Vol.2013-SLP-97 No.12
2013/7/27

ポスター会話における聴衆のマルチモーダルな振る舞いに基づく興味・理解度の推定

河原 達也[1,a)]　林 宗一郎[1]　高梨 克也[1]

概要：ポスターセッションにおける聴衆の興味・理解度を自動推定することを試みる。講演形式の発表に比べて、ポスター発表では、聴衆の視線や相槌などの振る舞いが顕著に見られる。これらの振る舞いは、興味・理解度と関係があると考えられる。また興味・理解度は、聴衆の質問や相槌などの発話行為からも推測できると考えられる。本研究ではまず、興味・理解度と発話行為の関係を分析する。次に、発話行為と聴衆の振る舞いとの関係を調べる。これに基いて、話題セグメント毎にマルチモーダルな振る舞いから、質問の生起とその種類の予測を行う。実験の結果、相槌と視線の特徴量が予測に有効であることと、それらを組み合わせることの相乗効果が確認された。

キーワード：マルチモーダルインタラクション、会話分析、視線、相槌

Estimation of Interest and Comprehension Level of Audience through Multi-modal Behaviors in Poster Conversations

Tatsuya Kawahara[1,a)]　Soichiro Hayashi[1]　Katsuya Takanashi[1]

Abstract: We address the estimation of the interest and comprehension level of an audience in poster sessions. Compared to lecture presentations, the audience's behaviors such as gazing and backchannels are more observable in poster presentations. These multi-modal behaviors are presumably related with their interest and comprehension level. We also assume that the interest and comprehension level can be judged by particular speech acts of the audience such as questions and reactive tokens. First, we make a preliminary analysis on their correlation. Next, we investigate the relationship between the audience's behaviors and the question type. Then, we conduct prediction of questions and their type based on the multi-modal behaviors during the relevant topic segment. Experimental results show that verbal backchannels and eye-gaze patterns are good predictors to this task, and also the combination of the multi-modal features is effective.

Keywords: multi-modal interaction, conversational analysis, eye-gaze, backchannel

[1] 京都大学
Kyoto University, Kyoto 606-8501, Japan
[a)] http://www.ar.media.kyoto-u.ac.jp/crest/

1. はじめに

人間どうしのコミュニケーションは本質的に双方向で全二重であり、聞き手のフィードバック行動が円滑なコミュニケーションに重要な役割を果たしている [1]。発表形式の会話を分析する際にも、聴衆のフィードバック行動は重要な手がかりとなる。聴衆がその発表に引きつけられているかは、フィードバック行動を見ることで推測できる。このような特性は、聴衆の数が少ない場合に一層顕著になる。聴衆は、頷きのような非言語的なフィードバックだけでなく、言語的な相槌を行うようになる。また、視線の振る舞いがより明確になり、聴衆による発話権の取得において重要な役割を果たす。

我々は、研究者が数人の聴衆に対して行うポスター発表における会話（=ポスター会話）のマルチモーダルな収録と分析を行っている [2], [3]。このようなポスター発表は、学会やオープンハウスなどで一般的に行われている。聴衆は発表の最中でも質問をすることができ、聴衆の反応、特に質問やコメントの頻度や中身をみることで、その発表が理解されているか気に入っているか推測することもできる。

これまでに我々は、「へー」や「あー」などの特定の音韻・韻律パターンの相槌が聴衆の興味と深く関係していることを明らかにした [4], [5], [6]。また、聴衆の発話権の取得と相槌や視線などのフィードバック行動との関係を分析し、聴衆の発話をある程度予測できることを示した [7], [8]。発話権の取得と非言語行動との関係は他の研究でも分析が行われている [9], [10], [11]。

本研究では、このような聴衆のマルチモーダルな振る舞いに基づいて、興味・理解度の推定を行うことを試みる。興味・理解度のアノテーションは容易でなく、主観的になりがちであるので、これらの心的状態と関係があると考えられ、客観的に観測できる発話行為に着目する。具体的には、聴衆による質問と特定のパターンの相槌に着目する。さらに、質問を確認質問と踏み込み質問に分類する。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。

以下、2 章と 3 章でマルチモーダルなコーパスと問題設定について述べる。4 章では、聴衆のマルチモーダルな振る舞いと着目している発話行為の関係を分析する。5 章では、相槌や視線の振る舞いから発話行為を予測する実験について述べる。

2. ポスター会話のマルチモーダルコーパス

我々はこれまでに、数多くのポスター会話の収録を行ってきた [12], [13]。本研究ではそのうちの 10 セッションを用いる。各セッションでは、1 名の発表者（A と表記）が 2 名（B,C と表記）の聴衆に対して自分の研究内容に関する発表を行っている。発表者と聴衆は大学院生もしくは若手研究者である。発表者と聴衆は別の研究室から選んでおり、互いの面識や研究内容の事前知識はほとんどない設定になっている。同一の発表者が

