一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
SP2012-51 (2012-07)

［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

河原　達也†

† 京都大学 学術情報メディアセンター
〒 606-8501 京都市左京区吉田本町
E-mail: †kawahara@i.kyoto-u.ac.jp, http://www.ar.media.kyoto-u.ac.jp/crest/

あらまし　学会等で一般的に行われているポスター発表に伴う会話 (ポスター会話) は、マルチモーダルな多人数会話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・分析・認識に関するプロジェクトの概要を紹介する。我々は特に、相槌・頷き・視線配布などの聴衆の反応 (聞き手行動) に着目している。具体的には，聴衆の相槌の韻律パターンや発表者と聴衆の視線などの情報から、いつ誰がどのような質問をするかを予測したり、聴衆の興味度を推定することを検討している。さらに、ポスター会話をマルチモーダルに収録した上で、会話参加者の行動やインタラクションパターンを捉える「スマートポスターボード」を設計・実装している。

キーワード　マルチモーダル, 音声対話, 会話分析, 話者交替, ポスターボード, マイクロフォンアレイ

[Invited Talk] Smart Posterboard: Multi-modal Sensing and Recognition of Poster Conversations

Tatsuya KAWAHARA†

† Kyoto University, Academic Center for Computing and Media Studies, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan
E-mail: †kawahara@i.kyoto-u.ac.jp, http://www.ar.media.kyoto-u.ac.jp/crest/

1. はじめに

様々な音声対話システムが開発・実用化されるにつれて、工学的な観点での音声対話研究も、従来の人間と機械の音声インタフェースから新たな展開を見せている。一つの方向はマルチモーダルインタフェースで、これは携帯端末やキオスク端末等のディスプレイだけでなく、人間型ロボットも含む。もう一つの方向は多人数会話で、これに対応した受付エージェント [1] やロボット [2] の研究も行われている。これらが人間と機械の音声対話の発展であるのに対して、会議 [3], [4] や自由会話 [5] 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 [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、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、マルチモーダルなコミュニケー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認識である。認識のタスクとしては、話者インデキシング、(主に説明者の) 音声認識、聴衆の理解度や興味度のアノテーションなどを考えている。これらは音声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ことが期待される。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、どういう質疑が行われたか、どこが聴衆にとってわかりにくいところであったか、といった要求に応えることができる。また本研究は、将来自動でプレゼンテーションを行う知的な会話エージェントの基盤にもなる。

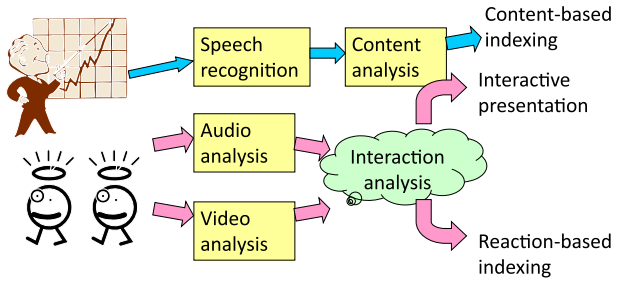

図 1　マルチモーダルなインタラクション分析の概要

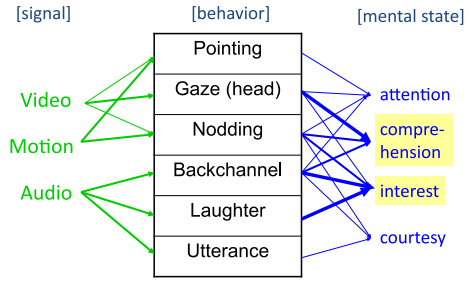

図 2　マルチモーダルなセンシングと分析の概要

これまで、発話者の音声の認識や言語解析などに基づいた、内容に基づく (content-based) インデキシングのアプローチが研究されてきたが、これに対して、聞き手の反応に着目した (interaction-based) アプローチを提案する。具体的には、相槌・頷き・視線配布などの非言語情報に注目する。現状の音声認識・言語解析システムと比べて、人間の聴衆の方がはるかに説明内容を理解しているのは自明であるので、その反応を捉える方が合理的と考える。この枠組みを図 1 に示す。

