6P-09

スマートポスターボードにおける視線情報を用いた話者区間検出及び相槌の同定

井上 昂治 [1]　若林 佑幸 [2]　吉本 廣雅 [3]　高梨 克也 [3]　河原 達也 [3]

1. 京都大学 大学院情報学研究科　2. 立命館大学 大学院情報理工学研究科　3. 京都大学 学術情報メディアセンター

1. はじめに

我々はポスターセッションにおける会話 (=ポスター会話) を対象として，マルチモーダルな分析環境であるスマートポスターボードの構築を進めている [1]．本稿では，ポスター会話において各々の参与者がいつ発話したかという情報（話者区間）とそのうちの相槌を同定する手法を述べる．これらの情報はポスター会話を後から参照するためのインデクスとして有用である．しかしながら，実際の会話では，周囲の雑音や自然な話し言葉などにより検出精度が低下する．そこで，従来用いられてきた音響情報に加えて，会話の発話権交替において重要な役割を担っている視線情報を統合することで，検出精度及び頑健性の向上を図る．これまでに話者区間検出の枠組み [2] を提案したが，本稿ではその改善と，相槌の同定手法について述べる．

2. 音響情報に基づく話者区間検出

音響情報のみを用いて話者区間検出を行う場合，MFCC や音声到来方向に基づく手法 [3] がある．ここでは，スマートポスターボードに搭載されている 19 チャネルマクロホンアレイを用いて音声到来方向を推定し，その方向に対応する参与者の発話とみなす手法をベースラインとする．音声到来方向推定には MUSIC[4] を用いる．この手法では観測信号の部分空間の直交性に基づいて MUSIC スペクトルを算出する．MUSIC スペクトルの大きさはその角度に位置する参与者が発話したかを示す手がかりとなる．

3. 視線情報の利用

実際のポスター会話では，周囲の雑音や自然な話し言葉などによる音響的影響を受けるため，音響情報のみに基づく手法では十分な検出精度が得られない．そこで音響情報だけでなく，会話参与者の視線情報も用いて話者区間を検出する．視線情報は画像処理により算出されるため，音響的影響を受けず，頑健な話者区間検出が期待できる．ここではスマートポスタードに搭載されている 2 台の Kinect から得られるカラー及び深度画像から各参与者の頭部位置と頭部方向を推定し [5]，頭部方向を視線として代用する．

3.1 視線情報に基づく発話の予測

多人数会話における視線のふるまいは，発話権取得と相関があることが知られている [6]．例えば，現話者から次話者へ発話権が移行する場面では，現話者は発話を終了する直前に次話者へ視線を向け，次話者は発話権を取得するために現話者へ視線を向ける傾向がある．この知見に基づいて，ポスター会話における視線のふるまいから各参与者の発話を予測できる [7, 8]．

ここでは視線情報として，その時点で各参与者が何 (他の参与者/ポスター) を見ているか (視線配布)，各参与者間での視線配布の組合せ (視線状態 [7]) を考慮する．視線状態は，共同注意 (互いにポスターを見ている) や相互注視 (互いを見ている) を表すことができる．

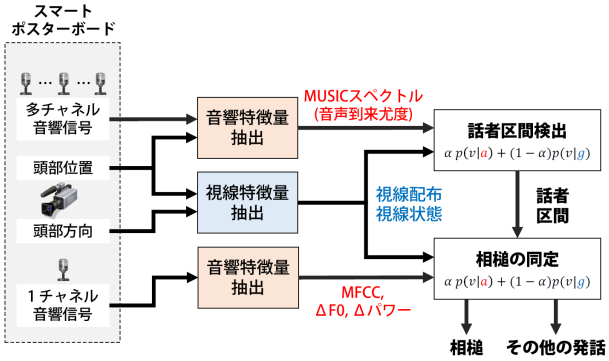

図 1: 提案手法の処理の流れ

3.2 音響情報と視線情報の統合法

音響情報と視線情報の統合処理の流れを図 1 に示す．話者区間検出 (図 1 上部) では，マイクロフォンアレイによる多チャネル音響信号と各参与者の頭部位置から音響特徴量を，各参与者の頭部位置と頭部方向から視線特徴量を算出する．続いて，これらの特徴量を確率的に統合し，各参与者の発話の有無をフレーム単位で推定する．

