Vol.2015-SLP-105 No.9
2015/2/28

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

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を対象として，各会話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検出する手法を述べる．実際のポスター会話では，自然な話し言葉や周囲の騒音などにより検出精度が低下する．著者らは話者区間検出において，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するマルチモーダルな手法を提案している．本稿では，視線特徴量と確率的統合モデルについて検討・改善を行った．また，検出した聴衆の発話区間が相槌であるかを，話者区間検出と同様のマルチモーダルな手法で判定し，相槌を発話区間から除去することで質問やコメントなどの発話のみを抽出する．実験結果から，音響情報と視線情報を統合することで雑音環境下での検出精度の向上が確認された．

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

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近年盛んに行われている．ミーティングを対象として，AMI[1] や VACE[2] 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．我が国でも，マルチモーダルな自由会話の収集と分析が行われている [3][4]．また，会話参加者が着席した状態ではなく，自由に動き回る状況を想定したマルチモーダルな収録環境として IMADE ルーム [5] が構築された．

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収集と分析を行ってきた [6]．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．説明者がポスターの内容について説明をする一方で，聴衆は相槌や頷き，あるいは質問やコメントなどの反応を随時行う．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 [7] や聴衆の興味・理解度の推定 [8] などの研究を行っている．

本研究では，ポスター会話における話者区間検出について取り組む．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，ポスター会話をアーカイブ化する上で基本的かつ重要である．さらに，検出した聴衆の発話に対して，それが相槌であるか否かを判定する．ポスター会話では，聞き手である聴衆の発話は，その多くが相槌である．したがって，聴衆の相槌を検出することで，質問やコメントなどの相槌以外の発話を抽出することが可能になる．しかしながら，実際のポスター会話では自然な話し言葉や周囲の騒音などにより話者区間及び相槌の検出精度が低下する．

そこで本研究では，話者区間及び相槌検出において，会話参与者の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，発話権の交替と相関があることが知られている [9][10]．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に視線を向ける傾向がある．したがって，視線のふるまいから各参与者の発話を予測することが可能と考えられ，実際に視線情報を用いた発話予測が研究されている [7][11]．これまでに著者らは話者区間検出に関して，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た [12]．本稿では，視線特徴量と確率的統合モデルについて検討・改善を行う．さらに，この統合手法を相槌検出に適用する．

[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

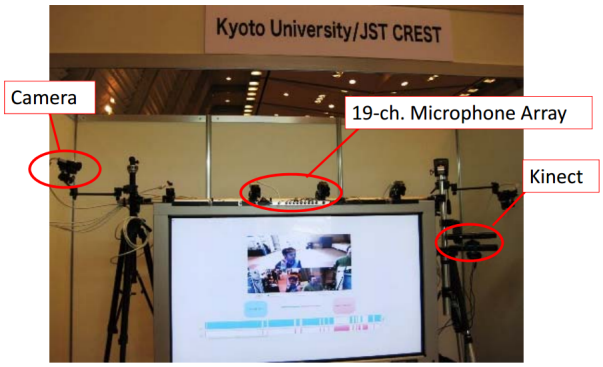

図 1　スマートポスターボードの外観

表 1　参与者毎の合計発話時間 [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

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 [6]．その外観を図 1 に示す．本研究では，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．聴衆は説明者についても研究内容についても事前に知らないように設定した．説明者と聴衆はすべてのセッションで異なる組合せである．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，各参与者が特別な装置を着用する必要はなく，実際のポスター会話に近い設定を実現する．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．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て，『日本語話し言葉コーパス』(CSJ) と同様の基準で各参与者の発話を書き起こした．相槌の認定基準は，「対話内に出現した感動詞のうち，話し手への反応として，聞き手が発しているもの」とした．「なるほど」や「そう」などの語彙的応答 [13] も相槌として含めた．また，磁気センサにより頭部位置および頭部方向のアノテーションデータを計測した．

各セッションにおける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，ポスター会話では聴衆が相槌を多くうつことがわかる．

3.　話者区間検出

提案する音響情報と視線情報を統合したマルチモーダルな話者区間検出について述べる．ここでは，新たに検討した視線特徴量と確率的統合モデルを中心に述べる．

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 (MFCC) や音声到来方向などが用いられてきた [14][15]．本研究では，スマートポスターボードに搭載された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，MUSIC 法 [16] 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 $P_{MU}(\theta)$ 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 を事前に求める必要があり，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毎の音源数を SVM により推定する [17]．

