Vol.2015-MUS-107 No.68
2015/5/24

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及び相槌の検出

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

概要：学会やオープンラボなどでなされる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，各々の会話参与者がいつ発話したかを表す話者区間の検出手法を述べる．ポスター会話のマルチモーダルなセンシング環境として，我々はスマートポスターボードの構築を進めている．従来法で用いられてきた音響情報に対して，会話における発話権取得で重要な役割を担う視線情報を統合した話者区間検出を行う．また，検出した聴衆の各発話区間に対して，それが相槌であるか否かを，話者区間検出と同様のマルチモーダルな手法で判定する．検出した相槌を発話区間から取り除くことで，質問やコメントなどの実質的な発話のみを抽出する．視線情報を用いることによる雑音環境下での話者区間検出の精度向上が実験により示されている．

キーワード：話者区間，相槌，マルチモーダル，視線，ポスター会話

1. はじめに

センシング技術の進展により，多人数会話のマルチモーダルな分析・処理に関する研究が可能になってきた．ミーティングを対象として，AMI / AMIDA[1] や VACE[2] などのプロジェクトでは，複数のカメラやマイクロホンを用いてマルチモーダルコーパスを構築する試みがなされてきた．マルチモーダルデータの分析では，言語的情報に加え，相槌，頷き，視線などの非言語的情報も含めた研究が行われている [3].

著者らはこれまでに，多人数会話の中でも特にポスターセッションにおける会話 (=ポスター会話) のインタラクションを対象にデータの収集と分析を行ってきた [4]．ポスターセッションは，学会やオープンラボなどにおいて，説明者が比較的少人数の聴衆に対してポスターを用いて説明を行うものである．ポスター会話のマルチモーダルな分析環境として，「スマートポスターボード」の構築を進めている．これは，大型液晶ディスプレイの周囲にマイクロホンアレイとカメラを配置したものである．この環境下で，ポスター会話における発話交替の予測 [5] や聴衆の興味・理解度の推定 [4] などの研究を行っている．

本研究では，ポスター会話での話者区間検出に取り組む．話者区間検出は，「いつ誰が発話したか」を検出する処理であり，多人数会話をアーカイブ化する上で基本的かつ重要な要素技術である．これまでに多くの手法が研究されており，これらの処理は音響情報に基づいている [6]．実際のポスター会話では，周囲の騒音や自然な話し言葉などにより検出精度が低下する．この問題を解決するために，音響情報に加えて，画像情報による動きベクトルや人物動作などを用いたマルチモーダルな手法が提案されている [7][8][9].

本研究では，話者区間検出において視線情報の利用を検討する．多人数会話における視線のふるまいは，発話権取得と関係があることが知られている [10][11]．例えば，発話権が交替する直前には，現話者と次話者は視線を交差（相互注視）する傾向にある．したがって，視線のふるまいから各参与者の発話予測が可能と考えられ，実際に予測手法が検討されている [12][5][13]. したがって，話者区間検出においては，音響情報から発話を検出し，それと同時に視線情報から発話を予測することで，音響的影響への頑健性が実現できると期待される．我々は音響情報と視線情報を統合するマルチモーダルな手法を提案し，視線情報の有効性を実験により示した [14]．さらに，検出した聴衆の各発話区間に対して，それが相槌であるか否かを，話者区間検出と同様のマルチモーダルな手法で判定した [15]．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，相槌と通常の発話権取得では視線のふるまいが異なると考えられ，視線情報を用いることで相槌の検出精度向上が期待される．ポ

[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

図 1　スマートポスターボード

スター会話では，聴衆の発話の多くが相槌であるため，相槌の検出及び除去により，質問やコメントなどの実質的な発話のみを抽出することが可能になる．

本論文の構成は以下の通りである．2 節では，ポスター会話の収録環境であるスマートポスターボードと収集したコーパスについて説明する．3 節では，話者区間検出について，音響情報と視線情報の統合法について述べる．4 節では，話者区間検出と同様のマルチモーダルな手法により聴衆の相槌を検出する．5 節では，ポスター会話コーパスを用いて提案法を評価する．最後に本論文のまとめを 6 節で述べる．

