グループディスカッションに参加するロボットのための頭部動作モデルの検討

Consideration of the Head Movement Model for a Robot in Group Discussion Context

木村 清也[1] 黄 宏軒[2] 岡田 将吾[3]
大田 直樹[2] 桑原 和宏[2]
Seiya Kimura[1] Hung-Hsuan Huang[2] Shogo Okada[3]
Naoki Ohta[2] Kazuhiro Kuwabara[2]

[1] 立命館大学大学院情報理工学研究科
[1] College of Information Science and Engineering，Ritsumeikan University
[2] 立命館大学情報理工学部
[2] Graduate School of Information Science and Engineering，Ritsumeikan University
[3] 北陸先端科学技術大学院大学情報科学系
[3] School of Information Science，Japan Advanced Institute of Science and Technology

Abstract: コミュニケーション能力は，他者との意思疎通を円滑に行うために必要な能力である．そのため，コミュニケーション能力を評価するために就職採用面接にグループディスカッション行う企業が増えている． しかし，グループディスカッションを題材としたコミュニケーション能力の訓練基盤はいまだ確立されていない． これらの問題を解決するために，我々は現在グループディスカッションに参加可能なロボットの実現に向けてプロジェクトを進めている．その基礎研究として本論文では，グループディスカッションデータコーパスを集め，そのデータからアテンション対象，発声ターン，韻律，頭部動作のマルチモーダル特徴量を抽出しエージェントの頭部動作モデルの検討を行った．エージェントの頭部運動モデルはアテンション対象モデルと頷き区間モデルの 2 つに分けエージェント，発話している場合，発話を聞いている場合，参加者全員が発話してない場合の 3 つの状況ごとに検討した．モデルの学習は SVM(Support Vector Machine) によって行い，F-Measure がアテンション対象モデルでは 0.4〜0.6，頷き区間モデルでは 0.5〜0.6 の精度を得た．

1 はじめに

コミュニケーション能力は人が相手と意思や思考の伝達を円滑に進めるために必要な能力であり，企業がプロジェクトを実行している場合，プロジェクトメンバーのコミュニケーション能力は他のメンバーとの人間関係に大きく影響し，その結果としてチームのパフォーマンスにも大きな影響を与えることになる．そのため，就職採用選考ではグループディスカッションを取り入れる企業も存在し，就職活動者は割り当てられた課題について他の参加者と議論する．そして面接官はその議論のプロセスから就職活動者のコミュニケーション能力の評価を行う．したがって，コミュニケーション能力を高めることは就職活動に成功する可能性を高めることにつながると考えられる．またコミュニケーション能力は反復練習によって向上すると考えられているが，反復練習には相応のパートナーが必要であり，そのパートナーを用意することは多くの学生にとって簡単なことではない．EU が設立したプロジェクトである TARDIS では就職面接の訓練のために人工的なパートナーを作成する研究が行われており，1 対 1 の面接訓練のための仮想エージェントの開発が行われている [5, 6, 3]．そして我々は現在，仮想エージェントやロボットとグループディスカッションを行うための訓練システムの開発を目指している．複数人会話では話し手と聴き手だけが存在する 1 対 1 の対話とは異なり，対話参加者間の役割の区別や，会話の主導権の切り替わりが複雑であるため，グループディスカッションに参加できるエージェントを実現するためには 1 対 1 の対話のためだけに設計された会話エージェントよりも多くの機能を組み込む必要がある．

この論文ではグループディスカッションに参加可能な

エージェントを実現するために必要な機能の一部として，頭部動作に関する 2 つのモデルを検討する．1 つ目はエージェントが議論中に他の参加者に対して適切な方向に注意を向けるアテンション対象モデルである．エージェントがグループディスカッション中に適切なアテンション対象を決定するためのモデルである．ここでは，注視方向と頭部方向の組み合わせをアテンションとして扱い，エージェントがアテンションを向ける目標（別の参加者または議論に使用されている資料）に着目する．2 つ目はエージェントが議論中に適切なタイミングで頷きを行う頷き区間モデルである．このモデルでは参加者の頷きの方法や速さ，角度に関係なく頷きという行動そのものに着目し，エージェントがグループディスカッションの中で頷きを行うべき区間を判断するモデルを生成する．[13] で言及された ”モナリザ効果”のために，ユーザーは，2D で表面的にレンダリングされるグラフィックエージェントの注視方向または頭部方向を実際には正確に判断することができない．したがって，この研究ではエージェントは仮想空間で 3D にレンダリングされたロボットあるいは物理的に体を持ったロボットに実装されることを想定する．また，この論文ではエージェントが発言しているとき，エージェント以外が発言しているとき，誰も発言していない時の 3 つのシチュエーションに分けて各モデルを生成する，人間が実際にグループディスカッション中に行う非言語行動から特徴量を抽出し，アテンション対象および頷き区間の予測モデルを生成する．またこの研究は自らが収集したデータコーパスに基づいている，そのコーパスは 4 人の参加者 10 グループ（各グループ 15 分）のグループディスカッションのビデオ/オーディオデータの他にモーションキャプチャなどのセンサーデータで構成されている．本稿は以下のように構成されている．第 2 章では関連する研究について紹介し， 第 3 章で用いたコーパスについての解説を行う．第 4 章ではアテンション対象モデルと頷き区間モデルの定義について述べ，第 5 章では，マルチモーダル特徴と機械学習技術を用いたモデルを生成および評価を行い， 最後に 6 章で本稿のまとめを行う．