3.2　視線情報を統合した話者区間検出

提案法では，話者区間検出において，参与者の視線情報を利用する．先行研究 [9][10][7][11] 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各参与者の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の実現を目指す．

音響情報と視線情報の統合処理の流れを図 2 上部に示す．はじめに，多チャネル音声信号と各参与者の頭部位置から音響特徴量 (3.2.1) を，各参与者の頭部位置と頭部方向から視線特徴量 (3.2.2) 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)．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．各参与者のインデクスを i とし，各参与者の検出

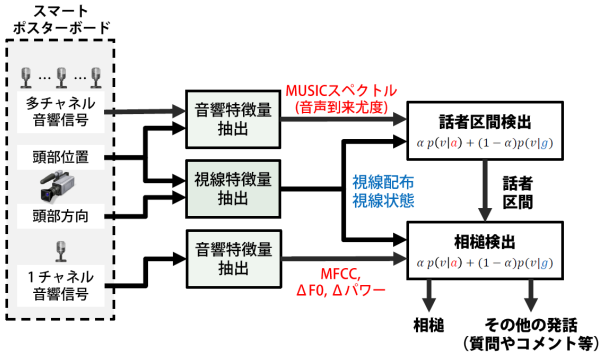

図 2　提案法の処理の流れ

は頭部位置推定に基づいて行う．

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 $\hat{\theta}_i$ を追跡している．追跡した頭部位置には誤差が含まれるため，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 スペクトルを参与者 i の音響特徴量ベクトル $\boldsymbol{a}_i$ とする．

$$\boldsymbol{a}_i = \left[P_{MU}\left(\hat{\theta}_i - \theta_B\right), \cdots, P_{MU}\left(\hat{\theta}_i\right), \cdots, P_{MU}\left(\hat{\theta}_i + \theta_B\right)\right]^T \quad (1)$$

3.2.2　視線特徴量

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方向を推定し，これを視線として代用する．頭部方向の推定手順 [18] 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．

参与者 i の視線特徴量ベクトル $\boldsymbol{g}_i$ は，参与者 i と対話相手 [*1] の視線配布に基づいて算出する．視線配布は，当該時間フレームにおいて，各対象物へ視線を向けているかどうかを表す．参与者 i が説明者のとき，視線配布先の対象物はポスター (**P**) または聴衆 (**I**) とする．また，参与者 i が聴衆のとき，視線配布先の対象物はポスター (**p**) または説明者 (**i**) とする．さらに，参与者 i と対話相手の視線配布の組合せである視線状態 [7] を特徴量に追加する．表 2 に視線状態の定義を示す．例えば，“**Ii**” は説明者と聴衆の相互注視を，“**Pp**” は共同注意をそれぞれ表している．また，当該時間フレームから過去 C msec の範囲で，参与者 i の各視線配布及び各視線状態が最大でどれくらい継続したか，さらに視線配布がどのように遷移したか (ユニグラム，バイグラム) も考慮する．参与者 i の視線特徴量ベクトル $\boldsymbol{g}_i$ の構成を以下にまとめる [*2]．

[*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．

表 2　説明者と聴衆の間での視線状態の定義

	説明者		
聴衆	視線配布先	聴衆 (**I**)	ポスター (**P**)
	説明者 (**i**)	**Ii**	**Pi**
	ポスター (**p**)	**Ip**	**Pp**

当該時間フレームの情報

(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

当該時間フレームから過去 C msec 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

3.2.3　確率的統合モデルによる発話判定

前節で定義した音響特徴量 $\boldsymbol{a}_i$ と視線特徴量 $\boldsymbol{g}_i$ 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 を判定する．発話イベント v_i は，当該時間フレームにおける参与者 i の発話 $(v_i = 1)$，または非発話 $(v_i = 0)$ を表す二値変数である．以下の確率モデルにおいて，参与者 i の発話尤度 $f_i(\boldsymbol{a}_i, \boldsymbol{g}_i)$ を計算し，この値の閾値処理により発話を判定する．ここでは，確率モデルとして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，結果統合する線形補間モデルを用いる．

$$f_i(\boldsymbol{a}_i, \boldsymbol{g}_i) = \alpha p_i(v_i = 1|\boldsymbol{a}_i) + (1-\alpha) p_i(v_i = 1|\boldsymbol{g}_i) \quad (2)$$

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．ただし，$\alpha \in [0, 1]$ は重み係数である．この重み係数は各識別モデルの学習とは別に設定可能である．また，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，音響と視線で学習データ量に違いがある場合でも学習が可能である．