2. ポスター会話マルチモーダルコーパス

著者らが構築を進めているスマートポスターボードでは，大型液晶ディスプレイの上部に 19 チャネルのマイクロホンアレイ，Kinect センサ，高精細度カメラが配置されている（図 1）．この環境下で 8 セッションのポスター会話を収録した．各セッションでは 1 人の説明者が 2 人の聴衆に対して自身の研究に関して説明を行った．各セッションの長さは概ね 20〜30 分である．本研究における話者区間検出は，スマートポスターボード上に搭載されたマイクロホンアレイと Kinect センサのみで実現する．これにより，参与者は特別な装置を着用する必要がなく，実際のポスター会話に近い形態を実現する．ただし，コーパスを構築する上で正確な情報を取得する必要があるため，各参与者にワイヤレスヘッドセットマイクと磁気センサを着用してもらった．

発話時間の統計量を表 1 に示す．すべてのセッションにおいて説明者の発話が大部分を占めているのがわかる．それに対して聴衆の発話時間は少なく，検出が容易でないことを示唆している．また，聴衆の全発話時間のうち，約 40%が相槌であり，ポスター会話では相槌が多くうたれることがわかる．

表 1　合計発話時間 [秒]（括弧内は相槌）

セッション ID	説明者	聴衆 1	聴衆 2	計
140206-01	1,251	19 (11)	227 (111)	1,497
140206-02	1,406	283 (138)	164 (15)	1,853
140206-03	1,333	328 (160)	170 (86)	1,831
140206-04	1,495	129 (57)	102 (35)	1,726
140207-01	1,343	164 (48)	123 (21)	1,630
140207-02	1,229	134 (52)	117 (26)	1,480
140207-03	1,205	106 (41)	267 (79)	1,578
140207-04	1,208	216 (113)	135 (81)	1,559
計	10,470	2,684 (1,074)		13,154

3. 話者区間検出

著者らが提案している音響情報と視線情報を統合するマルチモーダルな話者区間検出について述べる．

3.1 音響情報に基づく話者区間検出

従来の話者区間検出では，音響情報としてメル周波数ケプストラム係数（Mel-Frequency Cepstral Coefficients; MFCCs）[6][8] や音声到来方向（Direction Of Arrival; DOA）[16][17] などが用いられてきた．本研究では，マイクロホンアレイを用いて音声到来方向を推定し，話者区間を検出する手法をベースラインとする．

音声到来方向の推定手法として，MUltiple SIgnal Classification（MUSIC 法）[18] 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時間フレーム t の MUSIC スペクトル $M_t(\theta)$ を角度 θ 毎に算出する．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．また，MUSIC スペクトルの算出には，音源数 N を事前に設定する必要があり，ここでは空間相関行列の固有値分布から各時間フレームの音源数を SVM により推定する [19]．

3.2 視線情報を統合した話者区間検出

話者区間検出において各参与者の視線情報を利用する．先行研究 [10][12][5][13] により，視線のふるまいから各参与者の発話予測が可能であることが示されている．したがって，音響情報に基づく手法により発話を検出し，それと同時に視線情報から発話を予測することで，音響的影響に頑健な話者区間検出を実現する．

本研究では，Kinect センサから取得したカラー画像と深度画像から各参与者の頭部の位置と方向を推定し，頭部方向を視線として代用する．頭部方向の推定手順 [20] は以下の通りである．はじめに，Haar-like 特徴を利用した物体認識法により各参与者の正面顔を探索する．検出された頭部について，距離画像から 3 次元形状を，カラー画像からその色情報を計算し，頭部モデルとする．この頭部モデルをパーティクルフィルタにより追跡処理し，頭部の三次元位置と方向を獲得する．注視判定は，頭部位置からその方向へ延びる半直線と対象物との距離の閾値処理により決定する．

