Vol.2014-SLP-102 No.1
2014/7/24

多人数会話における視線情報を用いた話者区間検出

井上 昂治[1]　若林 佑幸[2]　吉本 廣雅[2]　河原 達也[1,2]

概要：多人数会話において視線情報を用いた話者区間検出手法を提案する．実世界の多人数会話では，マイクから離れた位置での発話や周囲の騒音などにより，話者区間検出精度が低下する．一方，会話参与者の視線情報は，これらの音響的影響を受けない．また，視線配布は会話の発話権取得に重要な役割を担っているため，発話の予測にも有用であると考えられる．提案手法は，音響と視線の情報を確率的に統合するもので，3 種類のモデル化を行う．実際に収録したポスター会話を用いた評価実験において，提案手法により，音響情報のみを用いたモデルに比べて，話者区間検出精度が向上した．

キーワード：話者区間検出，マルチモーダル，視線，ポスター会話，スマートポスターボード

1. はじめに

ミーティングや会話などの多人数インタラクションのマルチモーダルな分析や処理に関する研究が近年盛んに行われている [1], [2]．日常の自然会話では，言語的情報だけでなく，相槌，頷き，視線配布などの非言語的情報もやり取りされる．これらを考慮したマルチモーダルな多人数インタラクションは複雑ではあるが，会話参与者の非言語的ふるまいは多人数会話を分析する上で有用である．

我々はこれまでに，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分析を行ってきた [3], [4], [5], [6]．ポスター会話とは，学会やオープンラボなどでよくみられる会話形態であり，一人の説明者が複数名の聴衆に対してポスターを参照しながら説明を行う．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ため，マルチモーダルなインタラクションを観測することができる．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．スマートポスターボードでは，大型液晶ディスプレイの周囲にマイクロフォンアレイとカメラが配置されている．この環境下で，ポスター会話における発話権取得 [7] や聴衆の興味・理解度 [8] に関する分析を行った．

本研究ではポスター会話における話者区間検出 (speaker diarization) について検討する．話者区間検出とは，「いつ誰が発話をしたか」を検出する処理であり，これまでに様々な手法が提案されている [9], [10]．これらは主に，単一または複数のマイクロフォンから入力された音響信号に基づいて処理を行う．しかし，実際のポスター会話は人混みなどの雑音環境下である．また，参与者全てに接話マイクを装着させるのは非現実的であり，これにより参与者の発話はマイクロフォンアレイから離れた位置で観測される．また，自然会話のように発話が非明瞭 (音声認識を意識しない発話) になり，複数話者の発話が重なることもある．これらの要因によって話者区間検出精度が低下する．また，ポスター会話の参与者は，会議等とは異なり，立ち上がった状態で動きを伴いながら発話する．そのため各参与者の位置も同定及び追跡する必要がある．さらに，聴衆の発話は分析において重要な情報であるが，その頻度は説明者に比べて圧倒的に少ない．したがって，独立成分分析に代表される統計量による音源分離フィルタ [11] を構成することは困難で，音源分離による話者区間検出は容易でない．

本稿では，音響情報だけでなく，会話参与者の視線情報も用いる話者区間検出手法を提案する．視線のふるまいは多人数会話において発話権取得と相関があることが報告されている [7], [12], [13], [14]．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，視線のふるまいは発話予測に有用であると考えられるが，話者区間検出における視線情報の効果はまだ確認されていない．視線情報は上述の音響的影響に頑健であるため，視線情報と音響情報を統合することにより，話者区間検出の精度向上が期待される．

[1] 京都大学 大学院情報学研究科
[2] 京都大学 学術情報メディアセンター

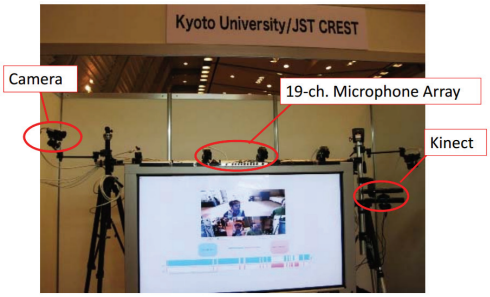

図 1　スマートポスターボードの概観

本稿では，まず 2 章でポスター会話を収録するマルチモーダル環境と収集したコーパスについて説明する．3 章では話者区間検出のための特徴量として音響情報と視線情報それぞれについて述べる．4 章では音響情報と視線情報の特徴量を確率的に統合するマルチモーダルな話者区間検出手法を提案する．ここでは 3 種類のモデルを検討する．5 章では収集したポスター会話コーパスを用いた話者区間検出実験により，提案手法の有効性を示す．

