

# スマートポスターボード：ポスター会話のマルチモーダルなセンシングと解析

## Smart Posterboard: Multi-modal Sensing and Analysis of Poster Conversations

河原達也

Tatsuya KAWAHARA

京都大学

Kyoto University

### Abstract

学会等で一般的に行われているポスター発表における会話（ポスター会話）は、マルチモーダルな多人数インタラクションに関する様々な興味深い研究テーマを提供してくれる。本稿では、著者らが進めているポスター会話のマルチモーダルなセンシングと解析に関するプロジェクト、及び構築しているスマートポスターボードシステムの概要を紹介する。本システムは、ポスターボードに設置した複数のセンサを用いて会話を記録し、誰がポスターに来て、どのような質問やコメントを行ったかを容易に検索できるようにすることを目指している。そのためにも、顔向き（視線）検出と話者区間検出を統合したマルチモーダルな信号処理を実現する。さらに、視線配布や相槌などの聴衆のマルチモーダルな振る舞いに着目することで、興味・理解度の推定を試みる。

### 1 はじめに

マルチモーダルな信号・情報処理に関する研究は従来、人間型ロボットを含むヒューマンマシンインターフェースの高度化を主な目標として行われてきた。一方、画像処理や音声処理が高度になり、上記のようなインターフェースを意識しない人間の自然なふるまいも扱えるようになり、いわゆるアンビエントなシステムを目指した研究開発も可能になっている。実際に、ミーティング [1] や自由会話 [2] などの人間どうしの会話を対象とした研究も行われている。

我々はポスターセッションにおける会話（＝ポスター会話）に焦点をおいたプロジェクトを進めている [3, 4, 5, 6]。ポスターセッションは、学会やオープンラボなどで一般的になっているが、未だに情報通信技術 (ICT) の導入が

ほとんどなされておらず、ICT 分野の会議でも紙のポスターを用いることが多い。一部液晶ディスプレイや携帯プロジェクトを用いる場合もあるが、センサを備えた環境は世界的にも前例がないと思われる。講演や講義の映像・音声収録・配信されることが一般的になっているのに対して、ポスターセッションを収録して分析した研究も皆無に近い。

ポスター会話は、講演と会議の中間的な形態と捉えることができる。すなわち、発表者が自身の研究内容について少人数の聴衆に説明する一方、聴衆の側も相槌や頷きなどでリアルタイムにフィードバックし、時折質問やコメントも行う。また会議と違って、参加者は立っており、動くこともできるので、マルチモーダルなインタラクションを行うことが多い。さらに、ポスター会話を扱う利点としては、話題や他の参加者に対する親近性を制御しながら、（研究者を集めてくれば）自然でリアルなデータを収集することが非常に容易であることが挙げられる。

本プロジェクトの目標は、人間どうしのインタラクションの信号レベルのセンシングとより高いレベルの解析である。人間型ロボットを含むヒューマンマシンインターフェースと比較して、長時間にわたる自然な振る舞いを扱う点が最大の違いである。認識のタスクとしては、人物・視線・話者・発話区間などの検出がある。これらは会話アーカイブに対する新たなインデキシングの枠組みを提供する。例えば、自身あるいは同僚のポスターセッションが終わった後で、どのくらいの聴衆がやって来て、どのような質疑・コメントが行われたかを振り返ることができるようになる。

さらに、どの部分に興味を持ってもらえたか、どこがわかりにくいところであったか、といった解析も研究する。動画投稿サイトなどの事例からもわかるように、我々はその人が興味を持ったものを視聴したくなるのが普通であるので、このようなアノテーションは有用であると考えられるが、アノテーションの基準や評価を含めて、これらを

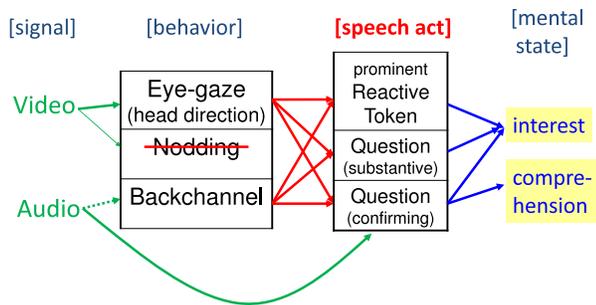


図 1: マルチモーダルなセンシングと解析の枠組み

明確に定義するのは非常に困難である。そこで、これらの心的状態に関係すると考えられ、客観的に観測できる発話行為 (speech act) に着目する。具体的には、聴衆による質問と特定のパターンの相槌 (reactive token) に着目する。さらに、質問を確認質問 (confirming) と踏み込み質問 (substantive) に分類する。マルチモーダルな振る舞いからこれらの発話行為を予測することで、興味・理解度の推定を近似できると期待している。提案する枠組みを図 1 に示す。

