found in the current window, the size of the window is increased by adding W_{min} until a segment boundary is found.
3. When a segment boundary is found, the next window begins with the detected boundary, using the minimal window size.

Table 1: Baseline training data set

cluster	training data
Speech (male and female)	JNAS [12]
Music	RWC-MDB [13]
Mix (male and female)	JNAS+RWC-MDB
Silence	JNAS, Self-synthesized
Laughter	IMADE corpus [14], Web

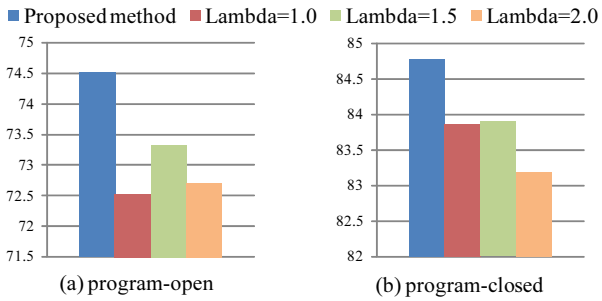

Figure 4: Frame-wise accuracy of eight-class segmentation and classification

4. Until the end of input stream, Step 2 and 3 are repeated.

4.3. Laughter, speech and music segment detection

Each segment obtained by the BIC-based segmentation is classified among laughter, male-speech, female-speech, male-mix, female-mix, music, and silence on the basis of the log likelihood of respective GMMs. Speech segments which are shorter than a constant duration t_{res} (2.0 sec) and have larger log likelihoods of the reactive-token-GMM than a threshold are extracted as candidates of reactive tokens.

4.4. Reactive token detection

Chang *et al.* [4] reported that three reactive tokens of "hu:N", "he:" and "a:" suggest interest level of listeners in Japanese dialogue. Apparently the prolonged vowel can be used as a clue to detect these tokens, thus we utilize the same algorithm as filled paused detection [11]. In addition, there is a lexical difference between these tokens and fillers, so we conduct simple ASR to identify the tokens.

In summary, detection of reactive tokens is realized as follows:

1. Filled pause detection is applied to each candidate, and the segments which include a prolonged vowel are selected.
2. ASR with a lexicon of filler entries and the three tokens is conducted, and reactive tokens are detected.

5. Experimental results

We tested our system on eight podcast episodes by choosing two episodes from four different programs. For training the GMMs, we used the baseline training data set as listed in Table 1. In addition, we prepared two sets of podcasts: (a) program-open set: 19 episodes from programs other than those in the test set, and (b) program-closed set: 23 (19+4) episodes including one previous episode of each program used in the test set.

To evaluate the effectiveness of proposed method, we measure the frame-wise accuracy, laughter detection accuracy, and

Table 2: Detection accuracy of laughter

Measures	R	P	F
Proposed method	65.0	**71.3**	**68.7**
$\lambda = 1.0$	**91.3**	26.4	30.5
$\lambda = 1.5$	74.2	42.2	45.9
$\lambda = 2.0$	60.0	57.5	57.5

Table 3: Detection accuracy of reactive tokens

Measures	R	P	F
Proposed method	34.0	**85.2**	**64.0**
$\lambda = 1.0$	**35.3**	67.9	54.7
$\lambda = 1.5$	33.1	79.3	59.9
$\lambda = 2.0$	29.2	81.2	57.5

reactive token detection accuracy of the following four methods.

1. Proposed method (switching penalty weight λ)
2. Using the constant penalty weight: $\lambda = 1.0$
3. Using the constant penalty weight: $\lambda = 1.5$
4. Using the constant penalty weight: $\lambda = 2.0$

The frame-wise accuracy is the accuracy of detection (segmentation and classification) of the eight classes (male speech, female speech, male mix, female mix, music, silence, laughter, and reactive tokens). In overlapping segments of two different events, if one of the correct classes is included, we regard that the correct class is labeled for the frame. As the detection accuracy of laughter and reactive tokens, we adopted the *recall* R, the *precision* P, and the *F-measure* F. F is defined as follows:

$$\text{F-measure} = \frac{(1+\alpha^2)RP}{R+\alpha^2 P} \tag{4}$$

where α is a measure of the relative importance of recall and precision. In this work, α was set to 0.5 because there are a large number of these events including indistinct ones, and we should focus on the detection of distinct ones rather than recall of everything.

