C-05

音リアクションイベント検出機能をもつポッドキャスト視聴インタフェース
Podcast Browser Based on Acoustic Event Detection

須見 康平 † 河原 達也 † 緒方 淳 ‡ 後藤 真孝 ‡
Kouhei Sumi Tatsuya Kawahara Jun Ogata Masataka Goto

1. はじめに

近年，インターネット上にはポッドキャストをはじめとした音声メディア（主に MP3 形式のオーディオファイル）や，映像が加わった動画コンテンツなどが多く存在するようになった．テキストや画像などのコンテンツは，一覧性に優れているため，部分抽出や検索などが容易である．それに対し音声・音からなるコンテンツは，一度すべてを聴かなければどこにどのような情報が存在するのかを把握することができない．つまり音声・音が一覧性に乏しい特性を持つため，音声メディアのオンデマンドな検索や閲覧が非常に困難であるという問題がある．

この問題に対して，音声認識を適用することで音声をテキスト化し，視聴だけでなく検索・閲覧を可能にするサービス（PodCastle[1]，Google Audio Indexing[2] など）が実現されている．PodCastle では，音声認識誤りを人手で修正可能なインタフェースを用いて，一般ユーザが修正した結果が音声認識の改善に反映される枠組みが構築されている [3, 4]．また Google Audio Indexing では，比較的音声認識が容易な政治演説を中心とした動画コンテンツを対象として，高精度な音声認識による検索と部分抽出を実現している [5]．しかしながら，ウェブ上に存在する音声メディアの多くは，純粋な音声だけでなく，音楽や音響効果，環境音，背景雑音などの多くの要素を含み，現状の技術によってテキスト化し，内容を完全に把握するのは困難である．また多様な形式のコンテンツが存在し，自由なスタイルの発話が多く，話し言葉特有の言い回しや多人数での同時発話などもあるため，音声認識は容易ではない．

そこで我々は，音声・音響データの一覧性を高めるための手段として，音声認識で対象となる言語情報ではなく，笑い声やあいづちなどの人のリアクションによって生じる音響的な非言語情報（音リアクションイベント）の検出を行うことで，視聴者が興味を持ちそうな箇所（＝ホットスポット）を特定することができるのではないかと考える．例えば，笑い声はおもしろいと思わせる発話が出現した直後に起こることが多い．またあいづちは聞き手の関心の度合いを表す機能をもち [6]，興味を引きそうな部分と密接に関わる音響イベントである．したがって，これらのリアクションイベントの直前にホットスポットの候補が存在する可能性が高いと考えられる．本研究では，ポッドキャストに対する音リアクションイベント検出手法 [7] を適用して得られる音響イベント系列をもとに，効率的な視聴を可能にするポッドキャスト視聴インタフェース「**Podspotter**」を提案する．

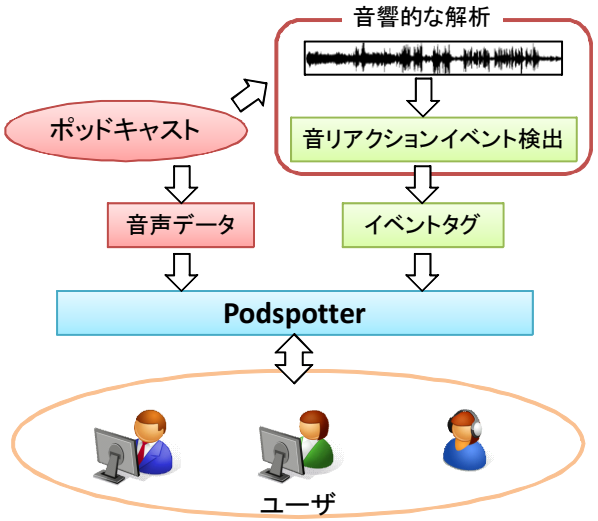

図 1: 視聴のための処理

以下，2 章において音リアクションイベントの検出と Podspotter の概要と特長について述べ，3 章では各機能実現のためのインタフェースの実装と詳細について説明する．最後に 4 章でまとめを述べる．