本稿では、まず 2 章でセンシング環境と収集したコーパスの説明を行う。3 章では、図 1 の左側の部分、すなわち信号レベルから振る舞いの検出を行う処理について述べる。4 章では、図 1 の右側の部分、すなわち発話行為と心的状態を関係づけて、興味・理解度を定義する。5 章では、図 1 の中央の部分、すなわち、聴衆のマルチモーダルな振る舞いからの発話行為の予測について述べる。

## 2 ポスター会話のマルチモーダルコーパス

### 2.1 収録環境：スマートポスターボード

我々は、ポスター会話における音声・映像と振る舞いなどのマルチモーダルな情報を収録するための環境の構築を進めてきた [7, 8]。また、ポスターボードを大型液晶ディスプレイで構成し、これに多様なセンサを設置することで、ポスター会話をセンシングするシステム (=スマートポスターボード) を構築している。スマートポスターボードの概観を図 2 に示す。

音声に関しては、ポスターボードの上に設置するマイクロフォンアレイを設計した。映像に関しては、参加者全員とポスターをカバーできるように、6~8 個のカメラをポスターボードに設置した。また、Kinect センサも設置した。<sup>1</sup>

ただしコーパス構築の上では、正確な情報 (ground

<sup>1</sup>簡易版は Kinect センサのみを用いる。



図 2: スマートポスターボードの概観

truth) を取得する必要がある。そのために、各参加者にワイヤレスのヘッドセットマイクを装着してもらうとともに、様々なセンサを着用してもらった。当初はモーションキャプチャシステムや視線計測装置を使用した。直近は磁気センサを使用している。

### 2.2 コーパスとアノテーション

上記の環境を用いて、これまで 5 ケ年度にわたって合計 43 セッションのポスター会話を収集してきた。ただし、いくつかのセンサデータが欠損したものも含まれる。

各セッションにおいては、1 名の発表者 (A と表記) が自身の研究に関する発表を、2 名の聴衆 (B, C と表記) に対して行う。聴衆は、発表者についても研究内容についても初めて接する設定となっている。セッションの長さは制御しているわけではないが、おおむね 20~30 分程度である。

音声データは、ヘッドセットマイクで収録されたものをポーズで区切られた発話単位 (IPU) に分割し、時間と話者ラベルを付与した上で、『日本語話し言葉コーパス』(CSJ) と同じ基準で書き起こしを行った。ただし、フィラー以外に相槌と笑いに対してもアノテーションを行った。

視線情報は、視線計測装置とモーションキャプチャシステム、または磁気センサのデータを用いて、視線ベクトルと他の参加者やポスターの位置との交差判定に基づいてアノテーションを行った。

以降の章の実験では、2012 年度に収集・アノテーションを行なった 4 つのセッションを主に用いている。これらのセッションにおける発表者と聴衆の組合せはすべて異なっている。

## 3 マルチモーダルな振る舞い (視線・発話) の検出

音声や映像の信号から、各会話参加者の振る舞いを検出する処理については、順次研究開発を進めている。ここでの

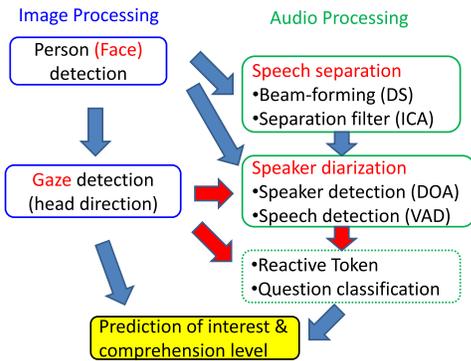


図 3: マルチモーダル統合による振る舞い検出の処理の流れ

目標は、参加者にマイクなどの装置を一切装着してもらわずに、ポスターボード等に設置したセンサだけで処理を行うことである。具体的には、マイクロフォンアレイで収録される音声信号、及び複数台のカメラ（Kinect センサ）で得られる映像信号・距離情報を用いる。

各処理の高度化と適用を行なうとともに、これらのマルチモーダルな情報を統合する方法を検討している。全体の処理の流れを図3に示す。まず聴衆の人物（顔）を検出し、各人の顔向き（視線）を検出・追跡する。人物の位置情報は、音響処理、具体的には音声強調と話者同定において利用される。また、視線情報は発話区間検出や質問種類の同定に用いる。

