情報処理学会論文誌　Vol. 52　No. 12　3363–3373 (Dec. 2011)

音声会話コンテンツにおける聴衆の反応に基づく音響イベントとホットスポットの検出

河原達也[†1]　須見康平[†1]
緒方淳[†2]　後藤真孝[†2]

ポッドキャストのような音声会話コンテンツの効率的な視聴のために，会話中の聞き手の反応に着目し，その音響イベントの検出に基づいてインデキシングを行う方法を提案する．本研究では，笑い声やあいづちを生起させる箇所（= ホットスポット）が，第三者である視聴者にとっても有益な情報を含んでいると考えて，それらの検出を行った．様々な会話・背景音楽が存在する状況で，このような短い音響イベントを頑健に検出するために，BIC に基づく音響セグメンテーションと GMM によるセグメントの分類，さらに有声休止検出器・音声認識器を組み合わせる．特に，BIC セグメンテーションにおける分割重みのパラメータを，背景音響条件に応じて自動的に推定して切り替える方法を提案する．提案手法により，フレームごとの分類精度および笑い声・あいづちの検出精度が有意に向上した．また，被験者実験によって各ホットスポットの妥当性を評価し，実際に被験者が興味・関心を持つような箇所であることが示された．さらに，これらのホットスポットに基づいて，効率的にコンテンツを視聴するためのインタフェースも作成した．

Detecting Acoustic Events and Hot Spots Based on Audience's Reaction in Conversational Speech Content

Tatsuya Kawahara,[†1] Kouhei Sumi,[†1] Jun Ogata[†2] and Masataka Goto[†2]

We present a novel scheme for indexing "hot spots" in conversational speech content, such as podcasts, 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. We also propose a novel method for automatically estimating and switching a penalty weight for the BIC-based segmentation according to the background acoustic environment. Experimental results show a significant improvement in detection accuracy by the proposed method. Furthermore, subjective evaluations suggest that hot spots associated with these acoustic events are mostly useful, attracting the viewer's interest. Finally, we design a new interface "podspotter", which provides efficient access to speech content based on these results.

1. はじめに

情報通信環境の飛躍的な発展により，多様な音声コンテンツが容易に視聴できるようになってきている．実際に，インターネット上にはポッドキャストや Web ラジオ，ボイスブログといったコンテンツが多く存在し，また様々なトーク番組などもストリーミングされるようになっている．このような大量の音声コンテンツに対して，視聴者が興味や関心に応じてスムーズにブラウジングできることが望ましい．ところが，テキストや画像と異なり，音声は不可視なメディアで，一覧性に乏しく，高速スキャンも困難なため，コンテンツに含まれている内容や欲しい情報の所在を速やかに把握するのが容易でない．したがって，音声から意味のある発話をあらかじめインデキシングできれば，利便性が大きく向上する．そのために，音声認識と自然言語処理の技術を用いたインデキシング・重要文抽出・要約などに関する研究が行われている[4]．しかし，様々な背景音や雑音の重畳した自由発話音声に対して現状の音声認識技術では精度が十分ではなく，また自然言語処理で想定されている文や句の構造も必ずしも明確でない．

これに対して本論文では，人が会話中に自然に起こす反応によって生じる非言語的な音響イベント（= 音リアクションイベント）を手がかりとして，視聴者にとって意味のある箇所を抽出するアプローチを述べる．会話音声には，言語的な情報だけでなく，心的態度や感情などのテキストでは表現できない非言語情報も多く含まれている．特に少人数の会話の場においては，聞き手（会話参与者）が受けた印象を音リアクションイベントによって頻繁に表出するため，これを自動的にインデキシングできれば視聴に際して有用であると考える．たとえば，Web 上の動画共有サイトでは視聴者が自身の反応をアノテーションできる機能を提供している場合があるが，当該コンテンツに登場している会話参与者の反応をあらかじ

†1 京都大学情報学研究科
School of Informatics, Kyoto University
†2 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)

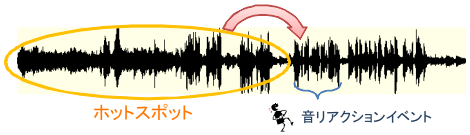

図 1　音リアクションイベントとホットスポット
Fig. 1　Acoustic events and hot spots.

め自動的にアノテーションできれば効果的であると考えられる．

本研究では，音声会話中の音リアクションイベントとして，笑い声とあいづちを対象とする．笑い声は独話や会話中でおもしろいと思わせる発話が出現した直後に起こり，あいづちは発話に対する聞き手の関心の度合いや心的状態（納得や驚きなど）を表すと考えられる．これらの音リアクションイベントは会話参与者間のインタラクションが活発に行われた箇所に出現しやすいため，笑い声やあいづちを生起させる発話は，会話中で特に意味のある箇所に該当する可能性が高いと考えられる．本研究ではこのような箇所を「ホットスポット」と定義し，笑い声やあいづちなどの音リアクションイベントを検出し，それによってホットスポットを抽出する枠組みを提案する（**図 1** 参照）．

この枠組みをポッドキャストに適用し，音響イベント検出の性能を評価するとともに，得られたホットスポットの有用性について被験者実験によって評価する．さらに，ホットスポットとして抽出された箇所の提示や，それに基づくブラウジング機能を持つ視聴インタフェースを作成したので，その紹介を行う．

2. 音声コンテンツに対するインデキシング

多様な音声コンテンツへの効率的なアクセスを目的として，音声認識に基づく言語情報を利用したインデキシングや，非言語情報を用いたインデキシングに関する研究が行われている．それぞれの研究について概観し，本研究の位置づけについて述べる．

2.1 音声認識に基づく閲覧・検索と要約

音声認識に基づいて Web 上の音声コンテンツの閲覧・検索を可能にするサービスとして，PodCastle [⋆1] や Google Audio Indexing がある．PodCastle では，テキストでの閲覧・検索だけでなく，音声認識誤りを誰でも容易に修正できるインタフェースを提供している[15]．Google Audio Indexing では，米国の選挙演説のコンテンツを対象として音声認識を実現し，検索と選択的視聴を可能にした[1]．

