Vol.2010-SLP-82 No.8
2010/7/23

音声会話コンテンツにおける聴衆の反応に基づいたホットスポットの抽出

須見康平[†1]　河原達也[†1]

本研究では，ポッドキャストとポスター会話といった音声会話コンテンツを対象として，会話音声中の聞き手のリアクションに基づいて，視聴者にとって有益な箇所を抽出する手法を提案する．笑い声やあいづちを生起させる箇所（＝ホットスポット）は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られる．そこで本研究では，笑い声とあいづちの検出を行い，検出されたそれぞれのリアクションに基づいて,「おもしろスポット」と「なるほどスポット」の 2 種類のホットスポットを定義し，それらの抽出を行う．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，これらの大半が実際に被験者が興味・関心をもった箇所であることを確認した．

Detection of Hot Spots based on Audience's Reaction in Conversational Speech Content

Kouhei Sumi[†1] and Tatsuya Kawahara[†1]

This paper presents a method to detect hot spots based on audible reactions in conversational speech content, such as podcasts and poster conversations. Hot spots, which are defined as the segments involving laughters or reactive tokens in conversations, would provide useful information to listeners of the audio content. Thus, we extract two kinds of hot spots based on detected laughters and reactive tokens. Subjective evaluations demonstrated that subjects acutually had interests in most of extracted hot spots.

†1 京都大学 情報学研究科
School of Informatics, Kyoto University

1. はじめに

近年，計算機やネットワークの普及により，多様な音声コンテンツを容易に聴取・鑑賞できるようになってきている．例えば，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コンテンツが多く存在する．また講義や講演などがデジタルアーカイブとして蓄積され公開されている．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，テキストや画像のコンテンツと異なり，音声では詳細な内容や情報の所在を視聴者が速やかに把握できない．すなわち音・音声は不可視であり，時間的な長さがあるため一覧性に乏しい．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と期待される．これに関して，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・要約・重要文抽出などがこれまでに研究されている．しかし，雑音や背景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ような情報の構造化も必ずしもなされていない．

これに対して本研究では，人が会話中に自然に起こす反応によって生じる非言語の音響イベント（＝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ことを検討する[1)]．会話音声は，表層的な言語情報だけでなく，発話者や聞き手の感情といったテキストでは表現できない非言語情報も含んでいる．特に少人数の会話の場においては，聞き手（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と考えられるため，このようなイベント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．例えば，ウェブ上の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，コンテンツに登場している参加者の反応をあらかじ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，会話音声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，認知状態（納得や同意，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，会話中で有益な情報を含んでいると考えられる．本研究ではこのような箇所を「ホットスポット」と定義し（図 1），笑い声やあいづちを含む音響イベント検出[2)]に基づいて，ホットスポットを抽出する手法を提案する．この手法をポッドキャスト及びポスター会話に適用し，それらから抽出された

Vol.2010-SLP-82 No.8
2010/7/23

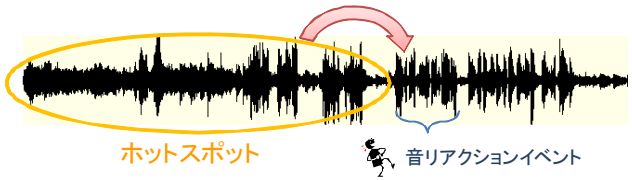


図 1 音リアクションイベントとホットスポット

ホットスポットについて，被験者実験による評価を行い，結果を報告する．

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，非言語情報のみを用いたインデキシングに関する研究が行われてきた．以下ではそれぞれの研究について説明した後，本研究の位置づけについて述べる．

2.1 音声認識に基づく検索・閲覧と要約

音声認識に基づく検索・閲覧を可能にするシステムとして，PodCastle[⋆1]や Google Audio Indexing[⋆2] などが実現されている．PodCastle では，テキストでの検索や閲覧だけでなく，音声認識の誤りを人手で容易に修正できるインタフェースを提供することで，一般視聴者が気軽に修正でき，さらに音声認識の改善にも反映される枠組みが構築されている[3),4)]．

一方，講義や講演音声を対象として，音声認識に基づいて重要文の抽出や要約を行う研究が報告されている[5),6)]．重要度の計算には，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・手がかり語の情報，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[7)]．