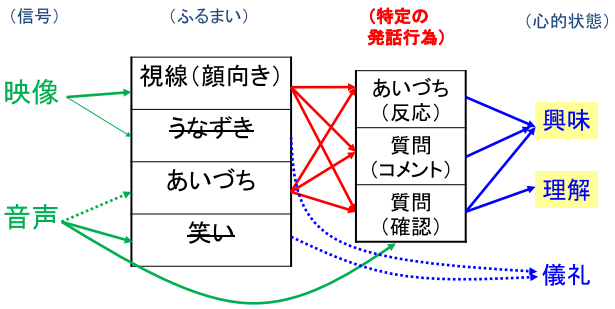

図 1　発話行為を介した興味・理解度推定の枠組み

2 つのセッションで発表を行った場合もあるが、その場合聴衆は異なっている。各セッションの長さはおおむね 20〜30 分である。

音声データは発話者毎に、ポーズで区切られた発話単位 (IPU) 及び文単位に分割し、書き起こしを行った。書き起こし基準は『日本語話し言葉コーパス』(CSJ) に準拠しているが、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ている。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。ただし一部のセッションでは、画像情報から自動で推定したものを人手で修正している。

各ポスターは、発表者の研究内容を他の分野の研究者や学生に紹介するように作成されており、比較的独立した 4 ないし 8 個の要素（=スライド話題単位）から構成されている。これは学会のポスターセッションで用いられるものとは若干異なるが、スライド話題単位毎に聴衆の興味・理解度の評価を行いやすいようになっている。ポスター会話は、通常スライド話題単位を一つずつ説明した後に、全体的な質疑や議論を行うことが多い。質疑・議論の段階では、どの話題について言及しているか特定するのが容易でないので、各スライド話題単位の説明セグメントのみを分析の対象とする。

本研究で用いた 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。

3. 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う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 [4], [5]。Ward ら [14] 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。

3.1 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。[*1] 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。[*2] 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。

3.2 質問の種類と興味・理解度との関係

直近（2012 年 12 月）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 2 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 [5] 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。

[*1] Strömbergsson ら [15] も同様の分類を行っているが、「後向き質問」と「前向き質問」と定義している。
[*2] ただし、実際に発表者が「はい／いいえ」のみで答えたとは限らない。

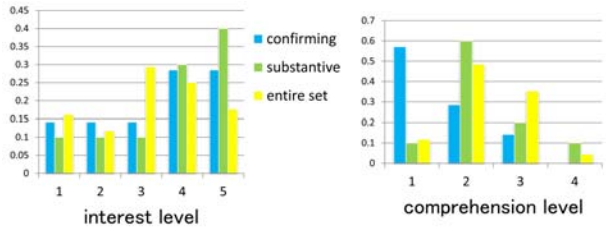

図 2 質問の種類毎の理解・興味度の分布

4. フィードバック行動と質問の関係

次に、聴衆の相槌や視線の振る舞いと質問の生起・種類の関係を調べた。

4.1 相槌

ポスター会話における相槌のうち、直接的に興味を表す顕著な相槌は 20%以下で、大半は「はい」などの聴衆が「聞いている」ということを示唆するものである。この種の相槌についても、著者らはその頻度が興味と一定の関係があることを示している [16]。すなわち聞き手は、会話に興味があるときにより多くの相槌を打つ傾向にある。

頷きは非言語的な相槌と考えられ、日常会話に比べてポスター会話ではより頻繁に見られる。しかしながら、頷きの多くは微妙な動作であるので、安定した検出・アノテーションが困難である。また、予備的な分析では、頷きの出現傾向に関して明確な傾向は見られなかった。そこで本研究では、言語的な相槌のみに絞って分析を行う。

相槌の出現回数を発表者の発話で正規化した頻度の平均値を、話題セグメント毎に求めた。これを質問の生起・種類に応じてまとめたものを表 1 に示す。半数以上の話題セグメントで質問が生起していないので、全セグメントにおける平均値は、質問が生起した場合に比べて小さくなっている。この結果、質問を行う時、特に踏み込み質問を行う時に、聴衆の相槌が増える傾向がわかる。

4.2 発表者に対する視線配布

話題セグメント毎に聴衆の各人の視線の対象と継続時間を求めた。視線の対象はポスターか他の会話参加者になるが、ポスター会話ではほとんどの時間ポスターを注視しているのが普通である。これは逆に、他の会話参加者に視線を配布することは、何らかの理由と効果があると考えられる。我々はこれまでに、視線情報が発話権の取得と関係があること、特に発表者の視線が発話権を管理していることを示している [8]。

本研究では、聴衆の視線に注目し、質問の種類との関係について分析する。そこで、聴衆が発表者に視線配布する回数と時間を求めた。具体的には、発表者の発話で正規化した頻度（回／発話）と継続時間の割合を求めた。その結果を表 2 に示す。確認質問を行う時は視線配布が少なくなり、踏み込み質問を行