音響情報と視線情報の統合処理は以下の通りである．はじめに，多チャネル音声信号と頭部位置から音響特徴量

(3.2.1 節) を，頭部位置と頭部方向から視線特徴量 (3.2.2 節) をそれぞれ抽出する．そして，これらの特徴量を確率的に統合し，各参与者が発話しているかを判定する (3.2.3 節)．これらの処理は各参与者で独立に，時間フレーム単位で行う．以下，それぞれの処理について述べる．ただし，参与者インデクスを i とし，各参与者の検出は推定頭部位置に基づく．

3.2.1 音響特徴量

音響特徴量はベースラインである MUSIC 法を基に算出する．スマートポスターボードでは，Kinect センサから取得した画像情報を基に参与者 i の頭部位置 $\theta_{i,t}$ を追跡している．頭部位置推定での誤差を考慮して，推定位置から一定の範囲内 $(\pm\theta_B)$ に参与者が存在するとみなす．この範囲の MUSIC スペクトルを，時間フレーム t における参与者 i の音響特徴ベクトル $\boldsymbol{a}_{i,t}$ とする．

$$\boldsymbol{a}_{i,t} = \left[M_t\left(\theta_{i,t}-\theta_B\right), \cdots, M_t\left(\theta_{i,t}\right), \cdots, M_t\left(\theta_{i,t}+\theta_B\right)\right]^T$$

3.2.2 視線特徴量

参与者 i の視線特徴量は，参与者 i と対話相手 [*1] それぞれの視線配布に基づき算出する．視線配布は，各対象物へ視線を向けているかを表す．参与者 i が説明者のとき，視線配布先の対象はポスター，または聴衆とする．また，参与者 i が聴衆のとき，視線配布先の対象はポスター，または説明者とする．さらに，参与者 i と対話相手との間での視線配布の組合せを視線状態 [5] と呼び，これを特徴量に追加する．これにより，説明者と聴衆の間での相互注視や共同注意を考慮することができる．また，時間フレーム t から過去 C ms の範囲で，参与者 i の各視線配布及び各視線状態の最大継続時間フレーム長，さらに視線配布の遷移 (ユニグラム，バイグラム) も考慮する．時間フレーム t における，参与者 i の視線特徴ベクトル $\boldsymbol{g}_{i,t}$ の構成を以下にまとめる [*2]．

時間フレーム t の情報

(1) 参与者 i の各視線配布の生起
(2) 参与者 i と対話相手との各視線状態の生起

時間フレーム t から過去 C ms までの情報

(3) (1) の最大継続時間フレーム長
(4) (2) の最大継続時間フレーム長
(5) 参与者 i の視線配布のユニグラム度数
(6) 参与者 i の視線配布のバイグラム度数
(7) 対話相手の視線配布のユニグラム度数
(8) 対話相手の視線配布のバイグラム度数

[*1] 参与者 i が説明者のとき対話相手は聴衆，また参与者 i が聴衆のとき対話相手は説明者とする．
[*2] (2)(7)(8) について，聴衆はまとめて扱い，1 人でも視線配布があれば計数する．

3.2.3 マルチモーダル統合モデル

音響特徴量 $\boldsymbol{a}_{i,t}$ と視線特徴量 $\boldsymbol{g}_{i,t}$ を確率モデルにより統合して，参与者 i の発話イベント $v_{i,t}$ を推定する．発話イベント $v_{i,t}$ は，時間フレーム t における参与者 i の発話 $(v_{i,t}=1)$，または非発話 $(v_{i,t}=0)$ を表す二値変数である．ここでは，音響特徴量と視線特徴量それぞれによる発話イベントを識別モデルにより推定し，事後確率を線形補間することで発話尤度 $f_{i,t}$ を計算する [8]．

$$\begin{aligned} f_{i,t}(\boldsymbol{a}_{i,t},\boldsymbol{g}_{i,t}) = {} & \alpha\, p(v_{i,t}=1|\boldsymbol{a}_{i,t}) \\ & + (1-\alpha)\, p(v_{i,t}=1|\boldsymbol{g}_{i,t}) \end{aligned} \quad . \qquad (1)$$

