ポスター会話中の音リアクションイベントに基づくホットスポットの抽出

須見 康平 †　　河原 達也 †
† 京都大学大学院 情報学研究科

1. はじめに

近年，講演やミーティング等をデジタルアーカイブとして蓄積し，再視聴することが可能となっている [1]．このような音声コンテンツは，一度全てを聴かなければ詳細な内容や情報の所在を把握できないため，有用箇所の抽出や検索は非常に困難である．特に多人数の会話に対しては，頻繁に生じる話者交替や話し言葉特有の言い回しのために，音声認識技術の適用も容易ではない．

そこで我々は，会話中に起こる人のリアクションによって生じる非言語の音響イベントに着目する．動画共有サイトなど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もあるが，本来会話に参与している参加者自身の反応をアノテーションできると効果的であると考えられる．

本研究では，アーカイブされたポスター会話 [2] を対象として，笑い声とあいづちを音リアクションイベントとして扱う．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り，あいづちは発話に対する聞き手の関心の度合いや，認知状態（納得や同意，発見など）を表すと考えられる [3]．これらの音リアクションイベントは，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，リアクションを生起させる発話は会話中で有益な情報を含んでいると考えられる．この発話を本稿では「ホットスポット」と呼ぶ（図 1）．

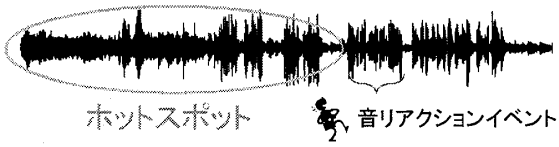

図 1: 音リアクションイベントとホットスポット

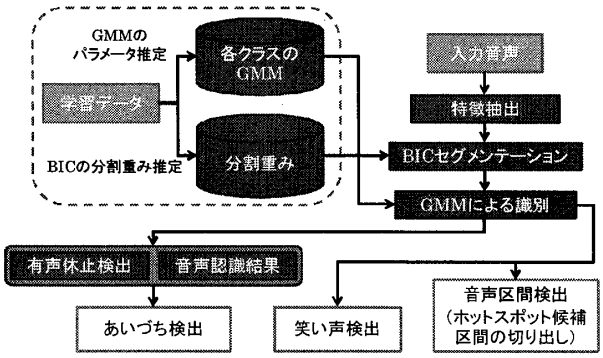

図 2: 音響イベント検出の概略

2. 音リアクションイベントの検出

音リアクションイベント検出の概略 [4] を図 2 に示す．Bayesian Information Criterion（BIC）に基づく音響セグメンテーションと Gaussian Mixture Model（GMM）に基づく識別を組み合わせることで，音リアクションイベント及びホットスポットの候補となる発話区間を検出する．本研究では，男性音声・女性音声・無音・笑い声・あいづちの計 5 つの音響イベントの検出・分類を行う．

2.1 学習フェーズ

26 次元の特徴ベクトル（12 次元 MFCC，ΔMFCC，対数パワー，Δ 対数パワー）を用いて，各クラスの GMM のパラメータと BIC の分割重み推定を行う．BIC の分割重み推定は，全学習サンプルを用いて得られる GMM 中のあるガウス分布の BIC 値と，それをさらに分割した二つのガウス分布に対する BIC 値との差分（Δ BIC）を用いて行う．[4] では音声・音楽・混合の 3 種の大分類について推定を行っていたが，ポスター会話データでは背景音楽は存在しないため，音声についてのみ推定した．

2.2 笑い声と音声区間の検出

まず推定された分割重みを用いた BIC に基づく音響セグメンテーションを行う．得られた各セグメントを，男性音声，女性音声，無音，笑い声及びあいづちの各 GMM 対数尤度に基づいて分類・識別する．あいづちについては，セグメント持続長が t 秒よりも短く，あいづち GMM 対数尤度が閾値 θ より大きい場合にあいづち候補とし，さらに以下の処理を適用する．

Detecting Hot Spots based on Acoustic Events in Poster Conversations: *Kouhei Sumi and Tatsuya Kawahara (Kyoto Univ.)*

2.3 あいづちの検出

本研究で扱うあいづちは，吉田ら [5] が挙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とする（例：あー，はー，へー，えー，おー，ふーん，など）．理由としては，引き延ばし型は韻律のバリエーションが多彩であり，興味や関心の度合いと深く関係していると考えられるためである．また音響的な特徴が有声休止 [7] と類似しているため，その特徴を利用できる．常ら [3] はポスター会話の分析を行い，「あー」，「へー」，「ふーん」の 3 つの引き延ばし型のあいづちについて，特に聴衆の関心との関係が大きいことを報告しているが，本研究では他の引き延ばし型も対象として扱う．

上記のあいづち候補区間に対して，まず有声休止の検出を行う．ただし，有声休止はフィラーや言い淀みなどにも含まれるため，音声認識を用いてそれらを取り除き，残ったものをあいづちとする．

3. 音響イベント検出精度の評価

IMADE ルーム [6] において収録されたポスター会話を対象として評価を行った．テストデータとして 2007 年収録の 4 セッションと 2009 年収録の 4 セッションを使用した．GMM の学習データには，新聞記事読み上げ音声コーパス（JNAS）とポッドキャストデータを用いた．

