Vol.2009-SLP-77 No.24
2009/7/18

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

須見康平[†1]　河原達也[†1]
緒方淳[†2]　後藤真孝[†2]

ポッドキャスト中の重要な箇所（ホットスポット）を抽出するための手掛かりとなる音響イベントの検出手法を提案する．本研究では，視聴者が興味を持ちそうな箇所と密接に関係すると思われる，発話者や対話参加者のリアクションに基づく笑い声やあいづちなどの音響イベント（音リアクションイベント）に着目し，ホットスポットの候補区間となる先行発話の区間とともに抽出することを考える．背景音楽が頻繁に混在するポッドキャストにおいて，頑健に区分化と分類を行うために，背景音に応じて分割重みを自動推定した BIC に基づく分割と GMM による識別を組み合わせた手法を提案する．評価実験において，大分類を行って分割重みを切り替える提案手法により，分類・識別の精度が改善され，笑い声やあいづちの検出精度も向上した．

Acoustic Event Detection for Finding Hot Spots in Podcasts

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

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 the proposed method compared to when using a constant penalty weight.

1. まえがき

近年，インターネット上にはポッドキャストやウェブラジオ，ボイスブログといった音声メディア（主に MP3 形式のオーディオファイル）や，それに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．そういった音声や音を含むコンテンツは，テキストや画像ベースのコンテンツと異なり，一度すべてを聴かなければどこにどのような情報が現れているのかを把握することができない．つまり音声・音は一覧性に乏しいため，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，検索・閲覧を可能にするサービス（Podscope[1)]，PodCastle[2)]，Google Audio Indexing[3)] など）が提案されている．PodCastle では，音声認識の誤りを人手で容易に修正可能なインタフェースを用いることで，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[4)–7)]．また Google Audio Indexing では，比較的音声認識が容易な政治家の演説を中心とした動画コンテンツを対象として，高精度な音声認識を実現し，検索と部分抽出を可能にしている[8)]．しかしながら，ウェブ上の音声メディアに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が存在するため，現状の音声認識技術を適用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非言語情報（音リアクションイベント）に着目する．発話者や対話参加者が意味のあるリアクションをとった箇所を検出することで，視聴者が興味を持ちそうな箇所（ホットスポット）の候補を特定することができないか検討する．例えば，笑い声は独話や対話の中でおもしろいと思わせる発話があった時に起こり，拍手は講演などで聞き手が感心するような発話の後に起こることが多い．また，対話中に起こるあいづちは聞き手の関心の度合いを表す機能をもち[9)]，興味を引きそうな部分と密接に関わるイベントである．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（図 1）．さらに，音声認識を行う際に障害となる音リア

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

Vol.2009-SLP-77 No.24
2009/7/18

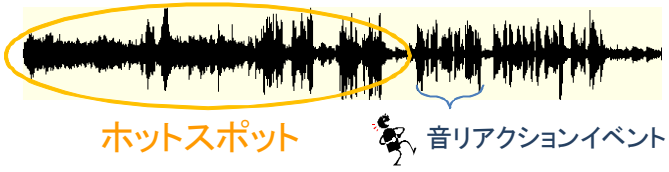

図 1 音リアクションイベントとホットスポット

クションイベントに対して対処できれば，音声認識においても有用となる．そこで本研究では，ウェブ上の音声メディア，主にポッドキャストにおいて，頻繁に使用される背景音楽や短時間の音響イベントに対して頑健に区分化と分類を行うために，音響条件に応じて適切な分割頻度を自動的に切り替える BIC に基づく分割と GMM による分類を組み合わせた手法を提案し，音リアクションイベントの検出とホットスポットの候補部分となる音声区間の検出を行う．

以下，2 章では従来の音響イベントの検出手法とポッドキャスト中のイベントを検出する際の問題点について述べる．3 章では，従来の典型的な手法である BIC に基づく分割において，分割の頻度を決定する要素である分割重みの自動推定について説明する．4 章では，提案するシステムの概要と各モジュールを具体的に説明し，5 章で評価実験結果を示した後，6 章で結論を述べる．