最終的に発話尤度 $f_{i,t}$ の閾値処理により発話を判定する．本研究では識別モデルの推定にロジスティック回帰モデルを用いる．その他のモデルとして，特徴量を直接統合する識別モデル $p(v_{i,t}|\boldsymbol{a}_{i,t},\boldsymbol{g}_{i,t})$ も考えられる．このモデルと比べて，線形補間モデルでは，音響と視線の識別モデルが独立しているため，学習データは音響と視線の対応づけをとる必要がなく，学習データ量が異なる場合でも学習可能である．さらに，重み係数 α により，SN 比などの音響環境に応じた事後確率の重みづけが可能である．例えば，背景雑音などが原因で音響情報の信頼性が低下した場合，視線情報の利用を優先させることができる．ここでは，環境の違いによるエントロピー変化に基づく推定手法 [21] を用いてオンラインで重み係数 α の値を決定する．音響特徴量に関して，発話イベントの事後確率の平均エントロピー h_c をクリーン環境である学習データを用いてあらかじめ求めておき，評価時にも同様のエントロピー h を求めて，これらの違いに基づいて以下の式で重みを決定する．

$$\alpha = \alpha_c \cdot \frac{1-h}{1-h_c} \qquad (2)$$

ただし，α_c はクリーン環境での理想重みである．推定した重みが 1 を越える場合は $\alpha=1$，0 を下回る場合は $\alpha=0$ とする．上記により，15 秒毎に過去 15 秒間の平均エントロピーを (2) 式の h として重みを更新する．

4. 相槌検出

前節により検出した聴衆の発話は相槌や雑音を含む可能性がある．本節では，話者区間検出の後処理である相槌の検出及び除去について述べる．この処理により，ポスター会話を後から振り返る際に，聴衆からの質問やコメントなどの発話のみの提示が可能になる．

従来の相槌検出手法として，GMM による MFCC のモデル化がある [22]．また，相槌検出ではなく，過去の情報から次に相槌がうたれるかの予測に関しては多くの研究がなされており [23][24]，視線情報も用いたマルチモーダルな手法も提案されている [25]．

相槌には，聞き手が発話権を取得せずに話し手の発話継続を促す役割があり，相槌と通常の発話権取得では視線の

ふるまいが異なると考えられる．したがって，これらの差異を学習することで，視線情報による相槌検出が可能にある．ここでは，前節のマルチモーダルな話者区間検出手法を相槌検出に適用する．視線特徴量（3.2.2 節）とマルチモーダル統合モデル（3.2.3 節）はそのまま用いるが，音響特徴量は相槌検出のために再設計する．多チャネル音声信号を入力として，遅延和法（delay-and-sum beamforming）により参与者 i の頭部位置 $\theta_{i,t}$ に応じた目的音強調を行う．強調されたシングルチャネル音声信号から以下を算出し，参与者 i の音響特徴量 $\boldsymbol{a}_{i,t}$ とする．

(1) MFCC （12 次元 MFCC，Δ MFCC） [22]
(2) パワー及び Δ パワー
(3) 先行発話末 100 ms での F0 とパワーの単回帰係数 [24]
(4) 発話区間の時間フレーム長（話者区間結果から算出）

ここでは発話イベント $v_{i,t}$ を 3 値変数に拡張し，相槌，雑音，相槌以外についての識別モデルをロジスティック回帰モデルにより推定する．話者区間検出の出力である聴衆の各発話区間において，各発話イベントの累積尤度を計算し，これを事後確率化（合計値が 1 になるように正規化）する．相槌及び雑音の事後確率の和を閾値処理し，閾値以上の発話区間を話者区間から除去する．