重み係数 α について，後述の実験では，環境の違いによるエントロピーの変化を利用した重み推定手法 [19] により値をオンラインで決定する．音響特徴量に関して，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．

$$\alpha = \alpha_c \cdot \frac{1-h}{1-h_c} \quad (3)$$

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

[*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．

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．上記の手法により，15 秒毎に過去 15 秒間の平均エントロピー h を用いて重みを更新する．

4. 相槌検出

2 で述べたように，ポスター会話では聞き手である聴衆は相槌を多くうつ傾向にある．ポスター会話を振り返る上で重要な情報は，聴衆からの質問やコメントなどの相槌以外の発話である．したがって，ここでは前節の手法により検出された聴衆の発話区間について，相槌か相槌以外の発話であるかを判定し，相槌以外の発話区間のみを取り出す．

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 [20]．また，相槌の検出ではなく，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，先行発話末の F0 やパワーの傾きなどからの予測 [21] や，視線情報も用いたマルチモーダルな予測 [22] などが提案されている．

ここでは音響情報と視線情報を統合するマルチモーダルな検出手法を提案する．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検出精度の向上が期待される．音響情報と視線情報の統合処理の流れを図 2 下部に示す．音響特徴量 $\boldsymbol{a}_i$ はシングルチャネルの音響信号から以下を算出する．

(1) 発話区間の時間フレーム数
(2) MFCC (12 次元 MFCC，Δ MFCC，パワー，Δ パワー)
(3) 先行発話末 100msec での F0 とパワーの単回帰係数 [21]

視線特徴量 $\boldsymbol{g}_i$ と確率モデル $f_i(\boldsymbol{a}_i, \boldsymbol{g}_i)$ は話者区間検出と同様のものを用いる．相槌イベント b_i を導入し，当該時間フレームにおける参与者 i の相槌 ($b_i = 1$)，または相槌以外の発話 ($b_i = 0$) を表す．相槌尤度 $f_i(\boldsymbol{a}_i, \boldsymbol{g}_i)$ について，相槌と相槌以外の発話の累積尤度を各発話区間で計算する．これを事後確率化した値の閾値処理により相槌か相槌以外の発話かを判定する．

5. 評価実験

視線情報の有効性を評価するために，提案法と音響情報のみに基づく手法とを実験により比較した．2 で述べた収録コーパスを用いて，ポスター会話 8 セッションの交差検定を行った．8 セッションのうち，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．(2) 式の確率モデルは説明者と聴衆で別々に学習した．

音響情報に関する設定は以下の通りである．音声データのサンプリングレートは 16 kHz である．MUSIC 法における 1 フレームの長さは 32 msec，フレームの移動幅は 16 msec，ブロックは当該時間フレームの前後 2 フレーム分（計 5 フレーム分）とした．MUSIC スペクトルは 19 チャネルのマイクロホンアレイの音声信号から算出した．

音響的雑音の影響を評価するために，19 チャネル音声データに，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．拡散性雑音は人混み環境下で実録音された 19 チャネル音声データである．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．

提案法に関する設定は以下の通りである．音響特徴量を抽出する際に，画像情報から推定された参与者の頭部位置から ±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，音響特徴量 $\boldsymbol{a}_i$ の次元は 21 次元である．視線特徴量 $\boldsymbol{g}_i$ 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 msec までとした ($C = 1000$)．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．

5.1 話者区間検出

話者区間検出誤り率 (Diarization Error Rate; DER)[24] により精度評価を行った．

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

ここで，#FA は非発話を発話と誤検出した時間フレーム数，#FR は発話を非発話と誤検出した時間フレーム数，#SE は話者誤りの時間フレーム数，#S は発話時間フレーム数をそれぞれ表す．また，DER は正解発話区間の開始と終了の前後 250 msec の区間は評価の対象としない．(2) 式の f_i に対する閾値を，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，検出した発話区間及び非発話区間に対して，短い発話 [*3] を非発話に，短い非発話 [*4] を発話にするハングオーバー処理を施している．また，(3) 式におけるクリーン環境での理想重み α_c は 0.9 に設定した．比較手法は，MUSIC スペクトルのピークを角度領域で GMM クラスタリングする手法 (=GMM クラスタリング)[14]，MUSIC スペクトルのピークと各参与者の頭部位置とを照合する手法 (=GMM クラスタリング+画像位置)[23] を用いた．また，提案法と同じ枠組みで視線特徴量を用いない場合 ($\alpha = 1$ の場合) も評価した．