2. 多機能ポッドキャスト視聴インタフェース: Podspotter

本研究で提案する Podspotter は，音リアクションイベント検出に基づくポッドキャストの視覚化と，2 種類のホットスポットの提示といった機能を持ち，より自由なポッドキャストの再生が可能なインタフェースである．

Podspotter を用いてポッドキャストを視聴するためには，ポッドキャストの音声データ（MP3 形式）とタイムスタンプが付与されたイベントタグが必要である．音声データはポッドキャストの各配信先からダウンロードすることが可能であり，イベントタグはそれらのデータに対して音響的な解析を行うことにより得られる（図 1）．

† 京都大学 大学院 情報学研究科,
Graduate School of Informatics, Kyoto University
‡ 産業技術総合研究所, AIST

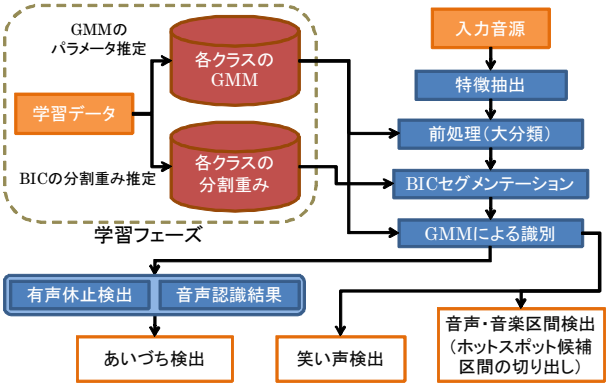

図 2: 音リアクションイベント検出の流れ

2.1 音リアクションイベントの検出

我々は，ポッドキャスト中の音リアクションイベントに関して，Bayesian Information Criterion（BIC）に基づく分割と，Gaussian Mixture Model（GMM）を用いた識別による検出手法を提案した [7]．その際，背景音楽が混在する区間としない区間の音響特性の違いを考慮して，音声区間，音楽区間，それらが混合された区間を大分類として粗く識別し，それぞれの大分類区間の分割重みの自動推定とその切り替えにより，笑い声やあいづちといった音リアクションイベントだけでなく，ホットスポットの候補区間となる音声や音楽区間についても高精度に検出可能であることを示した．

図 2 にその処理の流れを示す．学習フェーズでは，各クラスの GMM パラメータの推定と，各大分類の BIC 分割重みの推定を行う．識別フェーズでは，特徴抽出の後，大分類の区間を粗い分割により求め，各大分類ごとに分割重みを切り替えて BIC に基づく細かい分割を行う．得られたセグメントごとに GMM 対数尤度による識別を行い，あいづちはさらに有声休止検出 [8] と音声認識結果を用いて検出する．

本研究では，男性音声，女性音声，音楽，男性音声＋音楽（男性混合），女性音声＋音楽（女性混合），笑い声，あいづち，無音を 8 つの音響イベントとして扱う．これらのイベントに関して，この手法を用いてタグ付けを行ったデータをイベントタグとして用いる．イベントタグは BIC に基づく分割によって得られるセグメントごとに，8 つうちのいずれかの音響イベントタグが付与された形式になっており，この情報をもとにイベントブラウジングとホットスポットの抽出を行う．

2.2 イベントブラウジング機能

前述の音リアクションイベント検出を用いて得られる 8 つの音響イベントを，各セグメントごとに色分けして提示することで，音響的な情報を視覚化することが可能となる．例えば，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知ることができる．このことから音響イベントの視覚化によって，視聴を省力化できることが期待される．

図 4: Podspotter の概観

2.3 2 種類のホットスポッティング

本研究では「おもしろスポット」と「なるほどスポット」と名づけた 2 種類のホットスポットに着目して提示を行う．それぞれのスポットを以下のように定義する．

- **おもしろスポット**
  笑い声に基づくホットスポットで，笑いが起こる周辺に焦点をあてて切り出しを行う．これにより，なぜ笑い声が発生するに至ったかの経緯を聴くことが可能となる．そのエピソードの中で笑いが起こるような面白い箇所のみを抽出することを目指す．
- **なるほどスポット**
  「あー」,「へー」,「ふーん」といった関心を示すものを中心としたあいづちに基づくホットスポットで，対話中の聞き手の関心に焦点をあてた切り出しを行う．エピソード中の聞き手が関心をもつ箇所は，ポッドキャストの視聴者にとっても関心をもつ箇所である可能性が高く，特に役に立つ情報を提示することができると考えられる．