The results are shown in Figure 4 and listed in Table 2 and 3. The proposed method switching the penalty weight λ improved the frame-wise accuracy in both program-open and program-closed cases. The accuracy was drastically improved when using previous episodes of the same program (program-closed case) because there are often same speakers and same music in the same program even if the episode is different.

In detection of laughter and reactive tokens, there was not a significant difference between the program-open case and the program-closed case, thus we report the program-open case. As seen in Table 2, proposed method significantly improved the detection rate of laughter. Although the recall rate was degraded because it was difficult to detect subtle laughter, most loud laughter could be detected. Loud laughter is related more closely to hot spots than subtle laughter, thus we expect that the low recall does not cause major problems in finding hot spots.

The results listed in Table 3 show that the proposed method achieved best performance of detection accuracy of reactive tokens. The main reason for the low recall is strict constraints for rejecting fillers and falters. In fact, we obtained the recall of approximately 70% if we use only filled pause detection. A trade-off between the recall and the precision can be adjusted by tuning the thresholds in filled pause detection and GMM-based classification.

6. Conclusions

We have presented a method for detecting acoustic events in podcast programs based on the BIC-based segmentation which uses dedicated penalty weights for different global classes (speech, music and mix). We also enhance the detection of reactive tokens with the dedicated GMM, filled pause detecter and ASR. Experimental results for real podcast programs demonstrate that our system can detect eight acoustic events with frame-wise accuracy of 74.5%, laughter with F-measure of 68.7%, and reactive tokens with F-measure of 64.0%. It is also shown that the use of different penalty weights is effective in segmentation of podcasts.

Future work includes incorporation of clustering of speech segments for speaker diarization. We also plan to design and implement a "hot spot browser".

7. References

[1] M. Goto, J. Ogata, and K. Eto, "PodCastle: A Web2.0 Approach to Speech Recognition Research," in *Proc. of Interspeech*, pp. 2397-2400, 2007.

[2] J. Ogata, M. Goto, and K. Eto, "Automatic Transcription for a Web 2.0 Service to Search Podcasts," in *Proc. of Interspeech*, pp. 2617-2620, 2007.

[3] C. Alberti, *et al.*, "An Audio Indexing System for Election Video Material," in *Proc. of ICASSP*, pp. 4873-4876, 2009.

[4] Z. Chang, "Analysis on Morphological and Prosodic Features of Aiduti for Indexing Poster Presentations," Masters Thesis, Graduate School of Informatics, Kyoto University, 2009.

[5] B. Wrede and E. Shriberg, "Spotting "Hot Spots" in Meetings: Human Judgments and Prosodic Cues," in *Proc. of Eurospeech*, pp. 2805-2808, 2003.

[6] X. Zhou, *et al.*, "HMM-based Acoustic Event Detection with AdaBoost Feature Selection," *Multimodal Technologies for Perception of Humans*, pp. 345-353, 2008.

[7] A. Temko and C. Nadeu, "Classification of Acoustic Events using SVM-based Clustering Schemes," *Pattern Recognition*, Vol. 39, issue 4, pp. 682-694, 2006.

[8] M. T. Knox and N. Mirghafori, "Automatic Laughter Detection Using Neural Network," in *Proc. of Interspeech*, pp. 2973-2976, 2007.

[9] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion," in *Proc. of DARPA Broadcast News Transcription and Understanding Workshop*, pp. 127-132, 1998.

[10] A. Tritschler and R. Gopinath, "Improved speaker segmentation and segments clustering using the Bayesian Information Criterion," in *Proc. of Eurospeech*, pp. 679-682, 1999.

[11] M. Goto, K. Itou, and S. Hayamizu, "A Real-time Filled Pause Detection System for Spontaneous Speech Recognition," in *Proc. of Eurospeech*, pp.227-230, 1999.

[12] K. Itou, *et al.*, "JNAS: Japanese Speech Corpus for Large Vocabulary Continuous Speech Recognition Research," *J. Acoust. Soc. Jpn. (E)*, Vol. 20, No. 3, pp. 199-206, 1999.

[13] M. Goto, *et al.*, "RWC Music Database: Popular, Classical, and Jazz Music Databases," in *Proc. of ISMIR*, pp.287-288, 2002.

[14] T. Kawahara, *et al.*, "Multi-Modal Recording, Analysis and indexing of Poster Sessions," in *Proc. of Interspeech*, pp. 1622-1625, 2008.