2 関連研究

複数人会話に着目した関連研究として小山ら [2] は，複数人会話における参加者の振り向き，および発話タイミングを分析し，エージェントが複数人会話に参加した場合の聞き手としての振る舞いを検討した．このほかにも複数のユーザーとエージェントとのインタラクションについての研究もいくつか存在するが，その多くは人間のユーザーとはエージェントは異なる立場として扱われている [4，7]．一方我々の研究ではエージェントやロボットを他の人間の参加者と同様の立場としてディスカッションに参加させることに焦点を当てているため，それ専用の動作モデルが必要となる．また，頷きに着目した研究としてインタラクションにおいて頷きが相手に与える影響の大きさを分析した研究 [1] がある．この研究では頷きによる相槌は発話による頷きよりも対話に集中していることが伝えられることを示した．グループディスカッションを題材にした研究では岡田ら [8] が言語特徴量および非言語特徴量（音声韻律，発話ターン，頭部活動）を含むマルチモーダル特徴を用いてコミュニケーション能力のスコアを推定する回帰モデルを構築した．このほかに，Schiavo ら [9] はグループメンバーの非言語行動を観察することによって，参加型のディスプレイを通じて自動的に参加者に指示を与え，グループ活動のコミュニケーションの流れをサポートするシステムを提案した．

3 対話実験データの収集

3.1 実験概要

グループディスカッションに参加可能なエージェントを開発するために，人間同士が行うグループディスカッションのデータコーパスを収集する実験を行った．実験手順は，[11] に従ったがメガネ型アイトラッカーは使用しなかった．ディスカッションの議題は，日本企業の就職採用面接において頻繁に使用される「サバイバルタスク」型の課題を設定した．この課題は参加者に複数の選択肢が与えられ，その選択肢に対して提示された条件に基づいて順位づけを行う課題である．この課題では参加者の優先される選択肢の順序を論理的かつ明確に示すの能力を観察することができる．今回の実験で参加者が実際に行う議題は配布した資料の中に記載された 15 名の有名人の中から収益や集客を考慮し，最適だ と思われる人物の順位をつけていく「学園祭に招待する有名 人ランキング」を設定した．また議論が活発となるよう，議論に取り組む前にそれぞれの参加者単独で課題について考える時間を 5 分間設け，その後 15 分のディスカッションを行った．実験データは互いに顔見知りでない学生 4 人を 1 グループとした 10 グループの合計 40 人分のデータによって構成されている．また，就職活動経験者，就職活動中の学生に実践的なグループディスカッションを行ってもらうため，大学院生以下， 大学 3 年生以上の学生を実験参加者の対象とした．実験を実施にあたり男女の比率の違いによる発言の優劣をなくすため，グループ内の男女の割合は，すべて同性か，異性の数が等しくなるように設定した．

図 1: 実験環境

3.2 実験環境

実験参加者が，1.2m × 1.2m の正方形のテーブルの周りに座って議論を行う様子を， 2 つのビデオカメラと様々なセンサを使用して記録した．各参加者は，頭に加速度センサー（ATR-Promotions TSND121）とオーディオデジタイザ（Roland Sonar X1 LE）に接続されたヘッドセットマイク（Audio-Technica HYP-190H）を装着し，ウェブカメラ（Logicool C920）を 4 台使用し各参加者の顔を撮影した．また，モーションキャプチャ（OptiTrack Flex 3 with 8 台のカメラ）と Microsoft Kinect センサーを使用して，参加者の上半身の動きを記録した．記録実験の設定を図 1 に示す．実験の結果，15 分× 10 グループ= 150 分間のグループディスカッションデータが記録された． Praat[1] による分析の結果，このコーパスの統計情報は以下のようになった．各セッションで平均 767.1 発話（最大 913，最小 570，標準偏差 101.7）発声の長さは 0.898 秒（最大 10.6，最小 0.2，標準偏差 0.868）