そのために、多人数会話のマルチモーダルなセンシングと分析を行うための基盤を構築する。処理の概要を図 2 に示す。音響信号からは、発話とともに、笑い声と相槌を検出する。これらに加えて、視線・頷き・指示動作の情報を検出する。「正解」のコーパスを構築する際には、モーションキャプチャシステムや視線計測装置などの特殊な装置を使用するが、最終的なシステムは、カメラと遠隔マイクのみで実現することを想定している。

その上で、これらの情報を組み合わせることによって、理解度や興味度の推定を行う。動画投稿サイトなどの例からもわかるように、我々は他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、一方でアノテーションの基準や評価を含めて、これらを明確に定義するのは非常に困難である。そこで、これらに関係すると考えられ、より客観的に定式化できるマイルストーンをいくつか設定する。

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、興味度の推定に対して、笑い声と非語彙的な相槌を検出することでアプローチする。4 章と 5 章では、視線と頷きの情報も導入することで、聴衆の誰がいつどのような質問をするか予測するモデルの構築を試みる。これらの分析によって、ポスター発表が理解されたのか、興味を持たれたのかといった「会話理解」を行うための手がかりを得られることを期待している。6 章では、現在構築しているスマートポスターボードシステムの紹介を行う。

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

我々は、ポスター会話における音声・映像・動作・視線などのマルチモーダルな情報を収録するための環境 (IMADE ルーム) の構築を進めてきた [7], [8]。音声に関しては、各参加者に無線のヘッドセットマイクを装着してもらうとともに、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 個のカメラを部屋に設置した。この部屋には、モーションキャプチャシステムも設置されており、各参加者の位置と動きを把握するために使用した。各参加者には、このマーカに加えて、頭部に加速度センサと視線計測装置を装着してもらっている。このように多数の装置を装着してもらうのは参加者にとって負担となるが、大半の装置は帽子とウエストバッグに納めており、あまり気にならないように工夫している。(注1)実際のポスター会話の収録の様子を図 3 に示す。

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 3 ヶ年度にわたって合計 31 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。以降の章の分析では、アノテーションを整備した 4 つのセッションを用いている。これらのセッションにおける説明者と聴衆の組合せはすべて異なっている。各セッションにおいては、1 名の説明者 (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 分程度である。

ヘッドセットマイクで収録された音声データは、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、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衝突判定に基づいてアノテーションを行った。頷きについては、頭部の加速度センサのデータに基づいて自動アノテーションを行った。

3.　聴衆の相槌に基づく興味度の検出

聴衆は興味の度合いを聞き手行動によって表出していると考えられる。本章では、聞き手の相槌と笑いに着目する。日本語における相槌の典型的なものは「はい」であるが、「ふーん」や「へー」のように非語彙的で相槌にしか用いられないものも多い。本研究では主に後者に着目した。また、笑いと興味度の関係についても調べた。相槌と笑いの自動検出の方法と性能につ

(注1)：実際には視線計測装置等のキャリブレーションが大きな負担である。

図 3　ポスターセッションの収録の様子

いては、[9] [10] を参照されたい。

3.1　相槌の韻律パターンと興味度の関係

聴衆は興味の度合いを特定の相槌・韻律パターンで表現していると仮定し、分析を行った。韻律パターンはパラ言語情報や非言語情報において重要な役割を果たしていると考えられる。例えば、英語の “ok” などが応答であるのか相槌であるのかを韻律に基づいて判別する研究 [11], [12] も行われている。また、Ward [13] は、英語における非語彙的な相槌の韻律パターンとその意図・効用に関する分析を行っている。

本研究では、興味度と密接に関係する相槌・韻律パターンの同定を行った。そのために、「ふーん」「へー」「あー」の 3 種の相槌を選定した。これらはいずれも非語彙的でありコーパスにおける出現頻度も多く、また興味度と関係があると考えられたためである。

相槌の音声について、継続時間、F0(最大値とレンジ) 及びパワー (最大値) を求めた。これらの韻律特徴量は、話者毎に (平均を差し引いて分散で除することにより) 正規化している。