### 3.1 Kinect センサによる視線（頭部方向）検出

本研究では Kinect センサを用いて、ポスター会話に適した実用的な視線推定を実現した。なお、画像の解像度及び眼鏡等の影響により眼球そのものを安定に撮影することは困難であるため、視線方向を頭部方向で代用する。視線と頭部方向のずれは平均 10 度程度で、ポスターを注視する状況ではさらに小さくなる傾向がある [9]。処理は以下の手順で行っている。

#### 1. 顔検出

Kinect センサで撮影したカラー画像及び距離画像から、Haar-like 特徴を利用した物体認識法を用いて正面顔探索を行う。同時に複数人を処理することが可能である。

#### 2. 頭部モデル獲得

顔検出結果に従って、距離画像から頭部の 3 次元形状を、カラー画像からその色情報を計算する。計算結果はポリゴンとテクスチャ情報に変換し、頭部モデルとする。

### 3. 頭部追跡

頭部方向の推定を、画像への頭部モデルのフィッティングとして行う。具体的には、頭部を剛体とみなし、頭部の 3 次元位置と姿勢を表す 6 変数をパーティクルフィルタによる追跡処理で逐次計算する。

### 4. 注視対象特定

頭部追跡処理で得られた 6 変数から、3 次元空間で視線に対応する半直線を求める。この半直線と、ポスターボードや他の参加者との交差判定を行うことで、注視対象を決定する。

上記の処理は、GPU を使うことで、オンライン・リアルタイムに実行可能である。

### 3.2 マイクロフォンアレイによる音声の分離・強調

音声の分離と強調は、ブラインド空間的サブトラクションアレイ (BSSA)[10] によって実現する。これは、マイクロフォンアレイで得られる信号に対して、遅延加算 (Delay-and-Sum) 型ビームフォーミングを行うとともに、独立成分分析 (ICA) に基づいて各会話参加者の音声と背景雑音を分離し、目的信号以外の抑圧を行うものである。ポスター会話の設定では、発表者・聴衆・背景雑音の 3 つの成分への分離を行う（聴衆間の分離は行われない）。その際に、画像処理によって得られる各参加者の位置情報を用いることで、ICA のフィルタ計算の高速化を実現している。この処理を逐次的に行うことで、参加者が移動しても追跡できるようにしている。

19 チャンネルのマイクロフォンアレイを用いる場合は、高い品質の音声強調ができるが、リアルタイムには処理できない。Kinect センサ内蔵の複数のマイクロフォンを用いる場合は、音質は低下するが、リアルタイム処理が可能である。

### 3.3 音響情報と画像情報を統合した話者区間検出

話者区間検出 (speaker diarization) は、“いつ誰が発話したか”を検出する処理で、話者同定 (speaker localization) と発話区間検出 (voice activity detection) の 2 つの要素からなる。そのために、マイクロフォンアレイで得られる音響情報（音のパワーと位相情報）に加えて、画像から得られる各参加者の位置情報を利用する。マルチモーダルな発話区間検出として、口唇の動きを用いることも考えられるが、解像度の高い正面画像が必要なため、ポスター会話においては現実的でない。

マイクロフォンアレイを用いた音源の到来方向推定 (DOA estimation) の代表的な手法である MUSIC 法 [11] を用いる。MUSIC 法は、観測信号の部分空間の直交性に基づいて、同時に複数の音源をリアルタイムに推定することができ、各時刻  $t$ ・各方向  $\theta$  に関して、そこに音源が存在する尤度  $P_{MU}(t, \theta)$  を求めることができる。

表 1: 話者区間検出結果

	F 値 ( 発表者 )	F 値 ( 聴衆 )	DER
音声分離+パワー	0.880	0.515	38.5%
MUSIC 法	0.920	0.581	47.5%
ルールベース統合	0.921	0.591	42.3%
確率的統合	0.887	0.686	29.9%

ベースライン手法では、この尤度  $P_{MU}(t, \theta)$  の極大値を探索し、しきい値以上となるものを音源とみなす。このときに、画像処理による顔検出で得られる参加者の位置情報を利用する。すなわち、 $P_{MU}(t, \theta)$  がしきい値以上であり、かつ  $\theta$  が参加者の推定位置からしきい値以内である場合に、発話がなされたと判定する。