一方，講演や講義を対象として，音声認識に基づいて重要文の抽出や要約を行う研究も多数行われている[4),26)]．言語情報からの重要度計算には，tf-idf などのキーワードの統計量，話し言葉に特有の談話標識や手がかり語の情報[8)]，さらには講演全体の内容との相関などの情報が用いられている．また，韻律情報を利用することも検討されている．

2.2 韻律情報を用いた盛り上がり区間の検出

これに対して，言語情報を利用しないインデキシングに関する研究も行われている．Wrede ら[20)] は，ミーティング中で 2 人以上が会話に深く関わり，白熱した議論が行われている区間を “Hot Spot” と定義し，会話参与者の会話に対する関与の度合い（白熱度合い）が，ピッチやパワーといった韻律情報の偏差によって特徴づけられると報告している．これに対して本研究では，「ホットスポット」を，音リアクションイベントを生起させる箇所として定義するが，会話参与者間のインタラクションが活発に行われた箇所という点で共通している．また，Kennedy ら[10)] は，ミーティング中で強調された発話を韻律的な特徴量で特定する手法を提案しており，Gatica-Perez ら[5)] は，ミーティングにおいてグループ全体としての関心度の高まりを韻律情報と映像情報を組み合わせて検出することに取り組んでいる．

2.3 本研究の位置づけ

音声会話コンテンツには発話者の音声以外にも，背景音楽や音響効果，環境音・雑音などの多くの音が存在する．また，自由発話音声では音響的・言語的な変動も大きく，多人数会話では頻繁な話者交替・同時発話もみられる．一般的にこのような音声に対して実用的に十分な音声認識精度を得るのは容易ではない．したがって閲覧・検索はもとより，重要文抽出や要約についても，ポッドキャストなどの音声会話コンテンツで実現するのは難しい．また，特にエンターテイメント目的のコンテンツなどでは，重要かどうかが視聴者の主観に大きく影響されるほか，講演や講義のように，必ずしも話が整理・構造化されていないことから重要文抽出や要約自体が困難な場合もある．

従来の重要文抽出・要約は，発話者の音声認識や発話内容の解析に立脚しているが，本研究で対象とする音声会話には，発話者（話し手）以外にも会話参与者（聞き手）が存在する．聞き手が話し手の発話内容に対して様々な反応を示すことで会話は成立しており，特に興味・関心を持った発話に対しては，聞き手が大きな反応を示す場合が多い．そのような発話は，このコンテンツを後で視聴する第三者にとっても有益な情報を含んでいると期待できる．

そこで本研究では，聞き手の反応を表す非言語情報に着目したインデキシングを考える．

⋆1 http://podcastle.jp/

具体的には，盛り上がり区間の検出に関する研究のようにピッチやパワーといった単純な韻律情報を用いるのではなく，聞き手の反応・心的状態を表す非言語情報である笑い声とあいづちに着目する．これにより，単にインタラクションが活発に起こっているというだけでなく，インデキシングされた発話がどのような意味を持つかを表すラベルを付与できる．たとえば，笑い声は通常「おもしろい」という反応を表すため，笑い声に基づくホットスポットであると示すことによって，視聴者はおもしろい箇所が含まれていると想定することができる．

3. ホットスポット抽出のための音響イベント検出

音響イベント検出は，音響信号中に存在する音声・音楽・環境音などの様々な種類の音を「音響イベント」とし，それぞれの該当区間を検出する処理である．本研究では，笑い声やあいづちも音響イベントの 1 つとする．

音響イベント検出は，音響信号を各イベントごとに分割（セグメンテーション）する処理と，各セグメントがどのイベントであるかを分類する処理からなる．検出手法としては，GMM（Gaussian Mixture Model）や HMM（Hidden Markov Model），SVM（Support Vector Machines）などを用いたモデルベースの手法と，入力区間の類似度・距離に基づいてセグメンテーション・クラスタリングを行うメトリックベースの手法の 2 つに分類される．モデルベースの手法では，事前の学習データ収集が必要であり，学習データによくマッチした入力については高い精度が得られる反面，学習データでカバーされていない音響イベントは検出できない．本研究で対象とするポッドキャストでは，様々な背景音楽や音響効果が用いられるため，一般的なモデル学習は容易でない．一方，メトリックベースの手法では入力中の類似度のみを見るため，様々な対象に頑健に適用できる．その代表的な手法は，Bayesian Information Criterion（BIC）に基づくもので，放送ニュースや会議音声を対象とした音声認識の前処理として，背景音楽の区間や同一話者の発話区間に分割するために広く用いられている[6),24)]．BIC は話者インデキシングでも広く利用されている[14),17)]．ただし，メトリックベースの手法では，各セグメントがどのようなイベントであるかを識別（ラベル付与）することはできない．

本研究では，背景音楽などに対応しながら短時間の音響イベントを頑健に検出するために，メトリックベースのセグメンテーションとモデルベースの分類を組み合わせる．具体的には，音響信号に対して BIC に基づく音響セグメンテーション（BIC セグメンテーション）を適用することでおおまかなイベントごとに区切られたセグメントを求め，GMM による各セグメントの精密な分類を行うことで，笑い声・あいづちを含む音響イベントを検出する．特にポッドキャストでは，放送ニュースと異なり，発話中も長時間にわたって背景音楽が重畳される場合が多い．この場合，音響的な特徴が背景音がある場合とない場合で大きく変化するため，類似度に基づく BIC セグメンテーションを頑健に動作させることが困難になる．そこで，BIC セグメンテーションにおいて分割のされやすさを制御するパラメータの値を，音響条件ごとに自動的に切り替える手法を提案する．

笑い声の検出に関しては，ミーティング音声を対象にいくつか研究が報告されている．GMM や HMM などの統計モデルを用いる手法[13),18)]が一般的であり，ニューラルネットワークを用いた手法[12)]や SVM を用いた手法[11)]も提案されている．一方，あいづちに関して明示的に検出を行った研究はほとんどない．Ward[19)]はあいづちの韻律特徴の分析を行っている．また，英語において返答とあいづちの両方に用いられる “yes” について，両者の区別を韻律情報を用いて行った研究[7)]がある．日本語のあいづちについても，通常の単語と異なる韻律的な特徴を有すると考えられるが，音韻的にも通常の単語と異なるパターンが多数あるので，本論文ではこれらの情報を総合した手法を提案する．