2. ポスター会話のマルチモーダルコーパス

我々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 チャンネルのマイクロフォンアレイ，Kinect センサ，高精細度カメラが配置されている [5]．スマートポスターボードの概観を図 1 に示す．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する説明を行った．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み合わせである．各セッションの長さは概ね 20〜30 分程度である．

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フォンアレイと Kinect センサのみで実現する．そのため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形態を実現した．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．ワイヤレス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）に分割され，時間と話者ラベルを付与し，「日本語話し言葉コーパス」(CSJ) と同様の基準で書き起こしを行った．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説明者に比べて大幅に少なく，検出が容易でないことを示唆している．しかし，ポスター会話の分析において聴衆の質問やコメントは重要な情報であるため，聴衆の発話区間を正確に同定する必要がある．

表 1　各セッションにおける参与者毎の合計発話時間 [秒]

セッション ID	説明者	聴衆 (2 名)		計
140206-01	1,261	19	230	1,510
140206-02	1,417	285	166	1,868
140206-03	1,344	331	172	1,847
140206-04	1,507	131	104	1,742
140207-01	1,354	166	125	1,645
140207-02	1,239	135	119	1,493
140207-03	1,215	107	270	1,592
140207-04	1,218	218	137	1,573
計	10,555	2,715		13,270

3. 話者区間検出のための特徴量

音響情報と視線情報のそれぞれにおいて話者区間検出のための特徴量を設計する．

3.1 音響情報による特徴量

音響情報による話者区間検出として，音声到来方向に基づく方法 [15] がある．本研究では音声到来方向の検出手法として，MUltiple SIgnal Classification [16]（以下，MUSIC 法）を用いる．MUSIC 法は実環境での音源位置推定問題に応用されており [17]，その特徴として複数音源の到来方向を同時に検出することができる．そのため複数話者の発話が重なった場合でも対応が可能である．

MUSIC 法は観測信号の部分空間の直交性に基づいて音声到来方向を推定する．観測信号のフーリエ係数を $\boldsymbol{x} \in \mathbb{C}^M$，この空間相関行列の推定値を $\boldsymbol{R_x} = E\left[\boldsymbol{x}\boldsymbol{x}^H\right] \in \mathbb{C}^{M\times M}$ とする．ただし，M はマイクロフォンの数，$\cdot^H$ はエルミート転置，$E[\cdot]$ は設定されたブロックでの期待値をそれぞれ表す．空間相関行列 $\boldsymbol{R_x}$ の固有ベクトルを $\boldsymbol{e}_i\,(i = 1, \cdots, M)$，$\boldsymbol{e}_i$ に対応する固有値を $\lambda_i\,(\lambda_1 \geq \cdots \geq \lambda_M)$ とし，音源数が N の場合，角度 θ 方向の MUSIC スペクトル $P_{MU}(\theta)$ は雑音部分空間の基底ベクトル $\boldsymbol{e}_i\,(i = N+1, \cdots, M)$ とステアリングベクトル $\boldsymbol{t}\,(\theta) = [\exp(-j\omega\tau_1(\theta)), \cdots, \exp(-j\omega\tau_M(\theta))]^T$ を用いて次式のように表される．

$$P_{MU}(\theta) = \frac{||\boldsymbol{t}(\theta)||^2}{\sum_{i=N+1}^{M} |\boldsymbol{t}^H(\theta)\ \boldsymbol{e}_i|^2} \qquad (1)$$