5. 評価実験

視線情報の有効性を評価するために，2 節のコーパスを用いて，提案法と音響情報に基づく手法の性能を比較した．

5.1 実験条件

ポスター会話 8 セッションの交差検定を行い，1 セッションを評価用，残りの 7 セッションを学習用とした．ロジスティック回帰モデルは説明者と聴衆で別々に学習した．

提案法に関する設定は以下の通りである．MUSIC 法において音源数 N を推定するための SVM，及びクリーン環境での音響特徴量による発話イベントの事後確率のエントロピー h_c は上記の学習用データから推定した．音響特徴量では，頭部位置から ±10° の MUSIC スペクトルを特徴量とした ($\theta_B = 10°$)．この θ_B は，各参与者間で抽出範囲ができるだけ重ならないように設定した．MUSIC スペクトルは 1° 毎に算出した．したがって，音響特徴ベクトル $\boldsymbol{a}_{i,t}$ は 21 次元である．視線特徴量 $\boldsymbol{g}_{i,t}$ において，各視線配布及び各視線状態の最大継続時間フレーム長，視線配布のユニグラム，バイグラムを算出する時間範囲は当該時間フレームから過去 1,000 ms までとした ($C = 1,000$)．各特徴量の 1 秒当たりのフレーム数は，音響特徴量が 62.5，視線特徴量が 30 である．したがって，視線特徴量を最近傍補間することで，両者を 62.5 にそろえた．マルチモーダル統合モデルにおいて，(2) 式のクリーン環境での最適重み α_c は，話者区間検出では 0.9，相槌検出では 0.8 とした．

雑音の影響を評価するために，19 チャネル音声信号に，発話区間の信号対雑音比（SN 比）の平均が 20，15，10，5，0 dB　となるように拡散性雑音を重畳した．この拡散性雑音は人混み環境下で実録音された 19 チャネル音声信号である．学会等で行われる大きな会場でのポスター会話は 0〜5 dB と想定される．

5.2 話者区間検出結果

話者区間検出に関して，提案法と以下の手法を比較した．

(1) 極大値 + GMM クラスタリング [17]
　全時間フレームでの MUSIC スペクトルの極大値を GMM クラスタリングする．クラスタ数は参与者数に対応し，各極大値に対応する参与者が発話したとみなす．ただし，極大値が閾値以下の場合，その極大値は雑音とみなす．この手法は画像情報を使用しない．
(2) 極大値 + 頭部位置 [26]
　MUSIC スペクトルの極大値と各参与者の頭部位置とを比較する．頭部位置から $\pm\theta_B$ の範囲に極大値が存在する場合，その参与者が発話したとみなす．ただし，極大値がどの参与者にも対応しない，または極大値が閾値以下の場合，その極大値は雑音とみなす．
(3) 音響のみ識別モデル
　提案法の (1) 式において $\alpha = 1$ に固定する，つまり音響特徴量のみを用いる．

評価指標として，話者区間誤り率（Diarization Error Rate; DER）[27] を用いた．DER は誤受理（False Acceptance; FA），誤棄却（False Rejection; FR），話者誤り（Speaker Error; SE）からなる．

$$\mathrm{DER} = \frac{\#\mathrm{FA} + \#\mathrm{FR} + \#\mathrm{SE}}{\#\mathrm{S}}$$

ただし，#S は正解データでの発話時間フレームの総数を表す．また，DER は正解発話区間の開始と終了それぞれにおいて，前後 250 ms の区間は評価の対象としない．各手法における閾値を，交差検定の 8 セッションで同一になるように変化させて，得られた DER の最小値で評価した．ただし，検出した発話区間に対して平滑化（ハングオーバー）処理を施している．

表 2 に各手法の DER を示す．極大値に基づく 2 手法（極大値 + GMM クラスタリング，極大値 + 頭部位置）は他の確率的手法（音響のみ識別モデル，マルチモーダル）に比べて精度が低かった．これは，極大値に基づく手法がルールベースであり，頭部位置や MUSIC スペクトルの動的な変化に頑健でないことが原因と考えられる．音響のみ識別モデルと提案法を比較すると，雑音環境下（SN 比が 5 または 0 dB）では提案法が高い精度を示した．したがって，実際のポスター会話に近い状況において，視線情報の効果を確認することができた．