3.1 BIC セグメンテーションと分割重み推定

本節では，BIC セグメンテーションについて説明を行った後，提案する分割重みの自動推定手法について述べる．

3.1.1 BIC セグメンテーション

BIC[16)]はモデル選択の基準であり，各モデル $M_1, M_2, \cdots, M_m$ に対して，データセット $\{D_1, D_2, \cdots, D_N\}$ が与えられたとき，モデル M_i の BIC 値は以下のように定義される．

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

ここで，d_i はモデル M_i の自由パラメータ数であり，P はデータセットに対するモデル M_i の尤度である．このとき，BIC 値が最大になるものを最適なモデルとして選択する．

BIC セグメンテーション[2),3)]では，ある入力区間（N サンプル）に対して，それを 1 つのモデル $M_0 = \mathsf{N}(\mu_0, \Sigma_0)$ で表した場合の BIC 値 $BIC(M_0)$ と，ある点 j（$1 < j < N$）を境界とした 2 つのモデル $M_{12} = \{M_1, M_2\} = \{\mathsf{N}(\mu_1, \Sigma_1), \mathsf{N}(\mu_2, \Sigma_2)\}$ で分割して表した場合の BIC 値 $BIC(M_{12})$ を比較する．モデル化にはガウス分布を用いるのが一般的であり，入力特徴量系列を $X = \{x_1, \cdots, x_N\}$ とすると，それぞれ $M_0 : X = \{x_1, \cdots, x_N\} \sim \mathsf{N}(\mu_0, \Sigma_0)$，$M_{12} : \{x_1, \cdots, x_j\} \sim \mathsf{N}(\mu_1, \Sigma_1); \{x_{j+1}, \cdots, x_N\} \sim \mathsf{N}(\mu_2, \Sigma_2)$ となる．このとき，$BIC(M_0)$ は次式のようになる．

$$BIC(M_0) = -\frac{d}{2}N\log 2\pi - \frac{N}{2}\log|\Sigma_0| - \frac{N}{2} - \frac{1}{2}\lambda\left(d + \frac{1}{2}d(d+1)\right)\log N \quad (2)$$

なお，d は特徴量ベクトルの次元数である．$BIC(M_{12})$ についてはモデルパラメータ数が 2 倍となるが，同様に以下のようになる．

$$BIC(M_{12}) = -\frac{d}{2}N\log 2\pi - \frac{j}{2}\log|\Sigma_1| - \frac{N-j}{2}\log|\Sigma_2| - \frac{N}{2} - \lambda\left(d + \frac{1}{2}d(d+1)\right)\log N \quad (3)$$

したがって，これらの差分 $\Delta BIC(j)$ は次のようになる．

$$\begin{aligned}\Delta BIC(j) &= BIC(M_{12}) - BIC(M_0)\\ &= \frac{1}{2}(N\log|\Sigma_0| - j\log|\Sigma_1| - (N-j)\log|\Sigma_2|)\\ &\quad - \frac{1}{2}\lambda\left(d + \frac{1}{2}d(d+1)\right)\log N \end{aligned} \quad (4)$$

この λ を**分割重み**と呼ぶ．このとき，

$$j = \arg\max_j \Delta BIC(j) > 0 \quad (5)$$

であれば，点 j を分割境界とする．この手法は事前の学習を必要としないが，分割のされやすさを制御するパラメータ λ の値をタスクごとに調整する必要があるという問題がある[3]．

3.1.2　分割重みの自動推定

本研究では音響環境の大分類（タスク）ごとに，音響特徴量の分布を GMM で表現し，その GMM 中の各ガウス分布を用いて，上記の分割重み λ の適切な値を推定する方法を提案する．GMM が理想的な条件で学習されているとき，各要素分布は理想的な単一ガウス分布になっていると考えられる．すなわち，十分な学習データが存在し，混合数も十分大きければ，求まった GMM の各ガウス分布は，それ以上分割できない均一な 1 つのセグメントととらえることができる．したがって，このようなガウス分布を分割しないような λ は，同種類のタスクにおいて，未知の入力のセグメンテーションにおいても有効に機能すると期待できる．そこで本研究では，このようなガウス分布の集合を用いて BIC の分割重み λ を決定する．

混合数 M で学習された GMM 中のあるガウス分布 G_m と，それをさらに 2 つに分割（+ 再推定）した場合のガウス分布 G_{m1}，G_{m2} に対する ΔBIC は，式 (4) から次のようになり，これが 0 となる方程式を立てる．

$$\begin{aligned}\Delta BIC &= \frac{1}{2}((n_{G_{m1}} + n_{G_{m2}})\log|\Sigma_{G_m}| - n_{G_{m1}}\log|\Sigma_{G_{m1}}| - n_{G_{m2}}\log|\Sigma_{G_{m2}}|)\\ &\quad - \frac{1}{2}\lambda_m\left(d + \frac{1}{2}d(d+1)\right)\log(n_{G_{m1}} + n_{G_{m2}}) \approx 0\end{aligned} \quad (6)$$

ここで，$m = 1, \cdots, M$ はガウス分布のインデックスを表し，Σ_{G_m}，$\Sigma_{G_{m1}}$，$\Sigma_{G_{m2}}$ は各ガウス分布の共分散行列である．また $n_{G_{m1}}$ と $n_{G_{m2}}$ は，EM アルゴリズムによるパラメータ推定の過程で得られる G_{m1} と G_{m2} に寄与するサンプル数（EM カウント）である．$m = 1, \cdots, M$ のすべてのガウス分布に対して，式 (6) のように，ΔBIC を計算し，これが 0 と等しいとして得られる λ_m を各々について求める．最終的にそれらの平均を計算し，得られた値 λ を分割重みの推定値とする．

$$\lambda = \frac{1}{M}\sum_{m=1}^{M}\lambda_m \quad (7)$$

3.2　あいづちの検出について

本節では，本研究で対象とするあいづちについて述べ，提案するあいづち検出の方法について説明する．

3.2.1　本研究で扱うあいづち

本研究では，聞き手の関心や心的状態と深く関係し，かつ検出が比較的容易と考えられるあいづちを対象とする．具体的には，吉田ら[21]があげているうちの応答系感動詞と感情表出系感動詞の引き延ばし型のみを対象とする（例：「あー」，「はー」，「へー」，「ほー」，「ふーん」などで，「はーはー」，「ほーほー」などの引き延ばしの繰返し型を含む）．