ただし，ω は周波数，$\tau_i\,(\theta)\,(i = 1, \cdots, M)$ は角度 θ から音声が到来するときにマイクロフォンアレイの基準点と i 番目のマイクロフォン素子との間に作る相対的な遅延時間である．仮に角度 θ^* の方向から参与者が発話した場合，$\boldsymbol{t}(\theta^*)$ と $\boldsymbol{e}_i$ が直交し，MUSIC スペクトル $P_{MU}(\theta^*)$ は大きな値をとる．つまり，MUSIC スペクトルの大きさはその角度から参与者が発話したかを示す手がかりとなる．話者区間検出のベースライン手法では，角度領域において MUSIC スペクトルの極大値を探索することにより参与者の発話を検出する．

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る．話者区間検出では，当該時間フレームにおいて発話している参与者の数に相当する．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推定する．これまでに提案された推定方法として，赤池情報量基準 (AIC) 及び最小記述長 (MDL) に基づく手法 [18] や，固有値分布からサポートベクターマシン (SVM) により学習する手法 [19] がある．本研究では雑音への頑健性を考慮して後者を採用する．空間相関行列の固有値を，その最大値で正規化した値 $\lambda_i' = \lambda_i/\lambda_1 (i = 1, \cdots, M)$ のうち上位 L 個を取り出したものを SVM における特徴量として音源数を学習した．学習及び推定は各周波数ビン毎に行い，各周波数ビンでの推定結果を投票することで当該時間フレームにおける音源数を決定する．

次に，MUSIC 法により検出された音声到来方向がどの参与者からのものであるかを同定する必要がある．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各参与者の頭部位置 ($\hat{\theta}$) 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．この範囲内の MUSIC スペクトルを音響情報による話者区間検出のための特徴量ベクトル $\boldsymbol{a}$ とする．

$$\boldsymbol{a} = \left[P_{MU}\left(\hat{\theta} - \theta_B\right), \cdots, P_{MU}\left(\hat{\theta}\right), \cdots, P_{MU}\left(\hat{\theta} + \theta_B\right)\right]^T \quad (2)$$

3.2 視線情報による特徴量

Kinect センサから取得したカラー画像と深度画像から参与者の頭部方向を推定し [20]，これを視線として代用する．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探索を行う．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により決定する．本研究では，注視対象物をポスターと他参与者に限定する．

視線情報による特徴量ベクトル $\boldsymbol{g}$ は，説明者と各聴衆間での視線方向と視線状態 [7] により以下のように構成される．

(1) 視線配布
　当該時間フレームにおいて各参与者から各対象物への視線配布が生起したかを表す．各参与者における視線配布の対象物は以下の通りである．
　説明者：(P) ポスターまたは (I) 聴衆
　聴衆：(p) ポスター，(i) 説明者，(o) 他の聴衆

(2) 視線状態: “Ii”, “Ip”, “Pi”, “Pp”
　当該時間フレームにおいて各視線状態が生起したかを表す．視線状態は説明者と各聴衆間での視線配布の組み合わせにより定義される．ただし，各聴衆における (o) その他の聴衆への視線方向は除く．

(3) 視線状態のバイグラム
　視線状態の変化の頻度を表す．

(4) 各視線配布の継続時間 ((I) と (i) のみ)

(5) 各視線状態の継続時間 (“Pp” は除く)

ただし，バイグラムと継続時間は当該時間フレームから過去 C 秒の区間で算出される．

4. 確率的統合による話者区間検出モデル

前節で定義した音響情報と視線情報による特徴量を統合する 3 種類の話者区間検出モデルを提案する．特徴量は各参与者毎に算出され，これを確率変数とみなす．ここでは，a を音響情報による特徴量，g を視線情報による特徴量，v を発話イベント，それぞれの確率変数とする．ただし，v は発話 ($v = 1$)，非発話 ($v = 0$) を示す二値変数である．各参与者で独立に v の事後確率を推定することで話者区間検出を実現する．なお，各参与者の位置情報は Kinect センサで取得し，その位置情報に基づいて，音響情報と視線情報の対応付けを行う．

4.1 モデル 1 (結合識別モデル)

結合識別モデルでは，音響と視線の確率変数が共起すると仮定し，発話イベントの事後確率を推定する．

$$p(v|a, g) \quad (3)$$

ここでは，ロジスティック回帰モデル (LR) により，この識別モデルを直接推定する．

4.2 モデル 2 (独立識別モデル)