提案法では，(1) 式における重み係数 α は自動推定手法によりオンラインで決定したが，0.1 刻みで手動によるチュー

表 2　話者区間検出精度（DER [%]）

手法		SN 比 [dB]						
		∞	20	15	10	5	0	ave.
極大値 + GMM クラスタリング	[17]	16.94	23.14	31.66	47.92	67.03	88.80	45.92
極大値 + 頭部位置	[26]	8.34	14.45	22.31	36.09	55.80	78.05	35.84
音響のみ識別モデル	(1) 式 w/o $\boldsymbol{g}_{i,t}$	**6.16**	**7.28**	**9.36**	14.20	22.94	35.89	15.97
マルチモーダル	(1) 式	6.27	7.81	9.96	**13.69**	**18.18**	**21.61**	**12.92**

表 3　聴衆の発話区間検出精度（EER [%]）

手法		SN 比 [dB]						
		∞	20	15	10	5	0	ave.
後処理なし		13.37	15.80	17.86	20.86	25.77	31.80	20.91
時間フレーム長	4 節 (4) のみ	15.95	17.60	18.64	20.38	24.74	30.81	21.35
音響のみ識別モデル	(1) 式 w/o $\boldsymbol{g}_{i,t}$	**12.14**	**13.98**	15.47	**18.19**	23.34	30.20	18.89
マルチモーダル	(1) 式	12.23	14.11	**15.42**	18.29	**23.07**	**29.72**	**18.80**

ニングも行った．クリーン環境（SN 比が ∞ dB）では，最適な重み係数は 1.0 だった．一方，雑音環境下では SN 比が 5 dB のとき 0.6，0 dB のとき 0.5 が最適な重み係数となった．このことから，音響情報の信頼性が低下する雑音環境下では，重み係数 α を小さく，つまり視線情報の重みを大きくする必要があるといえる．手動チューニングによる平均 DER は 11.78%であり，自動推定の場合（12.92%）に比べてわずかな改善であった．したがって，音響環境に応じた適切な重み係数の調整が自動推定でも可能であるといえる．

5.3　相槌検出結果

前節の実験において出力された話者区間のうち聴衆の発話区間に対して，相槌と雑音のモデルによる後処理を行った．したがって，ここでの正解発話データでは，相槌は非発話区間とした．

相槌検出に関して，提案法と以下の手法を比較した．ただし，話者区間検出手法はマルチモーダルな提案法（表 2 最下段の結果）に固定した．

(1) 時間フレーム長
　発話区間の時間フレーム長を閾値処理する．つまり，4 節の特徴量 (4) のみを用いる．相槌の時間フレーム長は，通常の発話よりも短いと考えられるため，閾値よりも短い発話は相槌とみなす．

(2) 音響のみ識別モデル
　提案法の (1) 式において $\alpha = 1$ に固定する，つまり音響特徴量のみを用いる．

相槌検出の各手法の閾値は実験的に設定し，時間フレーム長は 800 ms，音響のみ識別モデルと提案法の事後確率は 0.8 とした．

相槌以外の発話に関して，聴衆どうしの発話が重なり合うことはほとんどないため，誤受理 (FA) と誤棄却 (FR) のみで計算される等価誤り率（Equal Error Rate; EER）を評価指標として用いた．これは誤受理率 (False Acceptance Rate; FAR) と誤棄却率 (False Rejection Rate; FRR) が一致する値であり，それぞれ以下の式で算出される．

$$\mathrm{FAR} = \frac{\#\mathrm{FA}}{\#\mathrm{NS}}, \quad \mathrm{FRR} = \frac{\#\mathrm{FR}}{\#\mathrm{S}}$$

ただし，#NS は正解データでの非発話時間フレームの総数を表す．ここでは話者区間検出での閾値を変化させて，各閾値で出力される聴衆の発話区間に対して相槌の検出及び除去を行い，FAR と FRR を計算した．得られた複数の FAR と FRR の組から EER を算出した．