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，言語情報を基本的に利用せずにインデキシングを行なう研究も行なわれている．Wrede ら[8)]は，ミーティング音声中で二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” と定義して，人が判定した “Hot Spot” と韻律情報との関係を分析している．その結果，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付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，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[9)]は，ミーティング音声中で強調された発話を韻律的な特徴量を用いて特定する手法を提案しており，Gatica-Perez ら[10)]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，音楽や音響効果，環境音，背景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，高精度な音声認識を前提としている重要文抽出や要約についても，多く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，講義や講演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会話音声中には，発話者（話し手）に対してそれを聴取する聞き手が存在する．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・関心をも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．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，聞き手の心的状態・反応を表す代表的な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もつ発話であるかを表すラベルを付与できる．例えば，笑い声は「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．

⋆1 http://podcastle.jp/
⋆2 http://labs.google.com/gaudi

3. 対象とする音声会話コンテンツ

我々は以前，ポッドキャストを対象とした笑い声とあいづちを含む音響イベントの検出手法を提案した[2]．本研究では，以下の 2 種類のデータに対して，この音響イベント検出手法を適用し音響イベント系列を求める．

- **ポッドキャスト**

ポッドキャストに含まれている音声の種類は番組によって様々であり，ラジオなどで放送されている番組のダイジェスト版がポッドキャストとして配信されることが多く，また一般のユーザでも簡単に配信できるため，個人的な音声ブログとしても利用される．そのため，特にコラム・自由対話形式の番組が多くみられる．本研究では，笑い声やあいづちを多く含むコラム・自由対話形式のポッドキャストを主な対象とする．

- **ポスター会話**

我々は，京都大学に設置された IMADE ルーム[11]において，多人数インタラクションのマルチモーダルな分析を目的として，学術的なポスター発表の収録を行っている[12]．発表者 A がポスターを使って研究テーマを 2 人の聴衆 B，C に説明し議論するという形式である．参与者の役割が固定された上で，発表者が一方的に話すだけではなく，聴衆も積極的に反応を示したり質問するといった特徴がある．インタラクションが活発に行われるため，笑い声やあいづちが頻繁に観測される．

4. ホットスポットの抽出

本研究では笑い声とあいづちに基づく 2 つのホットスポットを扱う．本節では，それぞれの定義と区間の決定方法について述べる．

4.1 ホットスポットの定義

本研究では,「ホットスポット」を直後に音リアクションイベントを生起させる箇所と定義する．具体的には，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，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．

4.2 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，長すぎると冗長であり，短すぎると重要な内容を欠くおそれがある．本来は，話されている意味的な内容を解釈した上で，各ホットスポットごとに適切に定めることが望ましいが，これを実現することは難しいため，本研究では各ホットスポット区間を，自動分割によって得られたセグメントの数と時間長に関するしきい値によって決定する．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．

久保田ら[13]らは，会話参与者が興味深く感じたミーティング中の会話シーンを会話に参加しながら切り出すために，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 45.96 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．これに対して本論文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ることができる．本研究ではセグメント長の制約 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 秒としてそれぞれ設定した．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，生起させる発話区間が長いことが予想されるため，なるほどスポットをやや長めに設定した．

5. 被験者実験によるホットスポットの評価

ポッドキャストとポスター会話を対象として，音響イベント検出に基づいて自動的に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．

5.1 実験条件と設問

テストセットして，ポッドキャストからお笑い番組，教養バラエティ，インタビューの 3 番組 2 エピソードずつの計 6 エピソードを，またポスター会話から 2009 年に収録された 4 セッションを用いた．

音響イベント検出手法[2]を適用して得られる笑い声とあいづちの検出精度は，ポッドキャストに対して F 値 0.687，0.640，ポスター会話に対して F 値 0.663，0.659 であった．結果の詳細を表 1 にまとめる．

これらのテストセットに対して，4 名の被験者が各人につきポッドキャストを 2 エピソー

表 1 音響イベントの検出精度

	笑い声			あいづち		
	再現率	適合率	F 値	再現率	適合率	F 値
ポッドキャスト	0.650	0.713	0.687	0.340	0.852	0.640
ポスター会話	0.396	0.797	0.663	0.412	0.775	0.659