それぞれのホットスポット区間は，BIC に基づく分割により得られたセグメント数 N_{max} と時間長 D_{max} の制約，さらにイベントの切り替わり点が存在するか否かに応じて決定される．笑い声かあいづちのいずれかの音リアクションイベントが検出されたセグメントの直前のセグメント系列 HS をホットスポットとして抽出する．その際 HS は，セグメント数 N_{max} 以下かつ時間長 D_{max} 以下を満たし，イベントの切り替わり点を含む場合は，セグメント数と時間長をできるだけ大きくするような切り替わり点までをホットスポットして切り出すこととした．現在の実装では，$N_{max} = 10$，D_{max} は音リアクションイベントごとに異なる値（笑い声の場合は 20 秒，あいづちの場合は 25 秒）とした．これは笑い声よりもあいづちの方が，複数の話者の発話にまたがることが多く，そのイベントを発生させる要因となる発話区間が長いと考えられるためである．

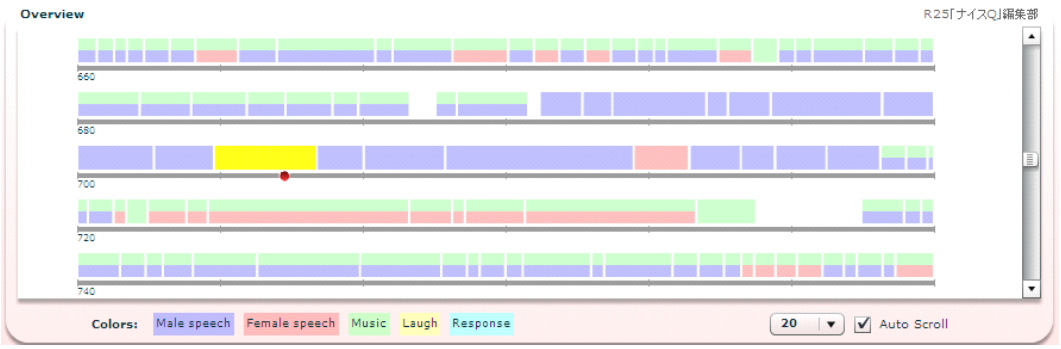

図 3: 全体表示パネル

3. Podspotter のインタフェース

Podspotter は，各機能を実現する 3 つのパネルとコントロール部から構成される（図 4）．このインタフェースを，Adobe Flex 及び ActionScript を用いて実装した．これにより，Flash 環境に対応したウェブブラウザ上で動作させることが可能であるため，OS に関わらず利用することができる．以下では各パネルについて詳細を述べる．

3.1 全体表示パネル

図 3 はポッドキャストの音響イベントを全体的に視覚化して表示するパネルであり，これによってイベントブラウジング機能を実現する．

このパネルではイベントタグに含まれる 8 つの音響イベントを男性音声（青），女性音声（赤），音楽（緑），笑い声（黄），あいづち（青緑）の色分けされたブロックによって表現する．音楽と音声の混合区間については，音楽と音声で半分ずづ色分けされたブロックで表し，無音区間ではブロックの表示を行わない．

ブロック列の下線はタイムラインを表し，左端から右端までの時間単位は右下のコンボボックスの値を切り替えることで，10 秒から 180 秒の範囲で変化させることができる．ブロックと下線はどちらもクリックを受け付ける仕組みになっており，いずれもクリックされた箇所からの再生が可能である．再生時には再生箇所のブロックのハイライトと，下線上に再生ポインタが表示される．その際右下のチェックボックスをチェックすることによって，再生箇所を自動で追従してスクロールするオートスクロール機能を有効にすることができる．