音響特徴量は，各参与者の頭部位置から ±10° の範囲の MUSIC スペクトルの値とする．視線特徴量は，1) その時点での視線配布と視線状態の生起，2) その時点から過去 1000msec の範囲での各視線配布と各視線状態の最大持続フレーム数及び遷移頻度 (1-gram，2-gram) とした．

抽出した特徴量を確率変数とみなし (音響は a，視線は g)，確率モデルにより発話イベント v(発話または非発話) の事後確率を推定する．確率モデルとして事後確率の結果統合を用いる．

$$\alpha\, p(v|a) + (1-\alpha)\, p(v|g) \qquad (1)$$

重み係数 $\alpha \in [0,1]$ は，環境の違いによるエントロピーの変化を利用した推定手法 [9] によりオンラインで決定する．各識別モデルはロジスティック回帰で推定する．

4. 相槌の同定

前節の手法により検出した各話者区間に対して相槌を同定する (図 1 下部)．相槌は聞き手が発する短い発話であり，発話権を取得せずに話し手の発話継続を促す役割がある．したがって，相槌も視線のふるまいと

Speaker diarization and detection of backchannels using eye-gaze information for smart posterboard: *Koji Inoue (Kyoto Univ.), Yukoh Wakabayashi (Ritsumeikan Univ.), Hiromasa Yoshimoto, Katsuya Takanashi, and Tatsuya Kawahara (Kyoto Univ.)*

表 1: 話者区間検出精度 (DER[%])

手法	SNR [dB]				平均
	∞	10	5	0	
音響	6.16	14.21	22.94	35.89	19.80
音響+視線	6.27	13.69	18.18	21.61	14.94

表 2: 相槌の同定精度 (F 値 [%])

手法	SNR [dB]				平均
	∞	10	5	0	
区間長	66.53	35.28	21.85	11.95	33.90
音響	78.22	42.15	25.14	12.50	39.50
音響+視線	78.58	44.64	27.41	13.10	40.93
#相槌区間	605	189	93	25	-

相関があると考えられ，視線情報を用いることで相槌の同定精度向上が期待される．

音響情報と視線情報を統合する枠組みは話者区間検出と同じである．音響特徴量は，1) 発話区間の長さ，2)MFCC，3) 先行発話末の F0 とパワーの傾き [10] とする．視線特徴量と確率モデルは話者区間検出と同様である．

5. 評価実験

スマートポスターボードを用いて収録したポスター会話 8 セッションによる交差検定で評価した．各セッションは，説明者 1 人，聴衆 2 人で構成され，時間は 20〜30 分である．雑音の影響を評価するために，評価用セッションの音響信号に拡散性の人混み雑音を重畳した．

話者区間検出は以下のダイアライゼーション誤り率 (DER)[11] で評価した．

$$\mathrm{DER} = \frac{\text{誤受理}\cdot\text{誤棄却}\cdot\text{話者誤りの区間長}}{\text{全発話区間長}} \times 100[\%]$$

各手法の閾値を，交差検定の 8 セッションで同時に変化させ，得られた DER の最小値で評価した．比較手法は，式 (1) において音響情報のみを用いる識別モデル $p(v|a)$(音響) である．表 1 より，信号対雑音比 (SNR) が小さくなるにつれて，提案手法 (音響+視線) により DER が小さくなった．したがって，雑音環境下において視線情報は有効であるといえる．

続いて，提案手法によって検出した話者区間 (聴衆のみの DER が最小時) のうち聴衆の発話に対して相槌の同定を行った．ただし，前処理として，発話の持続長が 2000msec 以上，または当該発話と先行発話の話者が同一のものは相槌でないとして除外した．評価指標は，相槌の発話区間に対する再現率 R と適合率 P を基に，以下の F 値 F を求めた．