独立識別モデルでは，音響と視線の確率変数が独立であると仮定し，それぞれを条件とする発話イベントの事後確率を線形補間する．

$$\alpha\, p(v|a) + (1 - \alpha)\, p(v|g) \quad (4)$$

パラメータ $\alpha \in [0, 1]$ は重み係数である．各識別モデルはモデル 1 と同様にロジスティック回帰で推定する．

4.3 モデル 3 (雑音のある通信路モデル)

モデル 1 における発話イベントの事後確率をベイズの定理により以下のように展開する．

$$p(v|a, g) = \frac{p(a|v, g)\, p(v|g)}{p(a|g)} \quad (5)$$

$$= \frac{p(a|v)\, p(v|g)}{p(a)} \quad (6)$$

ここでは，音響特徴量と視線特徴量の確率変数が独立であ

ると仮定している．分母の項は発話イベントの決定に影響しないため無視することができ，以下の尤度を得る．

$$l(v|a,g) = p(a|v)\ p(v|g) \tag{7}$$

生成モデル $p(a|v)$ は混合ガウスモデル (GMM) で，識別モデル $p(v|g)$ はロジスティック回帰で，それぞれ推定する．これは音声認識と同様の枠組みであると考えることができ，第一項は音響モデル，第二項は言語モデルに対応する．音声認識の枠組みのように，(7) 式の対数をとり，2 つのモデルのダイナミックレンジの違いを補償する重み係数 β を導入することで以下の対数尤度を得る．

$$ll(v|a,g) = \log p(a|v) + \beta \log p(v|g) \tag{8}$$

発話区間検出は対数尤度差の閾値判定により決定する．

$$ll(v=1|a,g) - ll(v=0|a,g) \begin{cases} \geq \Theta_{ll} & \text{発話} \\ < \Theta_{ll} & \text{非発話} \end{cases} \tag{9}$$

ただし，Θ_{ll} は閾値を表す．

5. 話者区間検出実験

収録したポスター会話コーパスを用いた話者区間検出実験により，マルチモーダルな提案モデルと，音響情報のみに基づくベースラインモデルとを比較した．

5.1 実験条件

音響情報に関するパラメータ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 5 フレームとした．本実験では，MUSIC スペクトルは 19 チャンネルの音声信号から算出した．SVM による音源数の推定では，正規化された固有値のうち上位 5 個を特徴量とした (L=5)．SVM におけるカーネルはガウシアンカーネルを用いた．

提案モデルでのパラメータ設定は以下の通りである．音響情報による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 ($\theta_B = 10°$)．この角度を設定するにあたり，各参与者間において抽出範囲が重ならないようにした．MUSIC スペクトルは 1° 毎に算出した．したがって，音響特徴量 $\boldsymbol{a}$ の次元は 21 次元である．視線情報による特徴量 $\boldsymbol{g}$ において，バイグラム及び継続時間を計算する対象範囲は当該時間フレームから過去 10 秒間とした ($C = 10$)．

モデルの学習及び評価はポスター会話 8 セッションの交差検定により行った．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て，ロジスティック回帰及び混合ガウスモデルを学習した．混合ガウスモデルの混合数は 8 とした．モデルの学習に続き，提案モデル 2 及び 3 での重み係数 (α と β) を同じ学習用セッションから学習した．また，音源数推定のための SVM も同じ学習用セッションから学習した．

ベースライン手法を含む比較手法は以下の通りである．

(1) ベースライン [15]
各時間フレームにおいて，MUSIC スペクトルを角度領域で極大値探索する．検出された全極大値を用いて角度領域での GMM クラスタリングを行う．このときの混合数は 3 であり，これは参与者数に対応する．つまり，GMM クラスタリングによって得られる各クラスタは各参与者に対応する．当該時間フレームに極大値が存在する場合，これに対応する参与者が発話したとみなす．ただし，MUSIC スペクトルの値が閾値以下の場合，その極大値は雑音 (非発話) とする．この手法は Kinect センサから得られる画像情報を一切使用しない．