3.2 ホットスポットパネル

このパネルでは，抽出した「おもしろスポット」と「なるほどスポット」を時間順で提示する（図 5）．各ホットスポットの提示は，アイコンとセグメント列を表すブロック群から構成される．「おもしろスポット」を表すアイコンと「なるほどスポット」を表すアイコンの違いによってそれぞれ区別し，ブロックの色分けは前述の全体表示パネルと同様である．

各アイコン・ブロックはクリック可能であり，アイコンがクリックされた場合はそのスポットの先頭から再生され，ブロックがクリックされた場合はそのブロックの先頭から再生される仕組みとなっている．

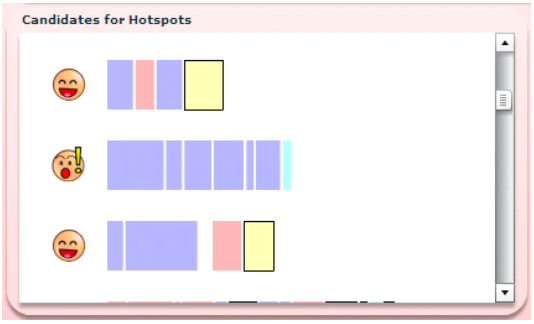

図 5: ホットスポットパネル

3.3 音声認識結果パネルとコントロール部

図 6 の上部は音声認識結果を表示するパネルであり，下部は音声データの各操作のためのコントロール部である．

再生するポッドキャストの音声認識結果がある場合は，このパネルに結果を表示することが可能である．表示されるテキストは，辞書に登録された単語ごとにブロックになっており，話し手の性別によって色分けされる．各ブロックをクリックすることで，そのブロックからの再生が可能である．

コントロール部の左側にはファイルを開くためのボタン，各パネルの表示と消去ボタン，さらにポッドキャストの再生・停止ボタンが配置され，中央ではファイル名と現在の再生時間・全再生可能時間を表示する．右側のアイコンを伴うボタンはスポットジャンプのためのボタンであり，再生中の箇所より後ろで最も近いホットスポットにジャンプすることが可能である．

4. おわりに

本稿では，音リアクションイベント検出を利用したポッドキャスト視聴インタフェース「Podspotter」を提案した．一覧性の低い音声データであるポッドキャストに対して，8 つの音響イベントを検出し，それぞれ色分けして提示することによ

図 6: 音声認識結果パネルとコントロール部

り，全体の流れをつかみやすくする効果を得ることができる．さらに笑い声とあいづちに基づく 2 種類のスポッティング機能を提供することで，より効率的な視聴環境を構築した．

今後の課題としては，ホットスポットの抽出について，一定の切り出し方法ではなく，統計的な手法での切り出し法の検討が挙げられる．また話者認識の枠組みを用いることで，性別だけでなく個別の話者の切り替わり点の表示を考えている．システムの実装に関しては，現在の仕様ではローカルの音声データファイルとイベントタグファイルを読み込む形態であるが，今後クライアントサーバモデルへの移行を考えており，ウェブサーバ上のデータをブラウザ上の Podspotter（クライアント）が取得することにより，広く利用することが可能になると考えられる．

参考文献

[1] PodCastle, http://podcastle.jp/.

[2] Google Audio Indexing, http://labs.google.com/gaudi.

[3] 後藤，緒方，江渡，“PodCastle の提案: 音声認識研究 2.0 を目指して,” 情処研報，SLP-65-7，2007.

[4] 緒方，後藤，江渡，“PodCastle の実現: Web2.0 に基づく音声認識性能の向上について,” 情処研報，SLP-65-8，2007.

[5] C. Alberti, M. Bacchiani, A. Bezman, et al., “An audio indexing system for election video material,” Proc. ICASSP, pp.4873–4876, 2009.

[6] 常，高梨，河原，“ポスター会話におけるあいづちの形態的・韻律的な特徴分析と会話モード間との相関の分析,” 人工知能研資，SIG-SLUD-A802-02，2008.

[7] 須見，河原，緒方，後藤，“ポッドキャストを対象とした音リアクションイベント検出,” 情処研報，SLP-77-24，2009.

[8] 後藤，伊藤，速水，“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,” 信学論，vol.83，no.11，pp.2330–2340，2000.