表 3 に各手法の EER を示す．後処理をしない場合に比べて，提案法は EER を大きく改善し，相槌及び雑音の除去の有効性を確認することができる．時間フレーム長を閾値とする手法は，雑音環境下では EER を改善することができたが，クリーン環境では EER が増加した．したがって，時間フレーム長のみの単純な閾値処理だけでは，検出が容易でないことがわかる．音響のみ識別モデルと提案法を比較すると，雑音環境下（SN 比が 5 または 0 dB）では提案法が高い精度を示した．したがって，話者区間検出のときと同様に，実際のポスター会話に近い状況において，相槌検出でも視線情報の効果を確認することができた．

6.　おわりに

本稿では，ポスター会話において視線情報を用いるマルチモーダルな話者区間検出を紹介した．さらに，検出した聴衆の発話区間から相槌及び雑音を検出・除去する手法についても述べた．実験結果より，雑音環境下において，提案法による検出精度の向上を確認した．

話者区間検出の応用例として，ポスター会話ブラウザ（図 2）を実装した．これは Web ブラウザ上で動作するアプリケーションであり，検出した話者区間，相槌，視線配

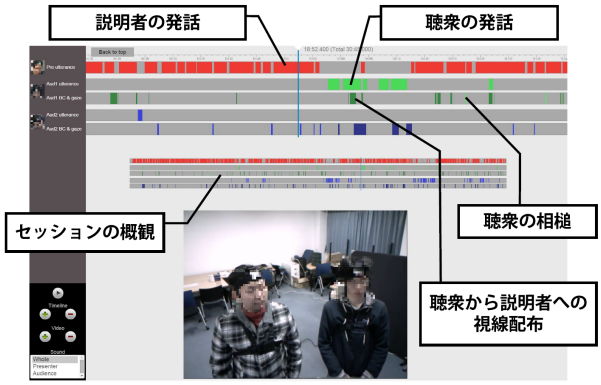

図 2　ポスター会話ブラウザ

布などを時間軸上に可視化することができる．可視化されたデータを映像や音声と共に確認することで，聴衆の質問やコメントの振り返りや，インタラクションの度合いの推察ができるようになっている．

参考文献

[1] Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D. and Wellner, P.: The AMI meeting corpus: A pre-announcement, *Machine Learning for Multimodal Interaction*, Springer, pp. 28–39 (2006).

[2] Chen, L., Rose, R. T., Qiao, Y., Kimbara, I., Parrill, F., Welji, H., Han, T. X., Tu, J., Huang, Z., Harper, M., Quek, F., Xiong, Y., McNeill, D., Tuttle, R. and Huang, T.: VACE multimodal meeting corpus, *Machine Learning for Multimodal Interaction*, Springer, pp. 40–51 (2006).

[3] Otsuka, K.: Conversation scene analysis, *IEEE Signal Processing Magazine*, Vol. 28, No. 4, pp. 127–131 (2011).

[4] Kawahara, T.: Smart posterboard: Multi-modal sensing and analysis of poster conversations, *Proc. APSIPA ASC*, pp. 1–5 (2013).

[5] Kawahara, T., Iwatate, T. and Takanashi, K.: Prediction of turn-taking by combining prosodic and eye-gaze information in poster conversations, *Proc. INTERSPEECH*, pp. 727–730 (2012).

[6] Tranter, S. E. and Reynolds, D. A.: An overview of automatic speaker diarization systems, *IEEE Trans. ASLP*, Vol. 14, No. 5, pp. 1557–1565 (2006).

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

[8] Friedland, G., Janin, A., Imseng, D., Miro, X. A., Gottlieb, L., Huijbregts, M., Knox, M. T. and Vinyals, O.: The ICSI RT-09 speaker diarization system, *IEEE Trans. ASLP*, Vol. 20, No. 2, pp. 371–381 (2012).