SN 比を変化させたときの各手法の DER を表 3 に示す．GMM クラスタリングを用いる 2 手法に比べて，提案法は全ての SN 比において高い精度を示した．GMM クラスタ

[*3] 208 msec （13 フレーム）以下
[*4] 80 msec （5 フレーム）以下

表 3　話者区間検出精度 (DER [%])

話者区間検出手法	SNR [dB]						平均
	∞	20	15	10	5	0	
GMM クラスタリング [14]	16.94	23.14	31.66	47.92	67.03	88.80	45.92
GMM クラスタリング + 画像位置 [23]	8.34	14.45	22.31	36.09	55.80	78.05	35.84
音響特徴量のみ ($\alpha = 1$)	6.16	7.28	9.36	14.20	22.94	35.89	15.97
マルチモーダル (提案法)	6.27	7.81	9.96	13.69	18.18	21.61	12.92

表 4　相槌検出精度 (F 値 [%])

相槌検出手法	SNR [dB]						平均
	∞	20	15	10	5	0	
時間フレーム長	67.07	54.28	44.46	35.35	22.99	12.61	39.46
音響特徴量のみ ($\alpha = 1$)	78.03	67.04	56.87	42.94	27.10	13.19	47.53
マルチモーダル (提案法)	78.69	68.00	59.13	45.51	29.64	15.27	49.37
#相槌区間	604	410	293	189	93	25	-

リングを用いる手法はルールベースであり，MUSIC スペクトルの大小や各参与者の位置などの動的な変化に対して頑健でないためと考えられる．一方，音響特徴量のみの確率モデルと提案法は，いずれも学習ベースの手法であり，異なるのは視線特徴量を用いるか否かである．この 2 手法を比較すると，SN 比が 5 dB 以下において提案法の DER が大きく改善している．クリーンに近い条件では両手法の性能差はほとんどなく，一般的なポスターセッションの会場で想定される 0〜5 dB において提案法は大きな改善を実現している．すなわち，音響特徴量の信頼性が低下する雑音環境下において，視線特徴量の有効性が確認できる．

5.2　相槌検出

前節の話者区間検出により得た聴衆の発話区間に対して相槌の検出を行った．話者区間検出には提案法を用い，聴衆のみの DER が最小となるときの発話区間を入力として，相槌検出に関して比較した．前処理として，発話の持続長が 2,000msec 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 を基にした以下の F 値 F である．

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

ただし，本研究は適合率を重視して，$\beta = 0.5$ とした．相槌検出の f_i に対する閾値を，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，得られた最大 F 値で評価した．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．各正解相槌区間の中央の時間フレームが，判定の対象である発話区間に含まれていれば，それを正解相槌区間とした．また，(3) 式におけるクリーン環境での理想重み α_c は 0.5 に設定した．表 4 下部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 2,576 である．相槌検出の比較手法は，発話区間の長さを閾値とする手法 (=時間フレーム長)，音響特徴量のみの確率モデル ($\alpha = 1$ の場合) である．

SN 比を変化させたときの各手法の F 値を表 4 に示す．いずれの条件においても，視線情報を用いるマルチモーダルな提案法が最も高い検出精度を示した．したがって，相槌検出においても視線特徴量の有効性が確認できる．また，相槌は比較的短い発話と考えられるが，時間フレーム長のみよりも音響や視線の特徴量を追加したほうがより高い検出精度を示すこともわかる．

5.3　聴衆の相槌以外の発話区間検出

検出した聴衆の相槌を発話区間から取り除き，相槌以外の発話の検出精度を評価した．したがって，ここでの正解発話ラベルは前節と異なり，相槌は非発話とみなされる．ここでは聴衆の相槌以外の発話の検出に焦点をおき，DER における話者誤り (SE) は考慮せずに誤受理 (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 は非発話時間フレーム数を表す．ここでは話者区間検出での発話尤度 f_i 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．そして各閾値で得られた FAR と FRR から EER を算出した．

話者区間検出には提案法を用い，相槌検出に関して比較した．ただし，前節と同様の前処理を行った．比較手法は，相槌検出をせずに話者区間検出結果をそのまま評価する方法 (=相槌検出なし)，発話区間の長さを閾値とする手法 (=時間フレーム長)，音響特徴量のみの確率モデル ($\alpha = 1$ の場合) である．時間フレーム長では閾値を 50 フレーム (0.8 秒)，音響特徴量のみの確率モデル及びマルチモーダルな提