さらに、確率的な統合 [12, 13] も検討する。すなわち、画像情報により得られる参加者の位置情報に関して、推定位置を平均、その信頼度を分散とする正規分布に基づいて尤度を算出し、 $P_{MU}(t, \theta)$  と統合する。

4 セッションに対する話者区間検出の結果を表 1 に示す。発表者はマイクに近く、発話が多いため、90%に近い検出精度が得られるのに対して、聴衆の発話区間検出は困難である。前節の音声分離結果に対してパワーに基づいて単純に発話区間検出する方法では F 値が 50%程度になっている。MUSIC 法ではそれより高い精度が得られるが、単純に画像による位置情報をヒューリスティックに用いてもほとんど効果が見られない。それに対して、確率的な統合により、70%程度まで改善している。ただし、雑音を重畳すると性能が低下するので、その対策も検討する必要がある。

#### 4 興味・理解度の定義

興味・理解度のアノテーションを行う最も自然な方法は、ポスターセッション終了後に聴衆の各人に、各々のスライド話題単位に対する興味と理解の評定を行ってもらうことである。しかしながら、このようなアンケート調査を大規模に行うことはあまり現実的でないし、既に収録済みのセッションに行くことは不可能である。またこのような評定は主観的で、その信頼性を評価することも難しい。

そこで本研究では、興味・理解度に関係が深いと考えられ、客観的に観測可能な発話行為に着目する。これまでに我々は、「へー」「あー」「ふーん」といった非語彙的・引き延ばし型で韻律的にも顕著な特徴を持つ相槌(=顕著な相槌)が聴衆の興味と関係があることを明らかにした [14, 15]。Ward ら [16] は英語の相槌に関して、そのパターンと役割の分析を行っている。

また経験的に、聴衆の質問の生起は興味と関係があると考えられる。すなわち、聴衆は発表に引きつけられるほ

ど、より多くの質問をするものである。また、質問の種類を調べることで、理解度を推測することもできる。例えば、既に説明されたことを質問しているなら、理解が困難であったことを示唆している。

#### 4.1 質問の種類のアノテーション

本研究では、質問を確認質問と踏み込み質問に分類した。確認質問は、現在の説明の理解が正しいか確認するために行うもので、「はい/いいえ」のいずれかで答えることができる。<sup>2</sup> これに対して踏み込み質問は、発表者の説明に含まれていなかったことに関して尋ねるもので、「はい/いいえ」のみで答えられるものでなく、何らかの補足説明が必要になる。踏み込み質問は、表層的には質問の形式をとっているが、実質的にコメントに近い場合もある。

#### 4.2 質問の種類と興味・理解度との関係

2012 年度に収録した 4 つのセッションについては、終了後に聴衆の各人に各スライド話題単位に対する興味と理解の度合いを評定してもらった。そこで、このデータを用いて、評定と質問との関係を調べた。

図 4 に、2 種類の質問( confirming: 確認質問; substantive: 踏み込み質問 )の生起毎、及び全話題セグメント(entire)の興味・理解度の分布を示す。興味度については、1(低い)から 5(高い)の 5 段階で評定してもらい、理解度については、1(低い)から 4(高い)の 4 段階で評定してもらっている。左のグラフから、質問の種類に関わらず、質問が生起している場合には全般に興味が高い(4 か 5) ことがわかる。また右のグラフから、確認質問の大多数(86%)が理解度が低い(1 か 2) ことと相関があることがわかる。

この分析結果と顕著な相槌に関する先行研究 [15] を踏まえて、分析対象の全話題セグメントに対して、以下のアノテーションの枠組みを採用した。

- 興味が高い ← (種類に関わらず) 質問もしくは顕著な相槌が生起している
- 理解度が低い ← 確認質問が生起している

これらのアノテーションは各話題セグメントに対して、聴衆 2 名の各人について行った。これらの心的状態の検出は、ポスター会話を後で振り返る際に有用であると考えられる。

#### 5 マルチモーダルな振る舞いに基づく発話行為の予測

聴衆の興味・理解度の推定を、関係する発話行為の予測を行う問題として定式化した。すなわち、興味の推定は質問

<sup>2</sup>ただし、実際に発表者が「はい/いいえ」のみで答えたとはいえない。

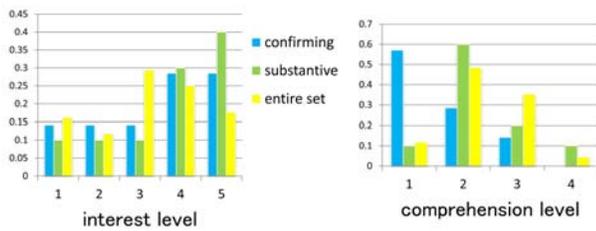


図 4: 質問の種類毎の理解・興味度の分布

と顕著な相槌の生起の予測に、理解度の推定は質問タイプの分類に帰着させた。この予測を、当該の発話行為が(話題セグメントの終わり頃に)実際に生起する前に、聴衆のマルチモーダルな振る舞いを元に行う。これにより、そのような発話行為が実際に生起しなくても、聴衆の心的状態の推定が可能になると期待できる。