3 種の相槌に関して、4 種の韻律特徴量各々の大きいもの 10 個と小さいもの 10 個のサンプルを抽出した。各々について、相槌とその直前の発話区間の音声を切り出した。これを 5 名の被験者に聴取してもらい、聴衆の心的状態について、「強くそう思う」から「そう思わない」の 4 レベルで評定を行ってもらった。評定項目は 12 個用意し、その中に興味度に関する項目 (「興味」と「関心」)、及び驚きに関する項目 (「驚き」と「意外」) が 2 つずつ含まれている。2 つの評定項目の両方について、上位と下位で有意さ ($p < 0.05$) が確認されたものを表 1 に示す。「ふーん」が引き延ばされた場合に興味と驚きを表すこと、「あー」のピッチとパワーが強調された場合に興味を表すことがわかる。一方、「へー」については、いずれの韻律パターンによっても興味と驚きの両方が表される。

相槌の多くは明瞭でなく検出も容易でないが、興味度に関係するのはパワーが大きいものや継続時間の長いものであり、これらは検出も比較的容易であるので、会話のインデキシングにおいて有用であると期待される。

表 1　相槌の韻律パターンと興味・驚きとの関係

		興味	驚き
「ふーん」	継続時間	*	*
	F0 最大値		
	F0 レンジ		
	パワー最大値		
「へー」	継続時間	*	*
	F0 最大値	*	*
	F0 レンジ		*
	パワー最大値	*	*
「あー」	継続時間		
	F0 最大値	*	
	F0 レンジ		
	パワー最大値	*	

3.2　ホットスポットの第三者による主観評価

次に、相槌や笑いを引き起こした区間を「ホットスポット」と定義し(注2)、これらの区間が第三者の視聴者にとっても、興味深い／おもしろいと感じられるのか調べた。

ポスター会話に参加していない 4 名 (前述の 5 名とは別) の被験者に、当該区間の音声のみを聴取してもらい、以下の項目について主観評価を行ってもらった。

Q1:　当該の相槌／笑いが出現した理由がわかりましたか？
Q2:　当該区間を興味深い／おもしろいと思いますか？
Q3:　このポスター会話を聴取するに際して、当該区間は必要または有用と思いますか？

Q1 に対して「はい」と答えた割合は、笑いについて 89%、相槌について 95%に達し、ホットスポットの大半は妥当であることが確認された。

Q2 と Q3 はより主観的な評価であるが、ホットスポットの有用性に関するものである。笑いを伴う区間に関して、「おもしろい」と評価されたものは約半数のみであり、35%については「おもしろくない」と評価された。おもしろいと感じるかどうかは主観的要素が大きい上に、そもそもポスター会話におもしろい区間はあまり存在しないためと考えられる。

これに対して、相槌を伴う区間の 90%以上について、「興味深い」(Q2)、「有用」または「必要」(Q3) という評価が得られた。この結果は、相槌の検出に基づくホットスポットの有用性を示すものである。

4.　視線と相槌の情報に基づく聴衆の発話の予測

多人数における会話において話者交替は自明でなく、ターン (発話権) が誰に譲られ、取得されるかをモデル化・予測することは、会話分析において重要なテーマとなっている [16]。また、多人数を会話相手とする音声対話エージェント・ロボット [1], [2] や、発話者に対してカメラのズームやマイクのビームフォームを制御するようなシステムの研究開発においても重要である。対話における話者交替に関する研究はこれまで数多くされているものの、多人数会話において話者交替を予測する工学的なモデル化はきわめて少ない [17], [18]。ポスターセッションにおけ

(注2)：ミーティングで複数の参加者が白熱した議論を行っている区間を「ホットスポット」と定義している先行研究 [14], [15] もある。

表 2　視線の継続時間 (秒) と発話交替の関係

	発表者 A による	聴衆による発話	
	ターンの保持	B	C
発表者 A が聴衆 B を注視	0.220	**0.589**	0.299
発表者 A が聴衆 C を注視	0.387	0.391	**0.791**
聴衆 B が発表者 A を注視	0.161	0.205	0.078
聴衆 C が発表者 A を注視	0.308	0.215	0.355