表 5　聴衆の発話区間検出精度 (EER [%])

相槌検出手法	SNR [dB]						平均
	∞	20	15	10	5	0	
相槌検出なし	14.82	16.69	18.39	21.52	26.15	32.32	21.65
時間フレーム長	17.04	17.93	18.86	22.58	26.68	32.27	22.56
音響特徴量のみ ($\alpha = 1$)	14.42	16.23	17.72	20.38	25.72	32.35	21.14
マルチモーダル (提案法)	14.43	16.38	17.91	20.64	25.29	31.79	21.07

案法では事後確率の閾値を 0.8 とした．

SN 比を変化させたときの各手法の EER を表 5 に示す．提案法は，相槌検出を行わない場合に比べて平均で 0.58%EER を改善した．また，音響特徴量のみと比べると SN 比が 5 dB 以下において改善がみられた．したがって，雑音環境下において，視線情報を用いた相槌検出により聴衆の相槌以外の発話の検出精度改善が確認できる．

6. おわりに

本稿では，ポスター会話における話者区間及び相槌検出について，視線情報を用いるマルチモーダルな手法を提案した．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．

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] Carletta, J. et al.: The AMI meeting corpus: A pre-announcement, *Machine learning for multimodal interaction*, Springer, pp. 28–39 (2006).
[2] Chen, L. et al.: VACE multimodal meeting corpus, *Machine Learning for Multimodal Interaction*, Springer, pp. 40–51 (2006).
[3] Campbell, N. et al.: A Multimedia Database of Meetings and Informal Interactions for Tracking Participant Involvement and Discourse Flow, *Proc. LREC* (2006).
[4] Otsuka, K.: Conversation Scene Analysis, *Signal Processing Magazine, IEEE*, Vol. 28, No. 4, pp. 127–131 (2011).
[5] 角　康之，西田豊明，坊農真弓，來嶋宏幸：IMADE:会話の構造理解とコンテンツ化のための実世界インタラクション研究基盤，情報処理，Vol. 49, No. 8, pp. 945–949 (2008).
[6] 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，SIG-Challenge-B303-01, pp. 1–6 (2014).
[7] 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).
[8] Kawahara, T., Hayashi, S. and Takanashi, K.: Estimation of Interest and Comprehension Level of Audience through Multi-modal Behaviors in Poster Conversations, *Proc. INTERSPEECH*, pp. 25–29 (2013).
[9] Kendon, A.: Some functions of gaze-direction in social interaction, *Acta psychologica*, Vol. 26, No. 1, pp. 22–63 (1967).
[10] 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).
[11] 石井　亮，大塚和弘，熊野史朗，松田昌史，大和淳司：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).
[12] 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).
[13] 吉田奈央, 高梨克也, 伝　康晴：対話におけるあいづち表現の認定とその問題, 人工知能研報, SIG-Challenge-B303-01, pp. 1–6 (2014).
[14] 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).
[15] 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).
[16] Schmidt, R.: Multiple emitter location and signal parameter estimation, *IEEE Trans. Antennas and Propag.*, Vol. 34, No. 3, pp. 276–280 (1986).
[17] 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).
[18] 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，信学論 (D)，Vol. J97-D, No. 8, pp. 1218–1227 (2014).
[19] 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，情処学研報，2009-SLP-76-24, pp. 1–6 (2009).
[20] 河原達也，須見康平，緒方　淳，後藤真考：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，情処学論，Vol. 52, No. 12, pp. 3363–3373 (2011).
[21] Kitaoka, N., Takeuchi, M., Nishimura, R. and Nakagawa, S.: Response timing detection using prosodic and linguistic information for human-friendly spoken dialog systems, *JSAI Journal*, Vol. 20, No. 3, pp. 220–228 (2005).
[22] Ozkan, D. and Morency, L. P.: Modeling wisdom of crowds using latent mixture of discriminative experts, *Proc. ACL*, pp. 335–340 (2011).
[23] Wakabayashi, Y., Inoue, K., Yoshimoto, H. and Kawahara, T.: Speaker Diarization based on Audio-Visual Integration for Smart Posterboard, *Proc. APSIPA* (2014).
[24] Fiscus, J. G., Ajot, J., Michel, M. and Garofolo, J. S.: The rich transcription 2006 spring meeting recognition evaluation, *Springer* (2006).