まず、各聴衆の振る舞いを特徴量にする必要がある。マルチモーダルな振る舞いとして、相槌と視線配布に着目した。相槌については、発表者の発話で正規化した平均頻度を求めた。発表者に対する視線配布については、発表者の発話で正規化した出現頻度と継続時間割合を求めた。

次に、識別のための機械学習の方法については、ナイーブベイズ分類器を用いた。これは、学習データが少なく、各特徴量の重みなどのパラメータを推定することが困難であるためである。特徴量ベクトル  $F = \{f_1, \dots, f_d\}$  に対するナイーブベイズ分類は、以下の事後確率に基づいて行われる。

$$p(c|F) = p(c) * \prod_i p(f_i|c)$$

ここで、 $c$  は分類カテゴリであり、ここでは「興味を持ったか否か」である。また、 $p(f_i|c)$  を計算するには、ヒストグラム量子化を用いた。これは、特徴量の値を量子化ビンに割り当てるもので、確率密度関数を仮定しないためモデルパラメータの推定を必要としない。特徴量の分布ヒストグラムを単純に 3 ないし 4 に分割して量子化ビンを設定する。その上で、各ビンの相対的な出現頻度を確率値に変換する。

本実験では 2009 年度と 2011 年度に収録したものを含めて合計 10 セッションを用いた。この 10 セッションには計 58 個のスライド話題単位があった。各セッションに 2 名の聴衆がいるので、興味・理解度を推定すべきスロット (= 話題セグメント) が合計 116 個あることになる。評価実験は、セッション単位の leave-one-out クロスバリデーションにより行った。

### 5.1 質問・顕著な相槌の生起の予測：興味度の推定

まず、各話題セグメントにおける聴衆の興味度を推定する実験を行った。これは聴衆が質問ないし顕著な相槌を生起するかを予測する問題に帰着される。すなわち、当該の

表 2: 質問・顕著な相槌を含む話題セグメントの予測結果 (興味度の推定)

	F 値	正解率
ベースライン	0.49	49.1%
(1) 相槌	0.59	55.2%
(2) 視線 頻度	0.63	61.2%
(3) 視線 時間	0.65	57.8%
(1)-(3) の組合せ	0.70	70.7%

表 3: 確認質問 / 踏み込み質問の同定結果 (理解度の推定)

	正解率
ベースライン	51.3%
(1) 相槌	56.8%
(2) 視線 頻度	75.7%
(3) 視線 時間	67.6%
(1)-(3) の組合せ	75.7%

発話行為を行った聴衆は、その話題セグメントに「興味を持った」とみなす。

種々の特徴量に対する結果を表 2 に示す。正解率は計 116 の話題セグメントで正しい判定が得られたものの割合である。なお、すべての話題セグメントに「興味を持った」とした場合の (chance rate) ベースラインは、49.1% である。

相槌と視線の特徴を用いることで、有意に高い正解率が得られ、両者を組み合わせることで 70% を上回る結果となった。ただし、視線に関する 2 つの特徴量 (頻度と時間) については一方を外しても結果は変わらなかった。以上、相槌と視線のマルチモーダルな統合効果を確認した。

### 5.2 質問の種類と同定：理解度の推定

次に、各話題セグメントにおける聴衆の理解度を推定する実験を行った。これは聴衆が質問を行った際に、質問の種類を同定する問題に帰着させる。すなわち、確認質問を行った聴衆は、その話題セグメントの「理解が困難であった」とみなす。

確認質問 / 踏み込み質問の分類結果を表 3 に示す。なお、このタスクでは各質問の出現頻度  $p(c)$  に基づく (chance rate) ベースラインは、51.3% である。

すべての特徴量が正解率の向上に一定の効果があったが、視線の出現頻度のみで最良の正解率が得られ、他の特徴量と組み合わせても相乗効果は見られなかった。質問が生起する直前の 2 発話における視線特徴量の時間的な変化を用いることも検討したが、改善は得られなかった。