表 3　発表者と聴衆の視線状態の定義

視線配布者	発表者		
聴衆	視線対象	聴衆 **(I)**	ポスター **(P)**
	発表者 **(i)**	**Ii**	**Pi**
	ポスター **(p)**	**Ip**	**Pp**

表 4　発表者と聴衆の視線状態と発話交替の関係

	発表者による	聴衆による発話		合計
	ターンの保持	(self)	(other)	
Ii	125	17	3	145
Ip	320	**71**	26	417
Pi	190	11	9	210
Pp	2974	147	145	3266

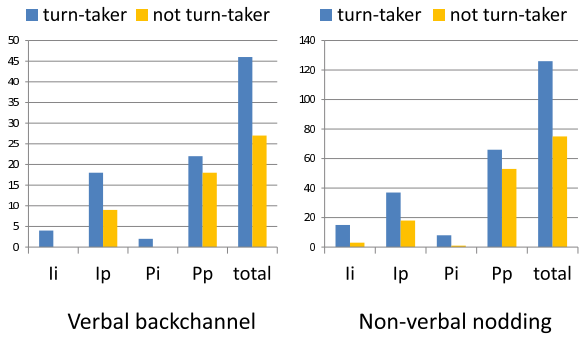

図 4　相槌・頷きと視線状態及び発話交替の関係

る会話は、先行研究で取り組まれたミーティングや自由会話などと異なり、発表者が大半の時間でターンを保持し、発話数も圧倒的に多い。しかしながら、聴衆の質問やコメントの発話の方がより重要と考えられる。したがって本研究では、聴衆によるターンの取得を予測する問題を設定する。具体的には、いつ聴衆の誰が発話するかの予測を行う。一般に、聴衆は当該発表に興味を持った時に質問やコメントを行うと考えられるので、聴衆による発話は興味度と関係があると期待される。

発話交替において視線配布が重要な役割を果たしていることが知られている [1], [18]～[20]。しかしながらポスター会話では、参加者の視線の大半はポスターに注がれる。地図やコンピュータなどを用いた会話でも同様と考えられる。そこで本研究では、相槌などの情報の利用も考える。

4.1　視線対象と発話交替の関係

発表者の発話 (IPU) 終了時におけるすべての参加者の視線対象を同定した。視線対象は、ポスターか他の参加者 (もしくは「なし」) のいずれかである。発表者の発話終了直前の 2.5 秒間における視線の継続時間も計測した。2.5 秒間に設定したのは大半の発話が 2.5 秒以下であるためである。発話交替との関係を表 2 に示す。発表者が聴衆にターンを譲る場合には、他の場合に比べて、その聴衆の方をより長く見ていることがわかる。しかしながら、聴衆の視線配布については、発話交替との明確な関係は見られない。

4.2　両者の視線状態と発話交替の関係

次に、発表者と聴衆の両者の視線配布に基づく視線状態を表 3 のように定義する。この表では「聴衆」と表記しているが、実際にはこれらの状態は聴衆の各人について定義される。例えば、“Ii” は発表者と聴衆の 1 人との相互注視である。また、“Pp” はポスターに対する共同注意に相当する。

発表者の発話末におけるこれらの視線状態の頻度と発話交替の関係を表 4 に示す。ここでは、各々の頻度は 2 名の聴衆に対して合計をとったものとなっており、発話交替に応じて分類している。発話交替は、当該視線状態に関係した人がターンを取得した場合 (self) とそうでない場合 (other) に分類している。相互注視の状態 “Ii” が発話交替に最も関係が深いと考えられたが、その頻度は多くない。聴衆が発表者を見ている状態 “Pi” の頻度も多くない。最も有用な情報と考えられるのは、発表者が聴衆を見ている状態 “Ip” であり、この場合にその聴衆にターンが譲られる場合が多い。この結果は前節の分析とも符合する。

4.3　相槌と発話交替の関係

3 章では、特定の相槌のパターンが聴衆の興味度と関係のあることを示した。頷きは、相槌の非言語的なものと捉えられ、2 人対話と比べてポスター会話ではより頻繁に観測される。

発表者の発話終了直前の 2.5 秒間における相槌と頷きの頻度と発話交替の関係を図 4 に示す。前節で定義した視線状態との関係も示している。聴衆が発話しようとする際 (turn-taker) には相槌・頷きとも多くなっている傾向が見られ、特に “Ii” や “Ip” といった発話交替との関係が深い視線状態においてより顕著である。