(2) ベースライン + 画像位置 [21]
ベースラインと同様に角度領域で MUSIC スペクトルの極大値探索を行う．得られた極大値の角度を，画像情報から推定した頭部位置と比較する．ある参与者の推定頭部位置から $\pm\theta_B$ 以内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，MUSIC スペクトルの値が閾値以下の場合，その極大値は雑音 (非発話) とする．

(3) 音響情報のみの識別モデル
提案モデル 1 及び 2 において視線情報を用いないモデルである．音響情報のみを用いて以下の事後確率をロジスティック回帰により推定する．

$$p(v|a) \tag{10}$$

(4) 音響情報のみの生成モデル
提案モデル 3 において視線情報を用いないモデルである．音響情報のみを用いて以下の条件つき確率を混合ガウスモデルにより推定する．

$$p(a|v) \tag{11}$$

発話区間検出は尤度比の閾値判定により決定する．

$$\frac{p(a|v=1)}{p(a|v=0)} \begin{cases} \geq \Theta_l & \text{発話} \\ < \Theta_l & \text{非発話} \end{cases} \tag{12}$$

混合ガウスモデルの混合数は 8 とした．

音響的雑音の影響を評価するために，19 チャンネル音声データに，信号対雑音比 (SNR) が 0，5，10 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際に録音された 19 チャンネル音声データである．

5.2 実験結果

話者区間検出精度を適合率と再現率で評価した．それぞれの定義を以下に示す．

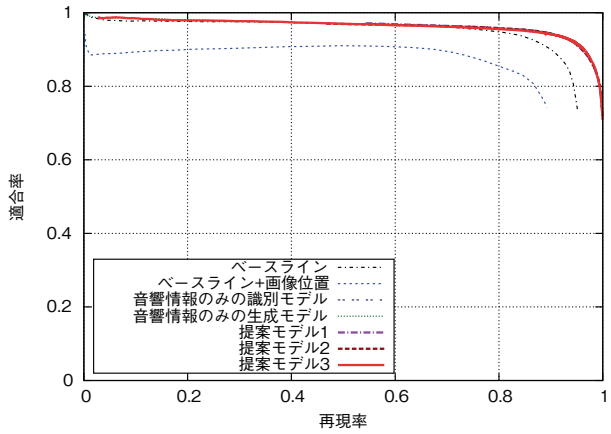

図 2 説明者に関する適合率-再現率曲線 (クリーン音声)

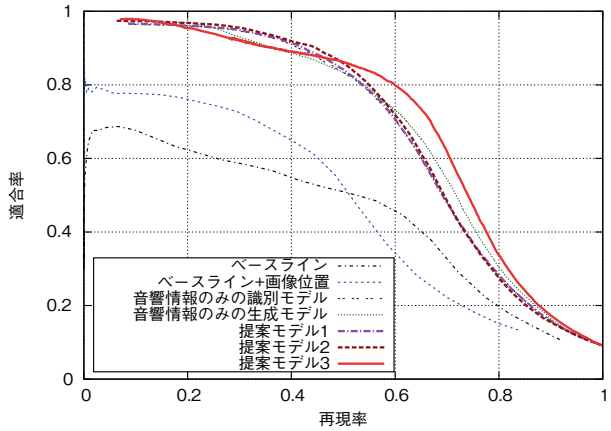

図 3 聴衆に関する適合率-再現率曲線 (クリーン音声)

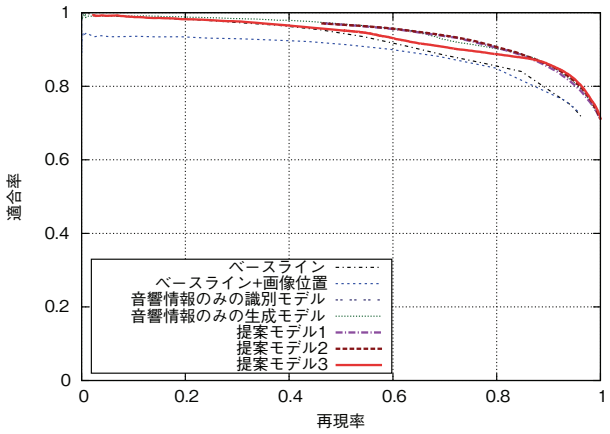

図 4 説明者に関する適合率-再現率曲線 (SNR = 0 dB)