評価尺度はフレーム毎の全 5 クラスの分類精度（フレーム精度）を用いた．さらに笑い声とあいづちに関する検出精度を個別に調べるため，再現率 R，適合率 P，F 値 F を求めて評価した．ここで F は以下のように求められる．

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

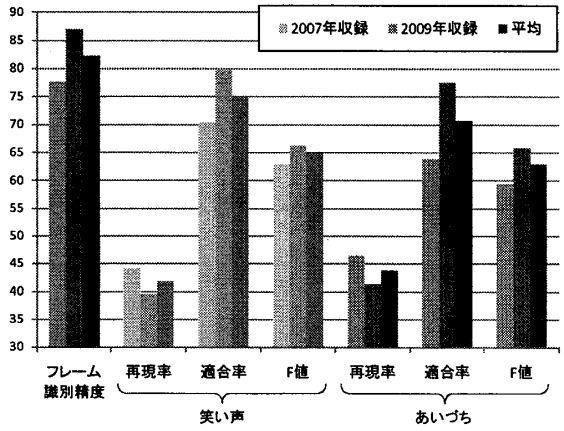

図3: 音響イベントの検出精度

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．検出することが困難な微かな笑い声やあいづちよりも，より明白なリアクションを重視すべきと考えて，本研究では$\alpha = 0.5$とし，適合率を重視した．

検出精度の結果を図3に示す．2009年のデータよりも2007年のデータに対する精度が全体的に低下しているのは，収録時に混入したノイズがはるかに大きかったためである．しかしながら，平均の検出精度はポッドキャストに対する実験結果[4]とほぼ同等となった．ポスター会話では，微かな笑い声やあいづちが比較的多く含まれるため再現率が低くなっているが，明白なリアクションに関しては大半を検出できていた．

4. ホットスポットの抽出

4.1 ホットスポットの定義

本研究で対象とするホットスポットとして，「おもしろスポット」と「なるほどスポット」の2種類を考える．

- おもしろスポット
  笑い声の直前の（生起させる原因となった）区間．
- なるほどスポット
  あいづちの直前の（生起させる原因となった）区間．

各ホットスポット区間を，セグメント数と時間長の制約を適用したセグメント境界から決定する．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，ホットスポットの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．現在の実装では，$N_{max} = 20$，$D_{max} = 25$としている．これらの値を大きくすれば，当該箇所が含まれる可能性は高くなるが，視聴に必要な労力も増加する．

4.2 被験者実験による評価

上記のように抽出したホットスポットを被験者に聴いてもらい，それぞれのスポットについてリアクション（笑い声とあいづち）が生起する理由がわかるかどうか調査した．生起する理由がわかるという場合に，ホットスポットを抽出できていると考えた．

本稿では2009年収録の4セッションを4名の被験者に聴取してもらい（ただし各人につき2セッションずつ），全出力スポットとそのうち音リアクションイベント検出が正解だったスポットについて検出率を求めた．視聴の際には，それぞれのスポットを個別に評価するのではなく，時間順に評価することで得られた情報を以後の判断に利用できるようにした．

表1: おもしろスポットの検出率

スポット（数）	検出率（数）
全出力スポット（91）	74.7%（68）
笑い声正解スポット（74）	89.2%（66）

表2: なるほどスポットの検出率

スポット（該当数）	検出率（該当数）
全出力スポット（148）	86.5%（128）
あいづち正解スポット（125）	95.2%（119）

結果を表1，2に示す．それぞれホットスポットを高い精度で抽出できている．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}の設定が妥当であったといえる．

4.3 議論

被験者に主観的な印象を聞いてみたところ，9割のなるほどスポットについて，実際に被験者自身が興味や関心，納得といった印象を受けたという答えが得られた．おもしろスポットについては，被験者が実際におもしろいと感じたのは6割程度であった．これはポスター会話中で笑い声がおもしろい箇所だけではなく，「意外」や「照れ隠し」などの感情を表現する場合にも多く出現するためであると考えられる．

また上記の評価では，音リアクションイベント検出に基づくスポットについて調べたが，全体の有用箇所中のどれくらいを抜き出せているのかを調査する必要がある．しかし，視聴者にとっての有用（おもしろい，興味のあるといった）箇所は個人の主観と大きく関係し，人によって捉え方は様々であるため，「正解」をアノテーションして評価を行うのは非常に困難である．

5. おわりに

本稿では，ポスター会話中の音リアクションイベントを検出することで，アーカイブ利用者にとって有用な箇所となり得るホットスポットを抽出する枠組みについて述べた．高精度な音リアクションイベント検出に基づいて，リアクションを生起させる発話区間を高い精度で抽出できることを示した．

参考文献

[1] J. Carletta *et al.*: The AMI Meeting Corpus: A Pre-announcement, in *Proc. MLMI*, 2005.
[2] 瀬戸口 他: 多数のセンサを用いたポスター会話の収録とその分析, 情処研報, 2007-SLP-67, pp. 31-36, 2007.
[3] 常 他: 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.
[4] 須見 他: 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, 2009-SLP-77-24, 2009.
[5] 吉田 他: 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.
[6] 角 他: IMADE: 会話の構造理解とコンテンツ化のための実世界インタラクション研究基盤, 情報処理, Vol.49, No.8, pp.945-949, 2008.
[7] 後藤 他: 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.