4.4　聴衆からの発話の予測

前節までの分析に基づいて、聴衆によるターンの取得を予測することができるか実験を行った。これは、話者交替の予測と次話者の予測の 2 つのタスクに分割して行った。話者交替の予測タスクでは、ターンが発表者から聴衆 (の誰か) に譲られるかの予測を行い、次話者の予測タスクにおいて聴衆の誰が実際に発話するのか予測する。これらの予測は、発表者の各発話 (IPU) の終了時、すなわち話者交替・次話者の発話が行われる前に、それまでの情報を用いて行う。

話者交替の予測タスクでは、発表者の発話の韻律特徴をベースラインとして用いた。具体的には、予測時点の直前の発表者の発話の F0(平均・最大値・最小値・レンジ) とパワー (平均・最大値) を計算した。相槌と頷きについては、予測時点の直前の各々の頻度を求めた。視線配布に関する特徴は、視線対象と視線状態の両方に関して、出現頻度と継続時間の両方で求めることができる。ただし実際には、これらの 4 通りの組合せの間で予測性能に有意な差は見られず、併用する効果もなかった。

表 5　話者交替の予測結果

特徴	再現率	適合率	F 値
韻律	0.667	0.178	0.280
相槌	0.459	0.113	0.179
視線	0.461	0.216	0.290
韻律+相槌	0.668	0.165	0.263
韻律+視線	**0.706**	0.209	0.319
韻律+相槌+視線	0.678	0.189	0.294

表 6　次話者の予測結果

特徴	予測精度
視線	66.4%
相槌	52.6%
視線+相槌	**69.7%**

話者交替を予測するモデルとして SVM を用い、4 セッションでクロスバリデーションによる学習・評価を行った結果を表5に示す。話者交替、すなわち聴衆によるターンの取得に関する再現率 (recall)・適合率 (precision)・F 値 (F-measure) を求めている。発表者の発話の大半は継続され、話者交替が発生する割合は 11.9%に過ぎないので、その予測はかなり困難である。本研究では、聴衆の発話を確実に捉えることが重要であると考えて、話者交替の再現率を最も重視する。

用いた各特徴を比較すると、韻律特徴が最も高い再現率を得ており、視線特徴が適合率と F 値で最高になっている。これらの 2 つの特徴を組み合わせることにより、再現率・適合率ともに改善することができた。これに対して、相槌・頷きの特徴は最も予測性能が低く、他の特徴と組み合わせてもかえって性能の低下を引き起こす結果となった。

次に、発話交替が起こった場合に、次話者の予測を行った。多人数会話において、(実際に次の話者が発話する前に) 次話者を予測することは非常に困難なタスクであり、これまでの研究例も皆無に近い。この場合、発表者 (現発話者) の韻律情報には次話者を示唆する情報は含まれないので、使用することができない。したがって、相槌・頷きと視線配布の特徴を用いた。これらの特徴は、話者交替の予測では聴衆全体について求めていたが、次話者の予測では聴衆の各人毎に求める。

予測結果を表 6 に示す。今回は、相槌・頷きの特徴を用いる効果が見られ、視線配布の特徴と組合わせることにより、約 70%の精度が得られた。

5. 聴衆の聞き手行動と質問の種類の関係

次に、聴衆の聞き手行動とその後にどのような種類の質問を行うのかとの関係を調べた。本研究では、質問の種類を確認質問と踏み込み質問の 2 つに分類した。確認質問は、発表者の説明に対する理解を確認するためのもので、「はい/いいえ」のいずれかで答えることができる(注3)。これに対して踏み込み質問は、発表者の説明に含まれなかったことに関して質問を行うもので、通常「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となる。