[9] Gebre, B. G., Wittenburg, P., Drude, S., Huijbregts, M. and Heskes, T.: Speaker diarization using gesture and speech, *Proc. INTERSPEECH*, pp. 582–586 (2014).

[10] Kendon, A.: Some functions of gaze-direction in social interaction, *Acta psychologica*, Vol. 26, No. 1, pp. 22–63 (1967).

[11] Duncan, Jr., S.: Some signals and rules for taking speaking turns in conversations, *Journal of personality and social psychology*, Vol. 23, No. 2, pp. 283–292 (1972).

[12] Jokinen, K., Harada, K., Nishida, M. and Yamamoto, S.: Turn-alignment using eye-gaze and speech in conversational interaction, *Proc. INTERSPEECH*, pp. 2018–2021 (2010).

[13] Ishii, R., Otsuka, K., Kumano, S. and Yamato, J.: Analysis and modeling of next speaking start timing based on gaze behavior in multi-party meetings, *Proc. ICASSP*, pp. 694–698 (2014).

[14] Inoue, K., Wakabayashi, Y., Yoshimoto, H. and Kawahara, T.: Speaker diarization using eye-gaze information in multi-party conversations, *Proc. INTERSPEECH*, pp. 562–566 (2014).

[15] 井上昂治，若林佑幸，吉本廣雅，高梨克也，河原達也：ポスター会話における音響・視線情報を統合した話者区間及び相槌の検出，情報処理学会研究報告，SLP-105-9 (2015).

[16] Anguera, X., Wooters, C. and Hernando, J.: Acoustic beamforming for speaker diarization of meetings, *IEEE Trans. ASLP*, Vol. 15, No. 7, pp. 2011–2022 (2007).

[17] Araki, S., Fujimoto, M., Ishizuka, K., Sawada, H. and Makino, S.: A DOA based speaker diarization system for real meetings, *Proc. HSCMA*, pp. 29–32 (2008).

[18] Schmidt, R.: Multiple emitter location and signal parameter estimation, *IEEE Trans. Antennas and Propagation*, Vol. 34, No. 3, pp. 276–280 (1986).

[19] Yamamoto, K., Asano, F., Yamada, T. and Kitawaki, N.: Detection of overlapping speech in meetings using support vector machines and support vector regression, *IEICE Trans. Fundamentals*, Vol. 89, No. 8, pp. 2158–2165 (2006).

[20] 吉本廣雅，中村裕一：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，電子情報通信学会論文誌 (D)，Vol. J97-D, No. 8, pp. 1218–1227 (2014).

[21] 岩野公司，松尾俊秀，古井貞熙：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，情報処理学会研究報告，SLP-76-24, pp. 1–6 (2009).

[22] Kawahara, T., Sumi, K., Chang, Z. and Takanashi, K.: Detection of hot spots in poster conversations based on reactive tokens of audience, *Proc. INTERSPEECH*, pp. 3042–3045 (2010).

[23] Ward, N. and Tsukahara, W.: Prosodic features which cue back-channel responses in English and Japanese, *Journal of pragmatics*, Vol. 32, No. 8, pp. 1177–1207 (2000).

[24] Kitaoka, N., Takeuchi, M., Nishimura, R. and Nakagawa, S.: Response timing detection using prosodic and linguistic information for human-friendly spoken dialog systems, *Journal of JSAI*, Vol. 20, No. 3, pp. 220–228 (2005).

[25] Ozkan, D. and Morency, L. P.: Modeling wisdom of crowds using latent mixture of discriminative experts, *Proc. ACL*, pp. 335–340 (2011).

[26] Wakabayashi, Y., Inoue, K., Yoshimoto, H. and Kawahara, T.: Speaker diarization based on audio-visual integration for smart posterboard, *Proc. APSIPA ASC* (2014).

[27] Fiscus, J. G., Ajot, J., Michel, M. and Garofolo, J. S.: *The rich transcription 2006 spring meeting recognition evaluation*, Springer (2006).