常ら[25]はポスター会話の分析を行い，「へー」，「ふーん」，「あー」の 3 つの引き延ばし型のあいづちが，聞き手の関心・興味や驚きと関係が深いことを報告している．ただし，たとえば「あー」と「はー」のように，長母音が等しく音響的にも類似しているあいづちは，人が聞き分けるのも難しい場合が多いため，本研究では，これら 3 種以外の引き延ばし型のあいづち全般を対象とする．逆に，「はい」「うん」などの引き延ばし型でなく，応答にも用いられるパターンのあいづちは対象としない．

3.2.2　あいづち検出の方法

引き伸ばし型のあいづちは長母音を含むことから，有声休止がこれらを検出するための手がかりとして有用であると考えられる．しかし，フィラーや言い淀みなどにも有声休止は含まれるため，それらの誤検出を防ぐ必要がある．

そこで本研究では，以下の 3 つの処理を適用してあいづちを検出する．

(1)　BIC セグメンテーションで得られるセグメントのうち，持続長が t 秒よりも短く，あいづち GMM の対数尤度が閾値 θ より大きい場合に，あいづち候補とする．
(2)　有声休止検出[23] を適用し，有声休止が検出されたものを候補として残す．
(3)　音声認識結果を用いて，区間中にフィラーを含む候補を除去し，残った候補をあいづち区間として出力する．

4. ポッドキャストにおける音響イベント検出

本研究では，ポッドキャストのうち，聞き手がいて，かつ自由に話しているコラム・対話形式のものを対象として，前章で述べた方法に基づいて音響イベント検出を行う．

4.1 ポッドキャストの特徴

この種のポッドキャストは主に音声と音楽から構成され，これらが混合した部分も少なくない．音声のみの区間，背景に音楽がある音声区間（= 混合区間），音楽のみの区間では，それぞれ音響特徴量の変動すなわち分散に違いがある．音楽区間では様々な楽器や音色・音高による多様な音楽がみられるため，音声区間よりも音響特徴量の変動が大きい．一方，混合区間で使用される背景音楽は，音楽のみの場合と比べて単調なものが多く，それが音声自体の変動もオフセットするため，全体として音響特徴量の分散は小さくなると考えられる．これらの特性を考慮して，BIC セグメンテーションにおける分割重みを設定する．

4.2 音響イベント検出の処理の流れ

上記の特性に基づいて，本研究では音声，混合，音楽を大分類として設定し，前処理としてこれら 3 つのクラスに対して GMM による粗い分類を行う．得られた各大分類の区間に対して，それぞれの分割重み λ_{spe}，λ_{mix}，λ_{mus} を用いた BIC セグメンテーションを行った後で，各区間をあらためて 8 クラス（男性音声，女性音声，男性混合，女性混合，音楽，無音，笑い声，あいづち）の GMM により分類する．

大分類（音声，混合，音楽），および最終的な識別対象の上記 8 クラスについて，各々の学習データを用いて GMM を推定する．学習に用いたデータベースを**表 1** に示す．GMM の共分散行列には対角成分のみを用い，混合数は 256 とした．また，各大分類の分割重みの値を 3.1.2 項の手法を用いてそれぞれ推定する．

ポッドキャスト中の音響イベント検出の処理の流れを**図 2** に示す．以下，各処理について説明する．

本研究では，12 次元のメル周波数ケプストラム係数（MFCC），ΔMFCC，対数パワー，Δ 対数パワーからなる計 26 次元の音響特徴量ベクトルを用いる．入力音響信号のサンプリング周波数は 16 kHz で，分析用のフレーム長を 25 ms，フレーム周期を 10 ms とした．

表 1　各クラスの学習データセット
Table 1　Training data set for acoustic events.

クラス	学習データ
音声（男性・女性）	新聞記事読み上げ音声コーパス（JNAS）
音楽	RWC 音楽データベース（RWC-MDB）
混合（男性・女性）	JNAS と RWC-MDB を合成
無音	JNAS から切り出した無音区間
笑い声	IMADE ポスター会話[9]，Web 上のデータ
あいづち	IMADE ポスター会話

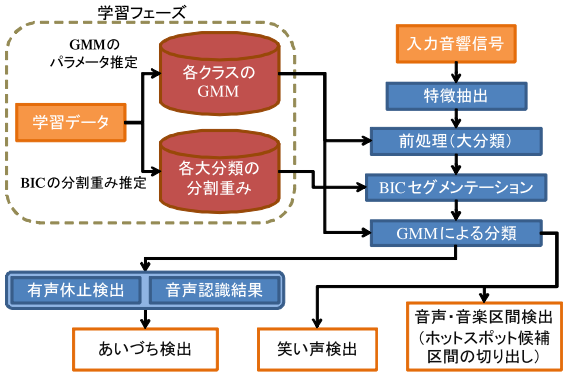

図 2　音響イベント検出の流れ
Fig. 2　Flow of acoustic event detection.