(注3)：ただし、発表者が実際に「はい／いいえ」のみで答えたとは限らない。

表 7　相槌の頻度 (秒当り) と質問の種類との関係

	確認	踏み込み
質問者	0.034	**0.063**
非質問者	0.041	0.038

表 8　頷きの頻度 (秒当り) と質問の種類との関係

	確認	踏み込み
質問者	0.111	0.127
非質問者	0.109	0.132

表 9　視線状態の継続時間 (割合) と質問の種類の関係

	確認	踏み込み
Ii	0.053	0.015
Ip	**0.116**	0.081
Pi	0.060	0.035
Pp	0.657	**0.818**

この質問の分類と対応する発表者の説明区間のアノテーションは、発表者がポスターに沿って説明している間は比較的容易であるが、発表を一通り終えて全体的な質疑に入ってしまうと非常に困難になる。そこで、質疑に入った後は除外して、説明者の発表中に行われた質問についてのみアノテーションを行った。また、視線や相槌などの聞き手行動について、前章では発話 (IPU) 毎に特徴を求めていたが、ここでは説明者の複数の発話をまとめた説明区間を人手で定義して求めた。これらの点が前章の分析との違いである。

5.1　相槌・頷きと質問の種類の関係

相槌と頷きについて、説明区間の長さ (秒) で正規化した出現頻度と質問の種類との関係を表 7 と表 8 に示す。実際に質問を行った参加者とそうでない参加者の比較も行っている。踏み込み質問を行う際には、相槌の頻度が有意に多くなっていることがわかる。しかしながら、頷きに関しては特徴的な傾向が見られなかった。

5.2　視線配布と質問の種類の関係

視線配布と質問の種類との関係についても調べた。前章で述べた様々な特徴量の中で、視線状態の継続時間が最も有用であった。説明区間の継続時間で正規化した割合と質問の種類との関係を表 9 に示す。確認質問を行う際には、“Ip” の時間が長い (“Pp” の時間が短い) ことがわかる。前章の分析と総合すると、発表者からの視線配布に伴ってターンが譲られた場合の大半は、聴衆の理解を確認するためのものであると推察される。

6. スマートポスターボード

我々は、JST CREST「マルチモーダルな場の認識に基づくセミナー・会議の多層的支援環境」プロジェクトで、「スマートポスターボード」の設計と実装を進めている。これは、ポスターセッションを収録し、参加者のふるまいをセンシングし、インタラクションのアノテーションを行うものである。2 章で述べたコーパス収録時のように、参加者にヘッドセットマイク

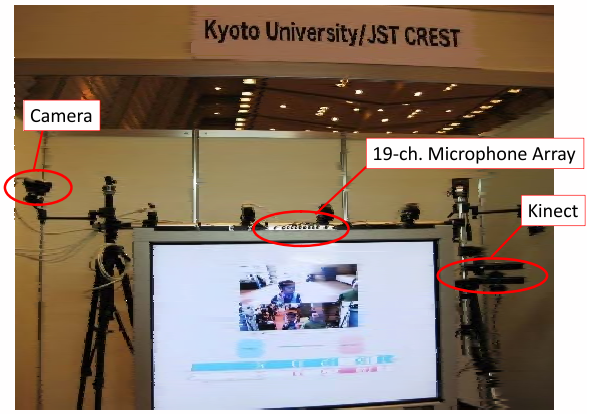

図 5　スマートポスターボードの概観

や視線計測装置などの特殊な装置を装着してもらうのは現実的でなく、また部屋に装置を設置するのも可搬性の点で問題となるので、すべてのセンサはポスターボード (65 インチ液晶ディスプレイ) に装着することにした。その概観を図 5 に示す。

ディスプレイの上部に 19 チャネルのマイクロフォンアレイを設置し、併設する三脚にカメラ 6 台と Kinect センサを設置している。音声の分離と強調は、遅延加算 (Delay-and-Sum) 型ビームフォーマと ICA に基づく雑音推定・抑圧を組み合わせたブラインド空間的サブトラクションアレイ (BSSA) [21] によって実現する。この処理によって、音声入力が発表者と聴衆のものに分離される。ただし、聴衆間の分離は行われない。分離された各音声チャネルに対して音声区間検出を行う。6 台のカメラから得られる画像情報を用いて、聴衆の数と位置、そして頭部方向 (視線の代用) を推定する。この情報を、前記の音声分離と話者特定の処理に利用することも予定している。