## 6 おわりに

本稿では、ポスター会話をセンシングするスマートポスターボードの紹介を行なった。ポスター会話では、複数人が動きながら遠隔で発話を行なう点でヒューマノイドロボットの場合と同様であるが、長時間の自然な会話であり、しかも発表者に比べて聴衆の発話が圧倒的に少ないので、より挑戦的なタスクと考えられる。現時点では、発表者に対して約 90%、聴衆に対して約 70%の話者区間検出精度となっている。

さらに、視線情報などをインタラクションの高度な解析に用いることを研究している。これにより、会話の流れを捉えるとともに、興味・理解度の推定も可能になると考えている。興味・理解度ともに 70%程度の精度で推定可能なことを示した。

これらの情報を可視化するブラウザも構築しており、長時間のポスター会話を振り返る上で有用であるか検証を進めていく予定である。

## 謝辞

本研究は、JST CREST「人間調和型情報環境」領域ならびに科学研究費補助金の支援を受けて実施されたものである。特に、スマートポスターボード(2章・3章)の研究開発は、CRESTプロジェクトに参画して頂いている京都大学の吉本廣雅研究員、Tony Tung 特定助教、若林佑幸研究員、井上昂治君と奈良先端科学技術大学院大学の猿渡洋准教授をはじめとする多くの方々のご貢献によるものである。また、4章・5章の研究は、京都大学の高梨克也研究員と林宗一郎君におうものである。

## 参考文献

- [1] S.Renals, T.Hain, and H.Bouclard. Recognition and understanding of meetings: The AMI and AMIDA projects. In *Proc. IEEE Workshop Automatic Speech Recognition & Understanding*, 2007.
- [2] K.Ohtsuka. Conversation scene analysis. *Signal Processing Magazine*, Vol. 28, No. 4, pp. 127–131, 2011.
- [3] T.Kawahara. Multi-modal sensing and analysis of poster conversations toward smart posterboard. In *Proc. SIGdial Meeting Discourse & Dialogue*, pp. 1–9 (keynote speech), 2012.
- [4] T.Kawahara. Smart posterboard: Multi-modal sensing and analysis of poster conversations. In *Proc. APSIPA ASC*, p. (plenary overview talk), 2013.
- [5] 河原達也. [招待講演] スマートポスターボード: ポスター会話のマルチモーダルなセンシングと認識. 電子情報通信学会技術研究報告, SP2012-51, 2012.
- [6] 河原達也. [特別講演] スマートポスターボード: ポスター発表における場のマルチモーダルなセンシングと認識. 電子情報通信学会技術研究報告, PRMU2012-167, 2013.
- [7] 瀬戸口久雄, 高梨克也, 河原達也. 多数のセンサを用いたポスター会話の収録とその分析. 情報処理学会研究報告, SLP-67-6, 2007.
- [8] T.Kawahara, H.Setoguchi, K.Takanashi, K.Ishizuka, and S.Araki. Multi-modal recording, analysis and indexing of poster sessions. In *Proc. INTERSPEECH*, pp. 1622–1625, 2008.
- [9] 矢野正治, 中田篤志, 福間良平, 角康之, 西田豊明. 非言語マルチモーダルデータを用いた会話構造の分析のための環境構築. 情処学研報, 2009-UBI-22-12, 2009.
- [10] Y.Takahashi, T.Takatani, K.Osako, H.Saruwatari, and K.Shikano. Blind spatial subtraction array for speech enhancement in noisy environment. *IEEE Trans. Audio, Speech & Language Process.*, Vol. 17, No. 4, pp. 650–664, 2009.
- [11] R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Trans. Antennas & Propagation*, Vol. 34, No. 3, pp. 276–280, 1986.
- [12] 中村圭佑, 中臺一博, 浅野太, 中島弘史, Ince Gokhan. マルチモーダル情報統合によるインテリジェント人追跡システム. 計測自動制御学会論文集, Vol. 48, No. 6, pp. 349–358, 2011.
- [13] K.Nakamura, K.Nakadai, F.Asano, and G.Ince. Intelligent sound source localization and its application to multimodal human tracking. In *Proc. IROS*, 2011.
- [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, K.Sumii, Z.Q.Chang, and K.Takanashi. Detection of hot spots in poster conversations based on reactive tokens of audience. In *Proc. INTERSPEECH*, pp. 3042–3045, 2010.
- [16] N.Ward. Pragmatic functions of prosodic features in non-lexical utterances. In *Speech Prosody*, pp. 325–328, 2004.