前処理として各大分類の GMM を用いて，粗い分割と分類を行う．具体的には，音声，音楽，混合の 3 状態をエルゴディックに遷移するマルコフモデルを考える．状態遷移が頻繁に行われないように，同一状態への自己遷移確率を 0.98，他の 2 状態への遷移確率を 0.01 ずつに設定した．このマルコフモデルを用いたビタビ探索により最尤系列を求めることで大分類の識別を行う．

次に，各大分類ごとに推定された分割重みを用いて，BIC セグメンテーションを行う．本研究では可変長窓を用いた分割を行う．その手順は以下のとおりである．

(1)　窓幅を最小窓幅 W_{min} に初期化し，入力の最初の点から分割境界の探索を開始する．
(2)　現在の窓幅で $\Delta BIC > 0$ となる分割境界が得られない場合，現在の窓幅に最小窓幅

を加算した新たな窓幅を用いて，分割境界が得られるまで同じ処理を繰り返す．

(3)　分割境界が得られた場合，その境界点を新たな始点として，窓幅を最小窓幅に設定し直したうえで，分割境界の探索を行う．

(4)　入力の終わりまで (2) と (3) の処理を繰り返す．

本研究では $W_{min} = 100$ フレーム（1.0 秒）とした．

最後に各セグメントを，男性音声，女性音声，男性混合，女性混合，音楽，無音，笑い声，あいづちの各 GMM の対数尤度に基づいて分類する．あいづちに関しては，3.2.2 項で述べた特別の処理を適用する．その際のパラメータの値は，実験的に $t = 1.8$，$\theta = -30.0$ と定めた．

4.3　音響イベント検出の評価実験

実際のポッドキャスト 4 番組（バラエティ番組 2，ビジネスインタビュー番組 2）から 2 エピソードずつの計 8 エピソードからなるテストセットを用いて，音響イベント検出の評価を行った．各エピソードの長さは 10 分〜40 分である．

4.3.1　実 験 条 件

最終的な識別対象 8 クラスの GMM の学習には表 1 のデータベースに加えて，ポッドキャストの音声も用いた．その際に，テストセットで用いる番組の過去のエピソード（各番組 1，計 4 エピソード）を使用しない場合（program-open；19 エピソード）と，使用する場合（program-closed；23 エピソード）を比較した．

提案手法により大分類ごとに自動推定された分割重み λ_{spe}，λ_{mix}，λ_{mus} の値はそれぞれ 1.68，1.22，3.48 となった．音声区間の値 λ_{spe} と比較して，音楽区間の値 λ_{mus} は大きくなっていることから分割されにくく，混合区間の値 λ_{mix} は小さくなっていることから分割されやすくなっている．これらの値の大小関係は，4.1 節で述べた各区間の特性を反映しており，妥当な値が得られたといえる．

このように大分類ごとに推定された分割重み λ を切り替える提案手法の有効性を評価するために，$\lambda = 1.0$，1.5，2.0 で固定した場合と比較を行った．

評価尺度として，全 8 クラスのフレームごとの分類精度を用いた．ただし，複数のイベントが重なっている区間に関しては，いずれか 1 つでも出力されている場合に正解と見なしている．また笑い声とあいづちに関する検出性能を調べるために，各々に関して，出力された区間が正解の区間に重なっていた場合を正解として，その再現率 R，適合率 P，F 値 F を求めた．F は以下のように求められる．

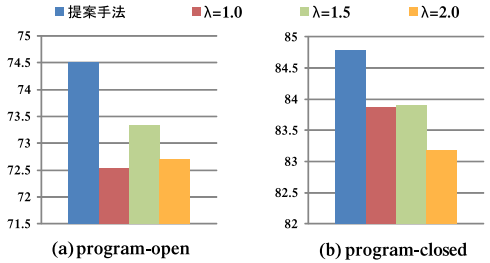

図 3　8 クラスのフレーム単位の分類精度
Fig. 3　Frame-wise classification accuracy of 8-class acoustic events.

表 2　笑い声とあいづちの検出結果（program-open）
Table 2　Detection performance of laughters and reactive tokens (program-open).

	笑い声の検出率				あいづちの検出率		
	再現率	適合率	F 値		再現率	適合率	F 値
提案手法	0.650	**0.713**	**0.687**	**提案手法**	0.340	**0.852**	**0.640**
$\lambda = 1.0$	**0.913**	0.264	0.305	$\lambda = 1.0$	**0.353**	0.679	0.547
$\lambda = 1.5$	0.742	0.422	0.459	$\lambda = 1.5$	0.331	0.793	0.599
$\lambda = 2.0$	0.600	0.575	0.575	$\lambda = 2.0$	0.292	0.812	0.575

$$F = \frac{(1+\alpha^2)RP}{R+\alpha^2 P} \tag{8}$$

ここで，α は適合率と再現率の相対的な重要度を示すパラメータである．本研究では，微かな笑い声やあいづちの検出はそれほど重視する必要はないと考えて，$\alpha = 0.5$ とした．

4.3.2　実験結果と考察

各手法による 8 クラスの平均分類精度を**図 3** に示す．過去のエピソードを使用しない場合（program-open）と使用する場合（program-closed）のいずれも，提案手法は音響条件に応じて分割重み λ を切り替えることで，固定された分割重みを用いる場合より分類精度が向上している．また，同一の番組では同じ話者や同じ音楽が出現することが多いため，過去のエピソードを学習に用いる場合の方が，10%程度分類精度が高くなっている．

笑い声とあいづちの検出性能を**表 2** に示す．過去のエピソードを使用する場合と使用しない場合でそれほど大きな差が見られなかったので，使用しなかった場合を示している．笑い声とあいづちの両方において，提案手法による精度の向上が示されている．提案手法では，適合率を重視した形になっているが，たとえば $\lambda = 2.0$ の場合と比べると，再現率・適

合率ともに高くなっていることから，提案手法の有効性は明らかである．正解には，たとえば無声に近い笑い声やあいづちといった微かなものが少なくなく，これらは検出することが困難であるうえに，検出できたとしてもホットスポット抽出においてあまり意味がないと考えられる．

5. ホットスポットの抽出

検出された音リアクションイベントに基づいて，ホットスポットを抽出する．

5.1 ホットスポットの定義

本研究では，ホットスポットは直後に音リアクションイベントを生起させる箇所として定義している．具体的に，「おもしろスポット」と「なるほどスポット」の 2 種類を考え，それぞれ以下のように定義する．

- **おもしろスポット**：笑い声の直前の（笑い声を生起させる原因となった）区間で，第三者である視聴者もおもしろいと感じうる箇所
- **なるほどスポット**：あいづちの直前の（あいづちを生起させる原因となった）区間で，第三者である視聴者も興味・関心を持ちうる箇所