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

ただし，本研究は適合率を重視して，$\beta = 0.5$ とした．各手法の閾値を，交差検定の 8 セッションで同時に変化させ，得られた最大 F 値で評価した．また，各 SNR で得られる推定話者区間が異なるため，推定話者区間に含まれる正解相槌区間のみを用いて再現率を計算した．表 2 に推定話者区間に含まれる正解相槌区間の数を示している．正解データの相槌区間の総数は 2534 である．比較手法は，発話区間の長さを閾値とする手法 (区間長)，音響情報のみの識別モデル $p(v|a)$(音響) である．表 2 より，いずれの条件においても，提案手法 (音響+視線) により同定精度が向上した．

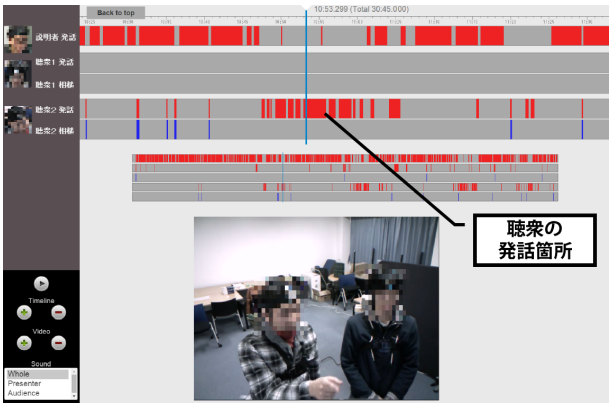

図 2: ポスター会話ブラウザ

6. ポスター会話ブラウザ

以上の検出結果を可視化し，ポスター会話を効率的に閲覧するブラウザ (図 2) を実装した．長時間のポスター会話の中で各参与者が発話した箇所を，相槌の同定結果も含めて時間軸上に表示している．これにより聴衆の発話区間などを容易に再生し確認することが可能になる．本システムは Web アプリケーションとして設計されており，ウェブブラウザが搭載されたすべてのコンピュータ上 (スマートフォンやタブレットを含む) で動作可能である．

謝辞 本研究は，JST CREST「人間調和型情報環境」領域の支援を受けて実施されたものである．

参考文献

[1] 河原達也, “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,” 人工知能研報, pp. 1-6, SIG-Challenge-B303-01, 2014.

[2] K. Inoue *et al.*, “Speaker diarization using eye-gaze information in multi-party conversations,” *INTERSPEECH*, pp. 562-566, 2014.

[3] S. Araki *et al.*, “A DOA based speaker diarization system for real meetings,” *HSCMA*, pp. 29-32, 2008.

[4] R. Schmidt, “Multiple emitter location and signal parameter estimation,” *IEEE Trans. Antennas and Propagation*, vol. 34, no. 3, pp. 276-280, 1986.

[5] 吉本廣雅 他, “未知剛体の形状と姿勢の実時間同時推定のための cubistic 表現,” 信学論 D, vol. J97-D, no. 8, pp. 1218-1227, 2014.

[6] A. Kendon, “Some functions of gaze-direction in social interaction,” *Acta psychologica*, vol. 26, no. 1, pp. 22-63 , 1967.

[7] T. Kawahara *et al.*, “Prediction of turn-taking by combining prosodic and eye-gaze information in poster conversations,” *INTERSPEECH*, pp. 727-730, 2012.

[8] 石井亮 他, “複数人対話における注視遷移パターンに基づく次話者と発話開始タイミングの予測,” 信学論 A, vol. J97-A, no. 6, pp.453-468, 2014.

[9] 岩野公司 他, “マルチモーダル音声認識におけるストリーム重みの教師なし推定法の検討,” 情処研報, pp. 1-6, 2009-SLP-76-24, 2009.

[10] N. Kitaoka *et al.*, “Response timing detection using prosodic and linguistic information for human-friendly spoken dialog systems,” *JSAI Journal*, vol. 20, no. 3, pp. 220-228, 2005.

[11] J. G. Fiscus *et al.*, “The rich transcription 2006 spring meeting recognition evaluation,” *Springer*, 2006.