2. ポッドキャスト中の音響イベント検出

本研究では，ポッドキャスト中に頻繁に出現し，ホットスポットと深く関係すると考えられる「笑い声」と「あいづち」に着目する．ホットスポットの区間を抽出するためには，発話境界を検出する必要があるが，音声，音楽，背景に音楽が存在する音声（混合），さらに無音もそれぞれ音響イベントとして捉えることで，ホットスポットを一連の音響イベント群として扱うことが可能である．音声については男性と女性の分類を行う．これにより，男性音声，女性音声，音楽，男性混合，女性混合，笑い声，あいづち，無音の計 8 つの音響イベントの検出・分類を本研究の目標とする．

2.1 音響信号の区分化・分類

音響イベント検出に関する研究は，音響信号の区分化・分類に関する研究と同様に，これまで多く行われており，大きく分けて大規模な学習データから学習したモデルを用いて分割と識別を行う手法と，非学習ベースで分割を行う手法が提案されている．これらは主にニュースやミーティングなどを対象に研究が行われてきたが，ポッドキャストでは，ニュースと比較するとより自然な発話や短い発話が多い．また，ミーティングでは背景音楽は存在しないが，ポッドキャストでは頻繁に使用される．したがって，ポッドキャスト中の音響イベントを検出するためには，背景音に対して頑健に，かつ短時間の音響イベントを検出しなければならない．以下各々の手法について，簡単に述べる．

2.1.1 学習ベースの手法

あらかじめ認識対象となる各音響イベントの学習用データを収集し，混合正規分布（GMM）や隠れマルコフモデル（HMM）[10]，またサポートベクターマシーン（SVM）[11] などのモデルを用いる手法が一般的である．また笑い声に関しては，ニューラルネットワークを用いて検出する手法[12] や，有声の笑い声と無声の笑い声を明確に区別してモデル化を行う手法[13] も提案されている．

ポッドキャストではより自然なスタイルで会話が行われるため，話者交代の頻度が高く，短時間のイベントも多く出現する．短時間のフレームから得られる音響特徴量は，局所的な変動に影響されることが多いので，学習したモデルを用いて分割・識別を同時に行うのは容易ではない．

2.1.2 非学習ベースの分割手法

異なるセグメントモデル間の距離を評価することで分割を行う．学習データは必要としないが，各セグメントを分類･識別することはできない．分割手法の中で最も広く利用されているのは，Bayesian Information Criterion（BIC）[14] に基づく手法である．BIC はモデル選択の基準であり，各モデル $M = M_1, M_2, \cdots, M_m$ に対して，データセット $D = D_1, D_2, \cdots, D_N$ が与えられた場合，モデル M_i の BIC 値は以下のように定義される．

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

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．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 に基づく音響信号の分割[15),16)] では，ある 1 つの区間（N サンプル）に対して，それを 1 つのモデル $M_0 = N(\mu_0, \Sigma_0)$ で表した場合の BIC 値 $BIC(M_0)$ と，ある点 j $(1 < j < N)$ を境に 2 つのモデル $M_{12} = \{N(\mu_1, \Sigma_1), N(\mu_2, \Sigma_2)\}$ 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が用いられるのが一般的である．このとき，$\Delta BIC(j) = BIC(M_0) - BIC(M_{12})$ は以下のようになる．

Vol.2009-SLP-77 No.24
2009/7/18

$$\Delta BIC(j) = \frac{1}{2}(N\log|\Sigma| - j\log|\Sigma_1| - (N-j)\log|\Sigma_2|) - \frac{1}{2}\lambda(d + \frac{1}{2}d(d+1))\log N \quad (2)$$

ここで d は特徴ベクトルの次元数である．また λ を分割重みと呼ぶ．このとき，$j = \arg\max_j \Delta BIC(j) > 0$ であれば，点 j を分割境界とする．ただしこの手法では，分割重み λ というパラメータが一般的に用いられ，この値はタスクごとに調整する必要があるという問題がある[16]．