5.2 ホットスポット区間の決定

ホットスポットとして提示する範囲は，長すぎると冗長となり，短すぎると内容の把握ができないおそれがある．久保田ら[22]らは，ミーティング中に会話参与者が興味深く感じた会話シーンをアノテーションするために，ボタン型の会話量子化器を提案し，実験で切り出された会話シーンの平均長は 46 秒であったと報告している．ここで扱われている会話シーンは 1 つのトピックに関する一部始終であるため，笑い声やあいづちは複数回出現する可能性がある．これに対して本研究で扱うホットスポットは，このような会話シーンがさらに細分化されたものと考えられる．

本研究では，前章までの処理との整合性を考えて，BIC セグメンテーションで分割されたセグメントの数と時間長に関してしきい値を設定した．具体的には，笑い声やあいづちの直前で，セグメント数 N_{max} 以下かつ時間長 D_{max} 秒以下を満たし，継続時間長が最大となるセグメント境界を切り出し位置とする．そのうえで，セグメント数の制約 N_{max} を 20 とし，時間長の制約 D_{max} はおもしろスポットで 20 秒，なるほどスポットで 25 秒とした．なるほどスポットの方が，複数のターンにまたがることが多く，長い文脈が必要であると考えた．

表 3　ホットスポット評価のためのアンケート項目
Table 3　Questionnaire for hot spot evaluation.

おもしろスポット

	設問	回答の選択肢
Q1	笑い声が出現する理由が分かったか？	はい/いいえ
Q2	被験者がおもしろいと感じたか？	意味不明/おもしろくない/前後なしで判断不可/おもしろみは感じる/おもしろい
Q3	当該スポットが視聴するうえで必要と思うか？	不要/ない方がよい/あった方がよい/必要

なるほどスポット

	設問	回答の選択肢
Q1	あいづちが出現する理由が分かったか？	はい/いいえ
Q2	被験者にとってどんな意味があったか？	無意味/前後なしで判断不可/納得・同意/関心・興味/新発見・驚き
Q3	当該スポットが視聴するうえで必要と思うか？	不要/ない方がよい/あった方がよい/必要

5.3 被験者実験によるホットスポットの評価

このようにして抽出された各ホットスポットを被験者に聴取してもらい，アンケート調査によりその妥当性の評価を行った．

5.3.1 実験条件と評価項目

テストセットとして，4 章の音響イベント検出の実験で用いたポッドキャストから 3 番組 2 エピソードずつの計 6 エピソードを用いた．4 名の被験者が各々2 エピソードずつ視聴（音声のみ）した．被験者は，専用 GUI 上で操作を行って，ホットスポットの（音リアクションイベントを含む）区間をエピソードごとに時系列順に聴き，それぞれのスポットについてアンケートに回答した．アンケートの設問は**表 3** に示すとおりで，選択肢から回答する形式とした．

Q1 はホットスポット抽出の成否（精度）に関する設問である．笑い声やあいづちの生起する理由が分かれば，本研究で定義したホットスポットを抽出できていると考えられる．Q2 と Q3 では，被験者自身がその箇所を聴いて主観的にどう感じるかを調査した．それぞれ Q1 の回答別に集計を行った．

5.3.2 実験結果と考察

Q1 の結果を**表 4** に示す．Q1 で「はい」と回答された割合，すなわちホットスポットが正しく抽出されていると考えられる割合（抽出精度）は，81.4%〜89.4%であった．笑い声やあいづちの音リアクションイベントが正しく検出されたホットスポットに限定すると，90%を上回っていた．このことから，ホットスポットを抽出するための N_{max} や D_{max} の設定が

表 4　ホットスポットの抽出精度
Table 4　Detection performance of hot spots.

各スポット	抽出精度（Q1 で「はい」の数/出力提示数）
おもしろスポット	81.4%（345/424）
おもしろスポット（正検出）	91.1%（338/371）
なるほどスポット	89.4%（143/160）
なるほどスポット（正検出）	90.5%（133/147）

（正検出）は笑い声/あいづちを正しく検出できていた場合

おおむね妥当であったといえる．なるほどスポットに比べて，おもしろスポットの方がやや低い値となっているのは，笑い声の検出精度（適合率）が低いためであり，正検出の場合にはほとんど差はない．

Q2 の集計結果を**図 4・図 5** に，Q3 の集計結果を**図 6・図 7** にそれぞれ示す．

なるほどスポットに関する Q2 の結果（図 5）から，正しく抽出できていた（Q1 で「はい」と答えた）場合の 9 割以上に対して，実際に被験者自身が驚きや興味・納得といった印象を受けたことが分かる．それらのなるほどスポットに対して，Q3 の結果（図 7）において，「必要」もしくは「あった方がよい」と回答されたものが多く，視聴すべき有益な情報を含んだ箇所であることが分かる．

これに対して，おもしろスポットについては，Q2 の結果（図 4），被験者が実際におもしろいと感じたのは 7 割弱であった．これは，「おもしろい」かどうかがより主観的であり，個人差によるものと考えられる．ただし，Q3 の結果（図 6）において，「必要」もしくは「あった方がよい」と回答された割合は 8 割を上回っており，たとえおもしろいと感じなくても，当該箇所が有益である場合が多いことが分かる．

6. ホットスポット提示機能を持つ視聴インタフェース

これまでに述べた，音響イベントとホットスポットの検出に基づいてポッドキャストを効率的に視聴するためのインタフェース Podspotter を作成した．その概観を**図 8** に示す．

Podspotter は，音リアクションイベント検出に基づく 2 種類のホットスポットの提示機能と各音響イベントの視覚化機能を有する．笑い声とあいづちの検出に基づいて，おもしろスポットとなるほどスポットをそれぞれ抽出して提示する（図 8 左下部）．現在の再生箇所から最も近いホットスポットにジャンプしたり，ホットスポットのみを再生したりすることも可能である．また，コンテンツの時間軸に沿って，各音響イベントをセグメントごとに色分けして提示する（図 8 上部）ことで，音響的な情報を視覚化している．たとえば，男性話者から女性話者への切り替わり点や，笑い声やあいづちの出現箇所を視覚的に知ることができ，効率的にスキップ・部分視聴ができる．

Podspotter の想定される利用シーンとして，要約的視聴と試聴が考えられる．ホットス

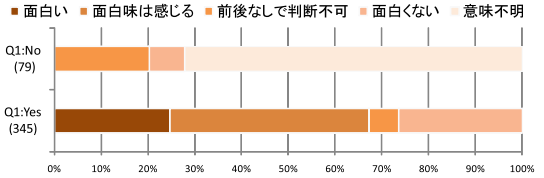

図 4　おもしろスポットに対する Q2 の集計結果
Fig. 4　Result of Q2 for funny spots.