図 2 アンケート入力のインタフェース

ド，ポスター会話を 2 セッションずつ聴取する．被験者は評価用インタフェース（図 2）上で操作を行って，抽出したホットスポット（音リアクションイベントを含む）をエピソード毎に時系列順に聴き，それぞれのスポットについてアンケートに回答する．設問形式は表 2 の通りであり，選択肢から答えるものとする．

Q1 はホットスポット抽出の成否に関する設問である．Q1 で生起する理由がわ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る．この集計に基づいて，イベント誤検出も含む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，それぞれ検出率を求めた．また Q2 と Q3 から，被験者自身がその箇所を聴いて主観的にどう感じたかを調査した．それぞれ Q1 の回答別に（生起箇所を抽出できているかどうか）集計を行った．

実験に際して，例えばポッドキャストを視聴する際には，RSS や配信先のウェブページなどからテキストによる番組情報を取得できるため，被験者には視聴の前にそのような予備知識（番組概略やセッションテーマなど）を教示した．またそれぞれのスポットを個別に評価するのではなく，通常の視聴と同じように時間順に聴いて評価することで，得られた情報を以後の判断に利用できるようにした．

表 2 アンケートの設問形式

おもしろスポット

No.	設問	選択肢
Q1	笑い声が出現する理由がわかったか？	はい／いいえ
Q2	被験者がおもしろいと感じたか？	意味不明／面白くない／前後なしで判断不可／面白味は感じる／おもしろい
Q3	該当スポットが視聴する上で必要と思うか？	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

No.	設問	選択肢
Q1	あいづちが出現する理由がわかったか？	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	該当スポットが視聴する上で必要と思うか？	不要／ない方がよい／あった方がよい／必要

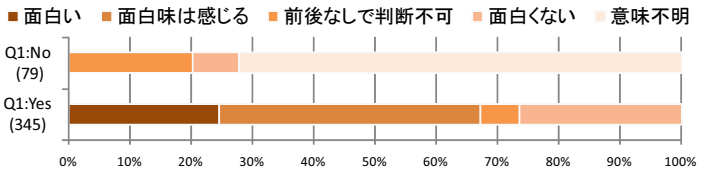
表 3 ホットスポットの検出率

	ポッドキャスト	ポスター会話
各スポット	検出率 (抽出数/出力数)	検出率 (抽出数／出力数)
出力された全おもしろスポット	81.4% (345/424)	74.7% (68/91)
(うち笑い声を正しく検出)	91.1% (338/371)	89.2% (66/74)
出力された全なるほどスポット	89.4% (143/160)	86.5% (128/148)
(うちあいづちを正しく検出)	90.5% (133/147)	95.2% (119/125)

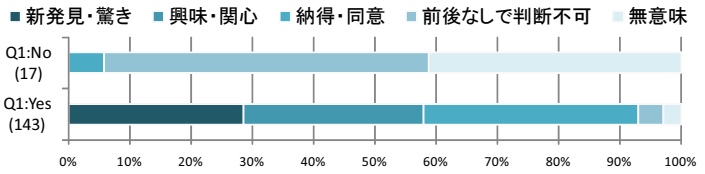
5.2 実験結果と考察

Q1 の結果を表 3 に示す．誤検出も含む全出力ホットスポットに対して，74.7%〜89.4%の検出率が得られ，そのうち正しく音リアクションイベントを検出したホットスポットに対しては，検出率は 89.2%〜95.2%であり高い精度で抽出できていた．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}の設定が妥当であったといえる．おもしろスポットとなるほどスポットの検出率を比較すると，おもしろスポットの方がやや低い値となっている．その理由として，笑い声が「おもしろい」場合だけではなく,「意外」や「照れ隠し」などを表現する場合にも多く出現するためである．特にポスター会話ではそのような笑い声が多かったため，検出率が低くなったと考えられる．

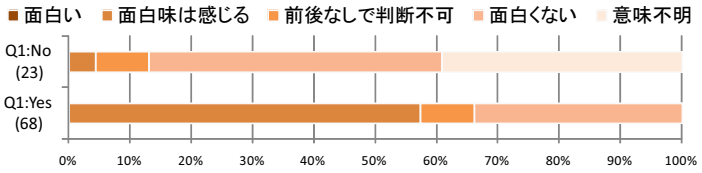
Q2 の集計結果を図 3 に，Q3 の集計結果を図 4 にそれぞれ示す．なるほどスポットに関して，Q2 の結果（図 3-(b),(d)）をみると，ポッドキャストとポスター会話のどちらにお