本研究では，短時間のイベントを検出することを考慮して，音響信号に対してまず BIC に基づく分割を行った後に，GMM による分類・識別を行う手法を用いる．しかしながら，ポッドキャストでは背景音楽が頻繁に使用され，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，BIC の分割重み λ の適切な値も変化する．そこで本稿では，音響的な条件によって分割重み λ の値を切り替える手法を提案する．

3. BIC における分割重みの自動推定

本節では，音響条件による特徴量の違いについて述べ，それぞれの条件に適した分割重みの推定手法を提案する．

3.1 音声・音楽・混合区間の特性の違い

音声のみの区間，背景に音楽がある音声区間（混合区間），音楽区間のそれぞれについて音響特徴量の分散に違いがある．音響特徴量の変動は，音声区間に比べて音楽区間で大きく，混合区間では小さくなる．音楽区間では様々な楽器や音色，音高などを含むバリエーション豊かな音楽が出現する．したがって，同じ値の分割重みを用いた場合，音声区間と比較すると分割されやすくなる．一方，混合区間で使用される背景音楽は，音楽のみのものと比べて一定なものが多く，全体として特徴量の分散は小さくなると考えられる．混合区間で音声の切り替わり点を検出するためには，より分割がされやすいように分割重みを設定しなければならない．

この特性に基づいて，音声，音楽，混合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・識別を行う．その識別された大分類に対して，それぞれの分割重み $\lambda_{spe}, \lambda_{mix}, \lambda_{mus}$ を適用した BIC ベースの分割を行う．

3.2 GMM 学習の情報を用いた分割重み推定

各大分類の適切な分割重みは GMM 学習フェーズであわせて推定する．各クラスの GMM のパラメータが同じ混合数で適切に求められている時，その GMM に含まれる各ガウス分布は適切な分布を表現（例えば音声 GMM ならば，各ガウス分布が学習サンプル内の各話者ごとの特徴空間の分布を表現）し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一つのセグメントと捉えることができ，それより混合数が大きく，小さなサイズのクラスタに対しては，結合した方がよいと考えられる．BIC の分割重みをこれらのガウス分布を用いて決定する．図 2 のように，最終的に得られる GMM の中のガウス分布と，それをさらに分割した二つのガウス分布に対する ΔBIC は，次のように定式化できる．

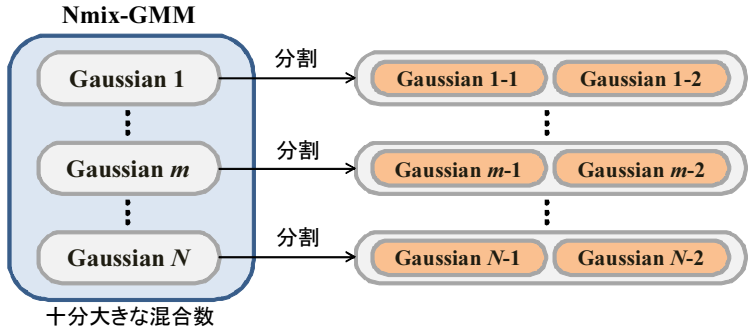

図 2 各分布の再分割

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

ここで，$m = 1, \cdots, N$ はガウス分布のインデックスであり，$n_{G_{m1}}$ 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ガウス分布 $m-1$ とガウス分布 $m-2$ に寄与するサンプル数である．この式（3）を用いて，$m = 1, \cdots, N$ のすべてのガウス分布に対して，ΔBIC を計算し，これが 0 と等しいとして得られる λ をそれぞれについて求める．さらに，それらの平均を各大分類で用いる分割重みの値とする．

3.3 大分類における分割重み λ の推定結果

5 章で述べる学習データセットを用いて，実際に学習時に推定された各大分類の分割重み $\lambda_{spe}, \lambda_{mus}, \lambda_{mix}$ はそれぞれ 1.68，3.48，1.22 となった．音声区間の値と比較して，音楽区