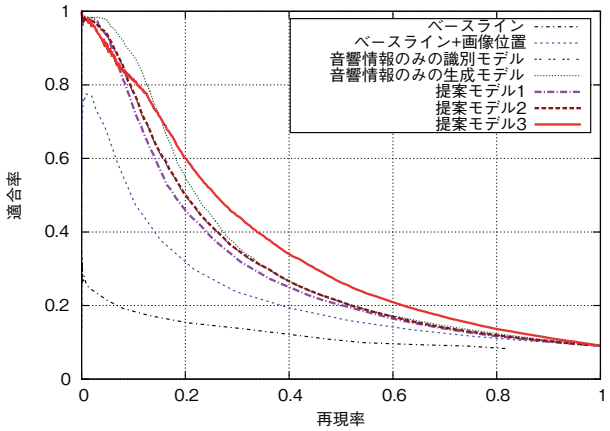

図 5 聴衆に関する適合率-再現率曲線 (SNR = 0 dB)

$$\text{適合率} = \frac{\#\text{TP}}{\#\text{TP} + \#\text{FP}}$$

$$\text{再現率} = \frac{\#\text{TP}}{\#\text{TP} + \#\text{FN}}$$

ただし，#TP は正解ラベルが発話のうち発話と正しく判定したフレーム数，#FP は正解ラベルが発話のうち非発話と誤って判定したフレーム数，#FN は正解ラベルが非発話のうち発話と誤って判定したフレーム数をそれぞれ表す．加えて，11 点補間平均適合率も計算した．これは 11 点の再現率 (0 から 1.0 までの 0.1 刻み) における補間適合率の平均値である．また，3 名の参与者 (説明者 1 名，聴衆 2 名) はそれぞれ個別に評価し，聴衆 2 名に関しては上記フレーム数を合算した．

雑音を重畳しない音声 (クリーン音声) における適合率-再現率曲線を，説明者について図 2，聴衆について図 3 に示す．また，雑音を付加した条件 (SNR=0 dB) における結果を同様に図 4 と図 5 に示す．各モデルにおいて，事後確率，尤度比，対数尤度差の各閾値を変化させ，これらの曲線の各点を得た．11 点補間平均適合率について，説明者を表 2，聴衆を表 3 に示す．これらは 8 セッションの結果の平均である．また，検出された発話区間及び非発話区間に対して，短い発話を非発話に，短い非発話を発話にするハングオーバー処理を施している．

上記の結果における視線方向は，Kinect センサから得た画像情報を用いた頭部方向推定によるものであり，磁気センサで計測したアノテーションデータとの平均誤差は位置が 12.2 mm，方向が 5.21° である．視線方向に関してアノテーションデータを用いた場合は，推定結果を用いた場合と比べて，11 点補間平均適合率で約 1〜3%高い結果となったが，自動推定による低下はほとんどないといえる．

5.3 考察

実験結果より，説明者の発話区間検出に関しては高い精度で実現できている．これは説明者がマイクロフォンアレイに対して比較的近くに位置し，セッションでの発話のほとんどを占めているためである．すべてのモデルが高い検出精度を示しており，これらの間に大きな差は見られなかった．

聴衆の発話区間検出は説明者の場合に比べて難しい．これは説明者とは逆の状況であることに起因しており，聴衆はマイクロフォンアレイから比較的離れた位置に立ち，発話時間が少ないためである．モデル 2(独立識別モデル) と音響情報のみの識別モデルでは大きな差が見られなかったが，モデル 3(雑音のある通信路モデル) は音響情報のみの

表 2 説明者に関する 11 点平均適合率

手法	SNR			
	クリーン	10 dB	5 dB	0 dB
ベースライン	0.876	0.874	0.874	0.835
ベースライン + 画像位置	0.745	0.821	0.821	0.808
音響情報のみの識別モデル	0.941	0.941	0.941	0.926
音響情報のみの生成モデル	0.946	0.949	0.949	0.933
提案モデル 1	0.942	0.940	0.940	0.925
提案モデル 2	0.941	0.941	0.941	0.925
提案モデル 3	0.946	0.947	0.947	0.923