4 頭部動作モデル

4.1 アテンション対象モデル

この研究ではエージェントの視線方向や頭部の向きのとの全体的な組み合わせをアテンション対象として扱う．これは人工的なエージェントはアクチュエータが人間と同じように機能しないことに起因する．特にロボットの場合，人間よりも自由度がはるかに低く，人間のようにスムーズに動く可動式の眼球もない．そのため我々は詳細に人間の動きを模倣したものより，より単純で抽象的かつ装置に依存しないアテンション対象モデルの生成を今回の研究課題とした．今回我々が行ったグループディスカッションデータ収集実験における参加者の注目可能なターゲットは主に他の参加者と配布した資料 (今回の実験では有名人リスト) とみなすことができる．各参加者の視点から見ると左側に座っている参加者，テーブルの反対側に座っている参加者，右側に座っている参加者の 3 人がいる．それらの参加者をそれぞれ *left*, *front*, *right* とし，アテンション対象として定義する．また，資料がテーブル上に置かれており参加者の視線がテーブル方向に向いていた場合，テーブルを見ているのか資料を見ているのは判別が困難であるため，それらを一様に *table* として定義する．これらの定義にアノテーションツール Elan [12] を使用してビデオコーパスによってすべての参加者のアテンション対象を手動でアノテーションを行った．また，アノテーションを行う者がアテンション対象を判断できなかった場合のクラスである *away* を定義した．
表 1 にアノテーションの結果を示す．結果から参加者は *table* へ最も頻繁かつ最も長くアテンションを向けていることが観察された．これは参加者が議論をしている間，有名人リストを見ている時間が長くなっていたことが要因としてあげられる．また，他の参加者へ向けられたアテンションの中でもテーブルの反対側に座っている参加者へ頻繁にアテンションを向けいるが，左右の参加者での明らかな違いは認められなかった．そして *away* クラスは他のクラスに比べてインスタンス数がはるかに少ないという結果になった．

表 1: 全参加者のアテンション対象ラベルデータ

クラス	個数	平均継続時間	最大継続時間	最小継続時間
table	1715	17.7 秒	359 秒	0.38 秒
front	1075	2.6 秒	31 秒	0.18 秒
right	662	2.3 秒	30 秒	0.18 秒
left	642	2.1 秒	25 秒	0.11 秒
away	4	1.8 秒	4 秒	0.10 秒

4.2 頷き区間モデル

頷きはそれを行う者によって多様な表現方法があり，速さや角度などについても様々である．また，アテンション対象モデルと同様のエージェントの自由度やスムーズによる問題から頷き区間モデルにおいても装置に依存しないモデルの生成を行う．そのためこの頷き区間モデルでは頷きの方法や速さ，角度に関わらずエージェントがグループディスカッションの中で頷きを行う

[1] http://www.fon.hum.uva.nl/praat/

べき区間を判断するモデルを生成する．参加者が頷き始めた時点から頷き終わるまでの区間を頷き区間とし，連続で数度頷いた場合は最初の頷きから最後の頷きが終わるまでを一つの頷き区間としてこれを *nod* として定義した．そしてそれ以外の区間を非頷き区間としてこれを *nomal* として定義した．これらのて意義をもとに Elan を使用してビデオコーパスによってすべての参加者の頷き区間と非頷き区間を手動でアノテーションを行った．表 2 にアノテーションの結果を示す．頷き区間の最大継続時間と最小継続時間には大きな差があり，頷きの多様性がうかがえる結果となっている．

表 2: 全参加者の頷きラベルデータ

クラス	個数	平均継続時間	最大継続時間	最小継続時間
nomal	655	51.6 秒	609.5 秒	0.3 秒
nod	615	1.3 秒	9.9 秒	0.2 秒

5　アテンション対象モデルの生成

5.1　シチュエーション

エージェントがグループディスカッションに参加した場合，話し手やあるいは聴き手など一つの役割を担いつつ議論に参加することになる．そこでロボットがグループディスカッションに参加した場合に起こりうる 3 種類のシチュエーションを想定しモデルの作成する．