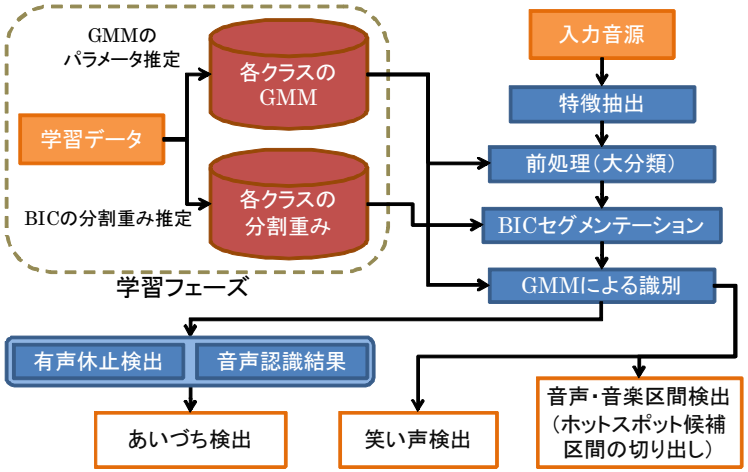

図 3 処理の流れ

間の値は大きくなっていることからより分割されにくく設定され，混合区間の値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．この値の大小関係は，3.1 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．

4. 提案する音響イベント検出システム

我々が提案する音響イベント検出システムの処理の流れを図 3 に示す．以下の節で，各処理の詳細を述べる．

4.1 学習フェーズと特徴抽出

学習フェーズでは，それぞれの GMM のパラメータを各クラスに属する特徴ベクトルを用いて推定する．特徴ベクトルをフレーム単位で求めた後，EM アルゴリズムを用いて各ガウス分布の平均と共分散，重みを推定する．ただし本研究において，共分散は対角成分のみを用い，混合数は 256 とする．また前述のように，各大分類に対しては，BIC の分割重みも同時に推定する．

音響特徴量として 12 次元 MFCC，12 次元 ΔMFCC，対数パワー，Δ 対数パワーからなる計 26 次元の特徴ベクトルを用いる．入力音響信号はサンプリング周波数 16kHz で，MFCC はフレーム長 25ms，フレーム周期 10ms として計算する．

表 1 各クラスの学習データセット

クラス	学習データ
音声（男性・女性）	JNAS[17]
音楽	RWC-MDB[18]
混合（男性・女性）	JNAS+RWC-MDB
無音	JNAS, 合成したノイズ
笑い声	IMADE ポスター会話[19], Web から収集

4.2 前処理と BIC に基づく分割

前処理として，各大分類に対する GMM（音声 GMM，音楽 GMM，混合 GMM）を用いて，粗い分割と分類を行う．BIC に基づく分割では，前処理によって得られた各セグメントごとに，適切な分割重みを選択し，さらに細かい分割を行う．精度の高い分割を実現するために，可変長窓を用いた分割手法を用いる．その手順は以下の通りである．

(1) 窓幅を最小窓幅 W_{min}（100 フレーム）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) 現在の窓幅で分割境界が得られない場合，その窓幅に最小窓幅を足したものを新たな窓幅とし，分割境界が得られるまで処理を続ける．
(3) 分割境界が得られた場合，その点を新たな始点として，最小窓幅の窓を用いて境界の探索を行う．
(4) 入力の終わりまで（2）と（3）の処理を繰り返す．

4.3 笑い声と音声・音楽区間の検出

BIC に基づく分割によって得られる各セグメントを，笑い声，男性音声，女性音声，男性混合，女性混合，音楽，無音の各 GMM の対数尤度に基づいて分類・識別する．このとき，t_{res} 秒よりも短く，あいづち GMM の対数尤度が閾値 θ_{res} よりも大きい音声区間については，あいづち候補区間として抽出する．

4.4 あいづちの検出

常らは「ふーん」,「へー」,「あー」の 3 つのあいづちが対話中の聞き手の興味と密接に関係することを報告している[9]．これら 3 つのあいづちが長母音を含むことから，有声休止がこれらのあいづち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．しかしながらフィラーや言い淀みなどにも有声休止は多く含まれるため，それらに対する誤検出を防ぐ必要がある．

そこで上記のあいづち候補区間に対して，まず有声休止検出アルゴリズム[20] を用いて，区間中に有声休止箇所を含む候補を絞り込み，さらに音声認識を行いて，フィラーとして認識