表 3 聴衆に関する 11 点平均適合率

手法	SNR			
	クリーン	10 dB	5 dB	0 dB
ベースライン	0.448	0.312	0.314	0.177
ベースライン + 画像位置	0.453	0.366	0.366	0.252
音響情報のみの識別モデル	0.659	0.531	0.531	0.333
音響情報のみの生成モデル	0.672	0.548	0.547	0.352
提案モデル 1	0.662	0.526	0.526	0.322
提案モデル 2	0.666	0.535	0.535	0.334
提案モデル 3	0.691	0.577	0.579	0.377

生成モデルよりも高い検出精度を示した．また，モデル 3 による検出精度は，比較手法を含む他のすべての手法と比べて有意に高くなった．これらの結果から，モデル 3 において視線情報は有効であり，音響的雑音に対する頑健性を向上させることができた．全般に音響情報については，識別モデル (ロジスティック回帰) よりも生成モデル (混合ガウスモデル) の方が高い性能を得ている．

6. おわりに

本稿では，多人数会話での話者区間検出について，音響情報と視線情報を確率的に統合する 3 種類のモデルを提案した．実験結果より，雑音のある通信路モデルが最も高い精度を示し，音響情報のみを用いた比較手法に比べて，大きな精度改善がみられた．

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] Gatica-Perez, D.: Automatic nonverbal analysis of social interaction in small groups: A review, *Image and Vision Computing*, Vol. 27, No. 12, pp. 1775–1787 (2009).
[2] Otsuka, K.: Conversation Scene Analysis, *Signal Processing Magazine, IEEE*, Vol. 28, No. 4, pp. 127–131 (2011).
[3] 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認識，電子情報通信学会技術報告，SP2012-51, pp. 7–12 (2012).
[4] 河原達也：スマートポスターボード:ポスター発表における場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報告，PRMU2012-167, pp. 167–172 (2013).
[5] Kawahara, T.: Smart posterboard: Multi-modal sensing and analysis of poster conversations, *Proc. APSIPA ASC*, pp. 1–5 (2013).
[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，人工知能学会研究会，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] Tranter, S. E. and Reynolds, D. A.: An overview of automatic speaker diarization systems, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 5, pp. 1557–1565 (2006).
[10] 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. on Audio, Speech, and Language Processing*, Vol. 20, No. 2, pp. 371–381 (2012).
[11] Pedersen, M. S., Larsen, J., Kjems, U. and Parra, L. C.: A survey of convolutive blind source separation methods, *Multichannel Speech Processing Handbook*, pp. 1065–1084 (2007).
[12] Kendon, A.: Some functions of gaze-direction in social interaction, *Acta psychologica*, Vol. 26, No. 1, pp. 22–63 (1967).
[13] 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).
[14] 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，電子情報通信学会論文誌 A, Vol. J97-A, No. 6, pp. 453–468 (2014).
[15] 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).
[16] Schmidt, R.: Multiple emitter location and signal parameter estimation, *IEEE Trans. on Antennas and Propagation*, Vol. 34, No. 3, pp. 276–280 (1986).
[17] Asano, F., Goto, M., Itou, K. and Asoh, H.: Real-time sound source localization and separation system and its application to automatic speech recognition., *Proc. EUROSPEECH*, pp. 1013–1016 (2001).
[18] Wax, M. and Kailath, T.: Detection of signals by information theoretic criteria, *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 33, No. 2, pp. 387–392 (1985).
[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. on Fundamentals of Electronics, Communications and Computer Sciences*, Vol. 89, No. 8, pp. 2158–2165 (2006).
[20] Yoshimoto, H. and Nakamura, Y.: Cubistic Representation for Real-time 3D Shape and Pose Estimation of Unknown Rigid Object, *Proc. ICCV Workshop*, pp. 522–529 (2013).
[21] 若林佑幸，井上昂治，河原達也，中井駿介，宮崎亮一，猿渡洋：スマートポスターボードにおける音響情報と画像情報の統合による話者区間検出，日本音響学会 2014 春季研究会，2-Q4-7 (2014).