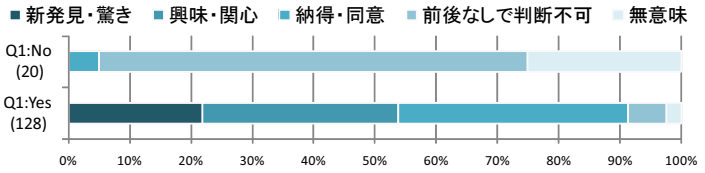
(a) ポッドキャスト中のおもしろスポット



(b) ポッドキャスト中のなるほどスポット

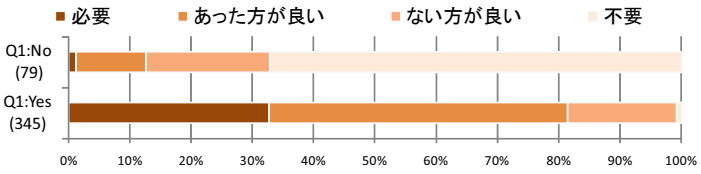


(c) ポスター会話中のおもしろスポット

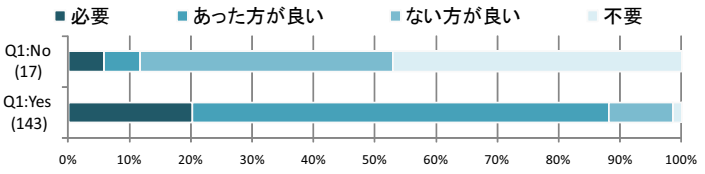


(d) ポスター会話中のなるほどスポット

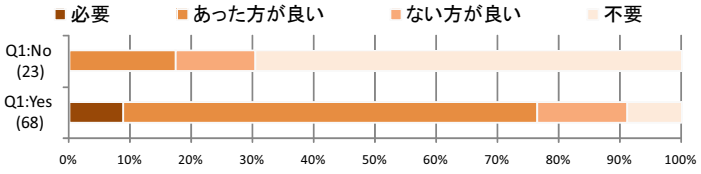
図 3 Q2 に対する集計結果



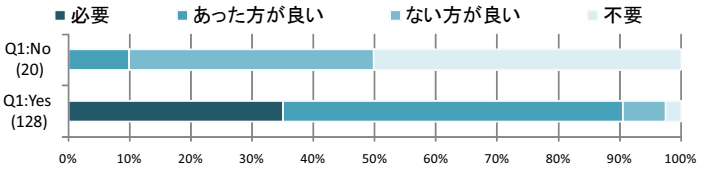
(a) ポッドキャスト中のおもしろスポット



(b) ポッドキャスト中のなるほどスポット



(c) ポスター会話中のおもしろスポット



(d) ポスター会話中のなるほどスポット

図 4 Q3 に対する集計結果

いても，生起箇所を抜き出せていた（Q1 で「はい」と答えた）場合の 9 割の箇所に対して，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．さらにそれらの大部分は，Q3 で「必要」もしくは「あった方がよい」と回答されており（図 4-(b),(d)），聴取すべき有益な情報を含んだ箇所を的確に抽出できていたといえる．

これに対して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」もしくは「面白味は感じる」と回答した箇所は，ポッドキャストで約 7 割，ポスター会話で 6 割程度であった（図 3-(a),(c)）．これは「おもしろい」かどうかという判断がより主観的であり，コンテンツ間での個人差が強く現れたためだと考えられる．しかし，おもしろくないと感じた場合でも，それらはコンテンツ内でインタラクションが活発に行われた箇所であるため，有益な情報が含まれている場合がある．そのため，Q3 で「必要」もしくは「あった方がよ

い」と回答されたおもしろスポットは約 8 割を占めていた（図 4-(a),(c)）．

また笑い声が聞き手の反応を表すだけでなく，緊張をほぐす用途で使われるなど，多人数インタラクションの中で円滑にコミュニケーションをとるための手段として使用されている箇所がいくつかみられた．本研究では，そのような笑い声とリアクションとしての笑い声を同様に扱っているが，計算機で自動判別することができれば，コミュニケーション分析のひとつの手がかりとしても，笑い声が効果的に利用できると考えられる．