Vol.2009-SLP-77 No.24
2009/7/18

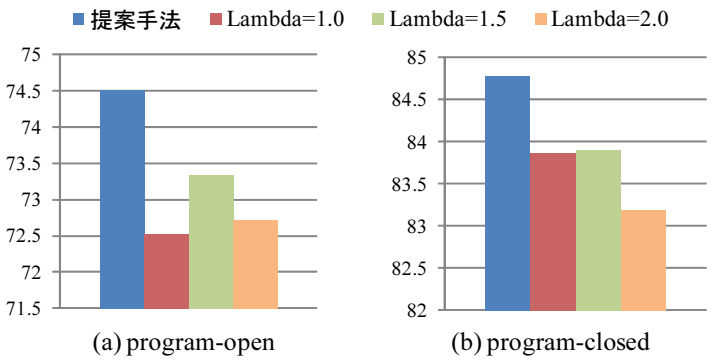

図 4 8 クラスのフレーム単位の分類精度

されたものを含む候補を取り除く．最終的に残された候補をあいづち区間として出力する．

5. 評価実験

提案手法を，実際のポッドキャスト 4 番組から 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評価を行った．GMM の学習には，表 1 のデータセットを用いた．これに加えて，実際のポッドキャストについても，テストセットで用いるエピソードの過去分を使用しない場合の 19 エピソード（program-open）と，過去分を使用する場合の 23 エピソード（program-closed）をそれぞれ区別して用いて，各クラスの GMM のパラメータ推定を行った．

提案手法の有効性を評価するために，分割重みの設定について以下の比較を行った．

(1) 提案手法（分割重み λ を切り替える手法）
(2) $\lambda = 1.0$ で固定した場合
(3) $\lambda = 1.5$ で固定した場合
(4) $\lambda = 2.0$ で固定した場合

提案手法の分割重み λ は，3.3 節で記述した結果を各大分類に用いた．

評価尺度はフレーム毎の全クラス（男性音声，女性音声，音楽，男性混合，女性混合，笑い声，あいづち，無音の計 8 クラス）の分類精度を用いた．ただし，オーバーラップを考慮せず，重なっている区間に対しては，どれかひとつでも正解が出力されている場合，そのフレームでは正解が出力されたとする．また笑い声とあいづちに関する検出精度を調べるために，それぞれに対して，出力された区間が正解の区間に一部でも重なった場合を正解として，その再現率 R，適合率 P，F 値 F を求めて評価を行う．F は以下のように求められる．

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

α は適合率の再現率に対する相対的な重要度を示すパラメータである．音リアクションイベントは実際のポッドキャスト中に多く含まれているが，そのすべてがホットスポットと結びつくわけではない．検出することが困難な微かな笑い声よりも，よりはっきりした大きな笑い声の検出を重視すべきという考えから，本研究では $\alpha = 0.5$ とし，適合率を重視した．

8 クラスの分類精度の結果を図 4 に示す．音響条件に応じて分割重み λ を切り替える提案手法によって，フレーム単位の認識率は過去のエピソードを使用しない場合とする場合のいずれも，一定値の分割重みを用いるよりも向上している．また，一つの番組内では同じ話者や同じ音楽が登場することが多いため，過去のエピソードを学習に用いることにより，精度が大幅に向上している．

笑い声とあいづちの検出においては，過去のエピソードを学習に用いた場合と用いなかった場合でそれほど差が見られなかったため，過去のエピソードを用いなかった場合の結果について表 2，表 3 に示している．表 2 において，笑い声検出において提案手法による精度の向上が示されている．微かな笑い声を検出することは困難なため，再現率は低いが，はっきりした大きな笑い声に関しては大部分を検出できていた．前述のように微かな笑い声よ