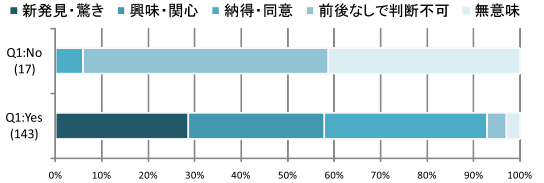

図 5　なるほどスポットに対する Q2 の集計結果
Fig. 5　Result of Q2 for interesting spots.

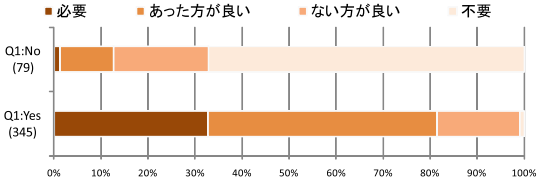

図 6　おもしろスポットに対する Q3 の集計結果
Fig. 6　Result of Q3 for funny spots.

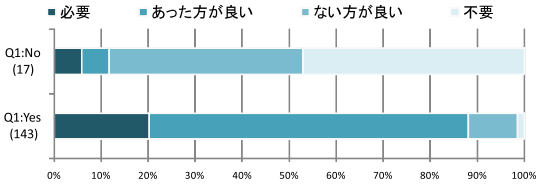

図 7　なるほどスポットに対する Q3 の集計結果
Fig. 7　Result of Q3 for interesting spots.

図 8　Podspotter の概観
Fig. 8　Outlook of Podspotter.

ポットのみを順に提示することによって，音声会話コンテンツを限られた時間で効率的に視聴できる．また，ユーザは多くの番組の中から興味を持つものを試聴したいと考えられるが，本インタフェースは，番組の特徴をすばやく把握するのに有用であると考えられる．

Podspotter は，Adobe Flex および ActionScript を用いて実装されており，Flash に対応した Web ブラウザ上で動作させることが可能であるため，計算機の OS に依存せずに利用することができる．

7. おわりに

本論文では，音声会話コンテンツの効率的なインデキシングを目的として，会話中の聞き手の反応である笑い声・あいづちとそれらに基づくホットスポットの検出手法を提案した．

笑い声やあいづちのような短時間の音響イベントを頑健に検出できるように，BIC セグメンテーションと GMM による分類を組み合わせた．また，背景音楽が頻繁に使用され，音響特性が大きく変化するポッドキャストにおいて，音響条件（音声，音楽，両者の混合）ごとにそれぞれの GMM から推定した分割重みを用いて BIC セグメンテーションを行う．さらに，有声休止検出と音声認識の情報を統合してあいづちの検出を行う．実際のポッドキャスト番組を用いた評価実験の結果，フレームごとの音響イベント分類精度や笑い声・あいづちの検出精度が従来手法に比べて有意に改善された．

本研究ではホットスポットとして，笑い声を生起させる箇所である「おもしろスポット」と，あいづちを生起させる箇所である「なるほどスポット」の 2 種類を定義した．被験者実験により評価したところ，おおむね想定されたホットスポットが正しく抽出されており，それらの多くに対して被験者が実際におもしろいと感じたり，興味・関心を持つことが示された．

また，これらのホットスポットに基づいた新たな視聴インタフェースである Podspotter を作成した．聞き手の反応に基づくこのようなインデキシングを活用することで，音声コンテンツ視聴における利便性が向上すると期待される．

謝辞　本研究の一部は，科研費特定領域研究（C）「情報爆発 IT 基盤」，および JST CREST「マルチモーダルな場の認識に基づくセミナー・会議の多層的支援環境」の一環として行われた．

参　考　文　献

1) Alberti, C., Bacchiani, M., Bezman, A., Chelba, C., Drofa, A., Liao, H., Moreno, P., Power, T., Sahuguet, A., Shugrina, M. and Siohan, O.: An Audio Indexing System for Election Video Material, *Proc. IEEE-ICASSP*, pp.4873–4876 (2009).
2) Cettolo, M., Vescovi, M. and Rizzi, R.: Evaluation of BIC-based Algorithms for Audio Segmentation, *Computer Speech and Language*, Vol.19, No.2, pp.147–170 (2005).
3) Chen, S. and Gopalakrishnan, P.: Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion, *DARPA Broadcast News Workshop*, pp.127–132 (1998).
4) Furui, S. and Kawahara, T.: Transcription and Distillation of Spontaneous Speech, *Springer Handbook on Speech Processing and Speech Communication*, Benesty, J., Sondhi, M.M. and Huang, Y. (Eds.), pp.627–651, Springer (online) (2008), available from ⟨http://www.springer.com/west/home/engineering?SGWID=4-175-22-173701714-0⟩.
5) Gatica-Perez, D., McCowan, I., Zhang, D. and Bengio, S.: Detecting Group Interest-Level in Meetings, *Proc. IEEE-ICASSP*, Vol.1, pp.489–492 (2005).
6) Gauvain, J., Lamel, L. and Adda, G.: The LIMSI Broadcast News Transcription System, *Speech Communication*, Vol.37, No.1-2, pp.89–108 (2002).