前章までに述べた理解度や興味度などのアノテーションは現段階で実装されていないが、上記の処理によって、インタラクションを可視化するポスター会話のブラウザを実現できる。

Kinect センサは、より手軽でオンラインのアプリケーション向けのものであり、話者の同定と頭部方向の推定はリアルタイムで行うことができる。

図 5 に示すように、このシステムのデモ展示を IEEE-ICASSP 2012 で行い、その後も引き続き改善を進めている。

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。本稿で紹介した研究 (特に 3 章〜5 章) は、瀬戸口久雄、常志強、土屋貴則、岩立卓真、高梨克也各氏の貢献によるものである。スマートポスターボード (6 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と奈良先端科学技術大学院大学の多くの方によるものである。

文　　献

[1] D.Bohus and E.Horvitz. Models for multiparty engagement in open-world dialog. In *Proc. SIGdial*, 2009.

[2] S.Fujie, Y.Matsuyama, H.Taniyama, and T.Kobayashi. Conversation robot participating in and activating a group communication. In *Proc. INTERSPEECH*, pp. 264–267, 2009.

[3] K.Ohtsuka. Conversation scene analysis. *Signal Processing Magazine*, Vol. 28, No. 4, pp. 127–131, 2011.

[4] C.Oertel, S.Scherer, and N.Campbell. On the use of multimodal cues for the prediction of degrees of involvement in spontaneous conversation. In *Proc. INTERSPEECH*, pp. 1541–1545, 2011.

[5] S.Renals, T.Hain, and H.Bourlard. Recognition and understanding of meetings: The AMI and AMIDA projects. In *Proc. IEEE Workshop Automatic Speech Recognition & Understanding*, 2007.

[6] T.Kawahara. Multi-modal sensing and analysis of poster conversations toward smart posterboard. In *Proc. SIGdial Meeting Discourse & Dialogue*, (keynote speech), 2012.

[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] T.Kawahara, K.Sumi, 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.

[10] 河原達也, 須見康平, 緒方淳, 後藤真孝. 音声会話コンテンツにおける聴衆の反応に基づく音響イベントとホットスポットの検出. 情報処理学会論文誌, Vol. 52, No. 12, pp. 3363–3373, 2011.

[11] F.Yang, G.Tur, and E.Shriberg. Exploiting dialog act tagging and prosodic information for action item identification. In *Proc. IEEE-ICASSP*, pp. 4941–4944, 2008.

[12] A.Gravano, S.Benus, J.Hirschberg, S.Mitchell, and I.Vovsha. Classification of discourse functions of affirmative words in spoken dialogue. In *Proc. INTERSPEECH*, pp. 1613–1616, 2007.

[13] N.Ward. Pragmatic functions of prosodic features in non-lexical utterances. In *Speech Prosody*, pp. 325–328, 2004.

[14] B.Wrede and E.Shriberg. Spotting "hot spots" in meetings: Human judgments and prosodic cues. In *Proc. EUROSPEECH*, pp. 2805–2808, 2003.

[15] D.Gatica-Perez, I.McCowan, D.Zhang, and S.Bengio. Detecting group interest-level in meetings. In *Proc. IEEE-ICASSP*, Vol. 1, pp. 489–492, 2005.

[16] 坊野真弓, 高梨克也（編）. 多人数インタラクションの分析手法. オーム社, 2009.

[17] K.Laskowski, J.Edlund, and M.Heldner. A single-port non-parametric model of turn-taking in multi-party conversation. In *Proc. IEEE-ICASSP*, pp. 5600–5603, 2011.

[18] K.Jokinen, K.Harada, M.Nishida, and S.Yamamoto. Turn-alignment using eye-gaze and speech in conversational interaction. In *Proc. INTERSPEECH*, pp. 2018–2021, 2011.

[19] A.Kendon. Some functions of gaze direction in social interaction. *Acta Psychologica*, Vol. 26, pp. 22–63, 1967.

[20] B.Xiao, V.Rozgic, A.Katsamanis, B.R.Baucom, P.G.Georgiou, and S.Narayanan. Acoustic and visual cues of turn-taking dynamics in dyadic interactions. In *Proc. INTERSPEECH*, pp. 2441–2444, 2011.

[21] 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.