表 2 笑い声の検出精度

Measure	R	P	F
提案手法	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

表 3 あいづちの検出精度

Measure	R	P	F
提案手法	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

りも大きな笑い声の方が，ホットスポットとより密接に関係するため，再現率の低さはホットスポットを検出する上で，それほど問題にならないと考えられる．

またあいづち検出に関しても，表 3 に示すように，提案手法を用いた場合に最も高い精度が得られた．有声休止検出によるあいづち区間の再現率はおよそ 70%であったが，その中に多く含まれるフィラーや言い淀みの区間を除去し，できるだけ適合率を上げるために，やや強めの制約をかけなければならなかった．そのため，すべての再現率が低い値となっている．しかしながら，有声休止検出時の閾値や GMM 尤度に関するペナルティを調整することで，再現率と適合率のバランスをある程度調整することが可能である．

6. むすび

本稿では，音響条件の異なった大分類（音声，音楽，両者の混合区間）に対して，あらかじめ自動推定した分割重みを切り替える BIC に基づく分割を用いた，ポッドキャスト中の音リアクションイベント検出手法を提案した．笑い声とホットスポット区間の候補となるイベントについては，GMM を用いて識別を行い，さらにあいづちに関しては有声休止検出と音声認識を導入した．実際のポッドキャストを用いた評価実験の結果，提案手法を用いた 8 クラスのフレーム単位の認識率は 74.5%で，笑い声とあいづち検出においても検出精度は向上しており，分割重みを切り替える手法の有効性を示した．

今後の課題としては，話者識別の枠組みを取り入れることによる各話者ごとの区分化や，オーバーラップへの対処が挙げられる．またホットスポットの抽出法を検討するとともに，それを提示するためのインタフェースを設計・実装する予定である．

参考文献

1) Podscope: http://www.podscope.com/.
2) PodCastle: http://podcastle.jp/.
3) Google Audio Indexing: http://labs.google.com/gaudi.
4) 後藤真孝，緒方　淳，江渡浩一郎：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).
5) Goto, M., Ogata, J. and Eto, K.: PodCastle: A Web 2.0 Approach to Speech Recognition Research, *Proc. Interspeech*, pp.2397–2400 (2007).
6) 緒方　淳，後藤真孝，江渡浩一郎：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).
7) Ogata, J., Goto, M. and Eto, K.: Automatic Transcription for a Web 2.0 Service to Search Podcasts, *Proc. Interspeech*, pp.2617–2620 (2007).
8) Alberti, C., Bacchiani, M., Bezman, A. et al.: An Audio Indexing System for Election Video Material, *Proc. ICASSP*, pp.4873–4876 (2009).
9) 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).
10) Zhou, X., Zhuang, X., Liu, M. et al.: HMM-Based Acoustic Event Detection with AdaBoost Feature Selection, *Multimodal Technologies for Perception of Humans*, pp.345–353 (2008).
11) Temko, A. and Nadeu, C.: Classification of acoustic events using SVM-based clustering schemes, *Pattern Recogn.*, Vol.39, No.4, pp.682–694 (2006).
12) Knox, M. 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. ICASSP*, pp.4765–4768 (2009).
14) Schwarz, G.: Estimating the Dimension of a Model, *The Annals of Statistics*, Vol.6, No.2, pp.461–464 (1978).
15) Chen, S. and Gopalakrishnan, P.: Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion, *Proc. of DARPA Broadcast News Transcription and Understanding Workshop*, pp.127–132 (1998).
16) Tritschler, A. and Gopinath, R.: Improved speaker segmentation and segments clustering using the Bayesian Information Criterion, *Proc. Eurospeech*, pp.679–682 (1999).
17) 伊藤克亘，山本幹雄，武田一哉ほか：大語彙連続音勢認識研究用日本語コーパス：JNAS，*Journal of the Acoustical Society of Japan (E)*, Vol.20, No.3, pp.199–206 (1999).
18) Goto, M., Hashiguchi, H., Nishimura, T. et al.: RWC Music Database : Popular, Classical, and Jazz Music Databases, *Proc. ISMIR*, pp.287–288 (2002).
19) Kawahara, T., Setoguchi, H., Takanashi, K. et al.: Multi-modal recording, analysis and indexing of poster sessions, *Proc. Interspeech*, pp.1622–1625 (2008).
20) 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 (音声情報処理：現状と将来技術論文特集)，電子情報通信学会論文誌，Vol.83, No.11, pp.2330–2340 (2000).