5.3 再現率に関する議論

以上は出力されたホットスポットの適合率の議論であるが，これに対して提案法により全体の有用箇所をどの程度抽出できるのかの再現性についても検討する．ただし，おもしろい，興味・関心のあるといった意味で有用となる箇所は視聴者の主観に依存し，人によって捉え方は様々であるため,「正解」をアノテーションすることは容易ではない．そのため，ここでは定性的な議論にとどめる．

視聴者にとって意味がある箇所としては，拍手や言語的な応答（「すごい」「なるほど」など）といった他のリアクションをともなう箇所や，会話のトピックに深く関連する単語（話題語）を含んだ箇所が考えられる．また今回利用したポスター会話のようなマルチモーダルなデータでは，表情の変化やポインティングが手がかりとなりうる．これらが笑い声やあいづちをともなっていなければ，本研究の枠組みで抽出を行うことはできない．しかし，一般的にそのような箇所は笑い声やあいづちをともなう可能性が高く，実際に本研究で扱ったポッドキャストやポスター会話では，これらの特徴的な箇所の多くで笑い声やあいづちが聞き手の反応として出現していた．

6. おわりに

本研究では聞き手の反応に基づく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 種類を定めた．音響イベント検出の結果に基づいて，笑い声やあいづちの直前のセグメントから時間長とセグメント数の制約をもとに各スポットを抽出する手法を提案した．被験者実験により評価したところ，高い精度でホットスポットが抽出されており，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・関心をもったことが示された．

我々は，興味・関心と関連の深いあいづちの韻律パターンの分析も行なっており[14)15)]，今後このような知見も統合することで，より精度の高いホットスポットの抽出ができるものと期待できる．

謝辞：本研究は JST CREST 及び科学研究費補助金によって行われた．

参 考 文 献

1) M.Pantic and Vinciarelli, A.: Implicit Human-Centered Tagging, *Signal Processing Magazine*, Vol.26, No.6, pp.173–180 (2009).
2) Sumi, K., Kawahara, T., Ogata, J. and Goto, M.: Acoustic Event Detection for Spotting "Hot Spots" in Podcasts, *Proc. Interspeech*, pp.1143–1146 (2009).
3) 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, pp.35–40 (2007).
4) 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, pp.41–46 (2007).
5) T.Kawahara, M.Hasegawa, K.Shitaoka, T.Kitade and H.Nanjo: 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).
6) Hirohata, M., Shinnaka, Y., Iwano, K. and Furui, S.: Sentence-extractive Automatic Speech Summarization and Evaluation Techniques, *Speech Communication*, Vol.48, No.9, pp.1151–1161 (2006).
7) 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，信学論， Vol.91-D, No.2, pp.238–249 (2008).
8) Wrede, B. and Shriberg, E.: Spotting "Hot Spots" in Meetings: Human Judgments and Prosodic Cues, *Proc. Eurospeech*, pp.2805–2808 (2003).
9) Kennedy, L. and Ellis, D.: Pitch-based Emphasis Detection for Characterization of Meeting Recordings, *Proc. ASRU*, pp.243–248 (2003).
10) D.Gatica-Perez, I.McCowan, D.Zhang and S.Bengio: Detecting Group Interest-Level in Meetings, *Proc. IEEE-ICASSP*, Vol.1, pp.489–492 (2005).
11) 角　康之，西田豊明，坊農真弓，來嶋宏幸：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，Vol.49, No.8, pp.945–949 (2008).
12) Kawahara, T., Setoguchi, H., Takanashi, K. et al.: Multi-modal Recording, Analysis and Indexing of Poster Sessions, *Proc. Interspeech*, pp.1622–1625 (2008).
13) 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，情処学論， Vol.48, No.12, pp.3703–3714 (2007).
14) 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).
15) T.Kawahara, Z.Q.Chang and K.Takanashi: Analysis on Prosodic Features of Japanese Reactive Tokens in Poster Conversations, *Proc. Int'l Conf. Speech Prosody* (2010).