7) Gravano, A., Benus, S., Hirschberg, J., Mitchell, S. and Vovsha, I.: Classification of Discourse Functions of Affirmative Words in Spoken Dialogue, *Proc. INTERSPEECH*, pp.1613–1616 (2007).
8) Kawahara, T., Hasegawa, M., Shitaoka, K., Kitade, T. and Nanjo, H.: Automatic Indexing of Lecture Presentations using Unsupervised Learning of Presumed Discourse Markers, *IEEE Trans. Speech & Audio Process.*, Vol.12, No.4, pp.409–419 (2004).
9) Kawahara, T., Setoguchi, H., Takanashi, K., Ishizuka, K. and Araki, S.: Multi-Modal Recording, Analysis and Indexing of Poster Sessions, *Proc. INTERSPEECH*, pp.1622–1625 (2008).
10) Kennedy, L. and Ellis, D.: Pitch-based Emphasis Detection for Characterization of Meeting Recordings, *Proc. IEEE Workshop Automatic Speech Recognition and Understanding* (*ASRU03*), pp.243–248 (2003).
11) Kennedy, L. and Ellis, D.: Laughter Detection in Meetings, *NIST Meeting Recognition Workshop* (2004).
12) Knox, M.T. and Mirghafori, N.: Automatic Laughter Detection Using Neural Networks, *Proc. INTERSPEECH*, pp.2973–2976 (2007).
13) Laskowski, K.: Contrasting Emotion-bearing Laughter Types in Multiparticipant Vocal Activity Detection for Meetings, *Proc. IEEE-ICASSP*, pp.4765–4768 (2009).
14) Nishida, M. and Kawahara, T.: Speaker Model Selection based on the Bayesian Information Criterion applied to Unsupervised Speaker Indexing, *IEEE Trans. Speech & Audio Process.*, Vol.13, No.4, pp.583–592 (2005).
15) Ogata, J., Goto, M. and Eto, K.: Automatic Transcription for a Web 2.0 Service to Search Podcasts, *Proc. INTERSPEECH*, pp.2617–2620 (2007).
16) Schwarz, G.: Estimating the Dimension of a Model, *The Annals of Statistics*, Vol.6, No.2, pp.461–464 (1978).
17) Tranter, S. and Reynolds, D.: An Overview of Automatic Speaker Diarisation Systems, *IEEE Trans. Audio, Speech, & Language Processing*, Vol.14, pp.1557–1565 (2006).
18) Truong, K.P. and Leeuwen, D.: Automatic Detection of Laughter, *Proc. INTERSPEECH*, pp.485–488 (2005).
19) Ward, N.: Pragmatic Functions of Prosodic Features in Non-Lexical Utterances, *Speech Prosody*, pp.325–328 (2004).
20) Wrede, B. and Shriberg, E.: Spotting “Hot Spots” in Meetings: Human Judgments and Prosodic Cues, *Proc. EUROSPEECH*, pp.2805–2808 (2003).
21) 吉田奈央，高梨克也，伝　康晴：対話におけるあいづち表現の認定とその問題点について，言語処理学会第 15 回年次大会発表論文集，pp.430–433 (2009).
22) 久保田秀和，齊藤　憲，角　康之，西田豊明：会話量子化器を用いた会話場面の記録，情報処理学会論文誌，Vol.48, No.12, pp.3703–3714 (2007).
23) 後藤真孝，伊藤克亘，速水　悟：自然発話中の有声休止箇所のリアルタイム検出システム，電子情報通信学会論文誌，Vol.83-D-Ⅱ, No.11, pp.2330–2340 (2000).
24) 三村正人，河原達也：会議音声認識における BIC に基づく高速な話者正規化と話者適応，情報処理学会研究報告，SLP-82-6 (2010).
25) 常　志強，高梨克也，河原達也：ポスター会話におけるあいづちの韻律的特徴に関する印象評定，人工知能学会研究会資料，SLUD-A901-06 (2009).
26) 中川聖一，富樫慎吾，山口　優，藤井康寿，北岡教英：講義音声ドキュメントのコンテンツ化と視聴システム，電子情報通信学会論文誌，Vol.91-D, No.2, pp.238–249 (2008).

(平成 23 年 4 月 11 日受付)
(平成 23 年 7 月 8 日採録)

河原　達也（正会員）

1987 年京都大学工学部情報工学科卒業．1989 年同大学院修士課程修了．1990 年同博士後期課程退学．同年京都大学工学部助手．1995 年同助教授．1998 年同大学情報学研究科助教授．2003 年同大学学術情報メディアセンター教授．現在に至る．この間，1995 年から 1996 年まで米国・ベル研究所客員研究員．1998 年から 2006 年まで ATR 客員研究員．1999 年から 2004 年まで国立国語研究所非常勤研究員．2001 年から 2005 年まで科学技術振興事業団さきがけ研究 21 研究者．2006 年から情報通信研究機構短時間研究員・招へい専門員．音声言語処理，特に音声認識および対話システムに関する研究に従事．京大博士（工学）．1997 年度日本音響学会粟屋潔学術奨励賞受賞．2000 年度情報処理学会坂井記念特別賞受賞．情報処理学会連続音声認識コンソーシアム代表，IEEE SPS Speech TC 委員，IEEE ASRU 2007 General Chair，言語処理学会理事，を歴任．情報処理学会音声言語情報処理研究会主査．日本音響学会，情報処理学会各代議員．電子情報通信学会，人工知能学会，言語処理学会，IEEE 各会員．

須見　康平（正会員）

2008 年京都大学工学部情報学科卒業．2010 年同大学院情報学研究科知能情報学専攻修士課程修了．在学中，音楽情報処理・音響信号処理等の研究に従事．現在，ヤマハ株式会社勤務．

緒方　　淳（正会員）

2003 年龍谷大学大学院理工学研究科博士後期課程修了．同年産業技術総合研究所に入所し，現在に至る．博士（工学）．音声認識，音声インタフェースに関する研究に従事．2000 年日本音響学会粟屋潔学術奨励賞，2001 年電子情報通信学会学術奨励賞，2004 年 WISS2004 ベストペーパー賞，2006 年 WISS2006 ベストペーパー賞，2006 年情報処理学会山下記念研究賞各受賞．電子情報通信学会，日本音響学会各会員．

後藤　真孝（正会員）

1998 年早稲田大学大学院理工学研究科博士後期課程修了．博士（工学）．同年電子技術総合研究所に入所し，2001 年に改組された産業技術総合研究所において，現在，情報技術研究部門メディアインタラクション研究グループ長．統計数理研究所客員教授，筑波大学大学院准教授（連携大学院），IPA 未踏 IT 人材発掘・育成事業未踏ユースプロジェクトマネージャーを兼任．ドコモ・モバイル・サイエンス賞基礎科学部門優秀賞，科学技術分野の文部科学大臣表彰若手科学者賞，情報処理学会長尾真記念特別賞等，25 件受賞．