表 1　聴衆の相槌の頻度（回/発話）と質問の生起・種類の関係

	確認質問	踏み込み質問	全セグメント
相槌の頻度	0.53	0.59	0.42

表 2　聴衆の発表者への視線（頻度，時間）と質問の生起・種類の関係

	確認質問	踏み込み質問	全セグメント
視線配布の頻度	**0.38**	**1.02**	0.64
視線の時間割合	0.05	**0.15**	0.07

う時は多くなる傾向がわかる。確認質問を行う際には、聴衆は内容を理解しようとポスターに集中しており、踏み込み質問を行う際には、発話権を取得するために発表者の気を引こうとしていることが推察される。

より詳細に分析すると、踏み込み質問を行う直前には、視線配布が徐々に増えていく傾向があるが、確認質問の際にはそのような傾向は見られなかった。

この結果は、質問の種類、さらには興味・理解度を予測する上で、視線情報が有用であることを示唆するものである。

5. 質問・顕著な相槌の生起の予測：興味度の推定

前章の分析に基づいて、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。3 章で述べたように、これは聴衆が質問ないし顕著な相槌を生成するかを予測する問題に帰着させる。すなわち、当該の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。

まず、各聴衆の振る舞いを特徴量にする必要があるが、前章で述べた特徴を用いる。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 $F = \{f_1, \ldots, f_d\}$ 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、$p(f_i|c)$ を計算するのには、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。

評価実験は、leave-one-out クロスバリデーションにより行った。種々の特徴量に対する結果を表 3 に示す。F 値は、「興味を持った」話題セグメントの再現率と適合率の調和平均である。ただし本実験では、再現率と適合率はほぼ同じ値になっている。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。

表 3　質問・顕著な相槌を含む話題セグメントの予測結果

	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 4　確認質問／踏み込み質問の同定結果

	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%

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。

6. 質問の種類の同定：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。3 章で述べたように、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。

特徴量と識別の方法は前章と同一である。確認質問／踏み込み質問の分類結果を表 4 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。これは、表 2 で示されている特徴量の差からもわかる。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。

相槌については単純な出現頻度ではあまり効果が見られなかったので、音韻や韻律のパターンを利用する必要があると考えられる。また、Strömbergsson ら [15] は、質問のピッチパターンに着目した分析を行っているので、この利用も検討する。

7. おわりに

本研究では、質問や顕著な相槌といった聴衆の発話行為が興味・理解度と関係があると仮定し、これらの発話行為とフィードバック行動との関係に着目した分析を行った。具体的には、興味度の推定を質問・顕著な相槌の生起を予測する問題に、理解度の推定を質問の種類を同定する問題に、それぞれ帰着させた。

まず、これらの問題設定の妥当性をアンケート調査によって確認し、それに基づいて話題セグメントのアノテーションを行った。次に、相槌や発表者への視線配布などのフィードバック行動の分析を行い、質問の生起や種類との関係を調べた。その上で、これらの発話行為を介して興味・理解度を推定する実験を行った。

マルチモーダルな特徴量を組み合わせることで、興味度の推定精度がベースラインの 50%から 70%まで向上することを示した。理解が困難であることを示唆する確認質問の同定は 75%の精度でできることを示した。理解ができない場合は聴衆はそもそも質問しないことも考えられるが、発表に興味を持ったにも関わらず理解が困難な箇所を特定することが特に重要と考えている。

我々は、カメラやマイクロフォンアレイを用いてポスター会話を収録する「スマートポスターボード」[2] の研究開発を行っているが、本稿で述べた成果は聴衆の反応のアノテーションを行い、収録した会話を検索する上で有用であると期待される。

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。

参考文献

[1] N.Ward, D.Novick, L.P.Morency, T.Kawahara, D.Heylen, and J.Edlund, editors. *Proc. Interdisciplinary Workshop on Feedback Behaviors in Dialog*, 2012.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 岩立卓真, 高梨克也, 河原達也. ポスター会話におけるパラ言語・非言語情報を用いた話者交替及び次話者の予測. 人工知能学会研究会資料, SLUD-B103-10, 2012.

[8] T.Kawahara, T.Iwatate, and K.Takanashi. Prediction of turn-taking by combining prosodic and eye-gaze information in poster conversations. In *Proc. INTERSPEECH*, 2012.

[9] N.G.Ward and Y.A.Bayyari. A case study in the identification of prosodic cues to turn-taking: Back-channeling in Arabic. In *Proc. INTERSPEECH*, pp. 2018–2021, 2006.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] S.Strombergsson, J.Edlund, and D.House. Prosodic measurements and question types in the spontal corpus of Swedish dialogues. In *Proc. INTERSPEECH*, 2012.

[16] T.Kawahara, M.Toyokura, T.Misu, and C.Hori. Detection of feeling through back-channels in spoken dialogue. In *Proc. INTERSPEECH*, p. 1696, 2008.