- Speaking モデル：エージェントが発話をしている
- Listening モデル：他の参加者の発話を聴いている
- Idling モデル：参加者全員が発話をしていない

モデルのトレーニングを行うにあたり，4 人の参加者のうち 1 人の参加者を対象として「センター参加者」として定義する．そして実験環境のテーブルは正方形であるため，グループ内の 4 人の参加者は座っている位置に関係なく全て同等に扱うことができる．したがって全ての参加者をセンター参加者のとして見なすことができるため 600 分（15 分× 10 グループ× 4 人）すなわち 10 時間分のデータを頭部動作モデルのトレーニングデータとして使用することができる．前章で収集したアテンション対象についてのアノテーションデータは各参加者からの視点でつけられているため，参加者のアテンション対象と頭部動作の関係を抽出するためには各参加者から見た絶対的なアテンション対象からセンター参加者から見た相対的なアテンション対象

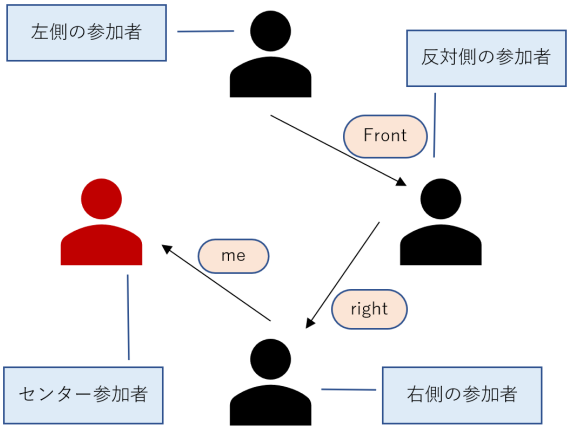

図 2: アテンション対象ラベルの変換例

への変換が必要となる．図 2 にはアテンション対象ラベルの変換例を示す．左側の参加者はセンター参加者から見てテーブルの反対側の参加者にアテンションを向けているため，*front* に変換する．また，例の右側の参加者のようにアテンションをセンター参加者に向けている場合のクラス *me* を新たに追加する．

5.2　特徴量抽出

エージェントのアテンション対象と頷き区間を決定する予測モデルを生成するために，グループディスカッション参加者の非言語行動からマルチモーダルな特徴量を抽出した．その特徴量によって作成したデータセットを用いてセンター参加者 (エージェント) のモデルのトレーニングを行う．今回の研究ではアテンション対象と頷き区間の 2 種類のモデルが存在するが両者とも同じ特徴量を用いてトレーニングを行う．また，アテンション対象モデルについては *away* クラスのインスタンスが非常に少ないためこのクラスを省略し，*left*, *front*, *right*, *table* の 4 つのクラスの予測モデルをトレーニングする．今回の研究では特徴量を,「アテンション」,「発話ターン」,「韻律」,「動作」の 4 つのモダリティに分け，センター参加者以外の参加者の行動から 0.1 秒ごとに抽出する．また特徴量によってはスライディングウィンドウを用いて抽出し，ウィンドウサイズ t は 1〜10 秒の間で 1 秒ごとに変化させる．特徴量は全部で 92 個存在し，その詳細は以下で説明する．

- アテンション (A)：他の 3 人の参加者がアテンションを向けている方向に関する特徴量．センター参加者以外の参加者から 5 つの特徴量の合計 3 x 5 = 15 個が抽出される．
  - 現在のアテンション対象

  - 議論開始時からセンター参加者にアテンションを向けていた時間の割合
  - 過去 t 秒の間でセンター参加者にアテンションを向けていた時間の割合 (t はウィンドウサイズ)
  - 議論開始時からアテンション対象を変更した回数
  - 過去 t 秒の間でアテンション対象を変更した回数 (t はウィンドウサイズ)
- 発話ターン (S):参加者の音声記録から音声解析ツール Praat を用いて有音と無音の識別を行い発話区間データとして収集．センター参加者以外の参加者から 7 つの特徴量と以下に太字で示されている，参加者からではなくグループ全体から抽出する特徴量 2 つの合計 3 x 7 + 2 = 23 個が抽出される．
  - 現在発話を行っているかそうでないか
  - 議論開始時から発話を行った回数
  - 過去 t 秒の間で発話を行った回数 (t はウィンドウサイズ)
  - 議論開始時から発話を行っていた時間の割合
  - 過去 t 秒の間で発話を行っていた時間の割合 (t はウィンドウサイズ)
  - 発話を行っていた場合，その発話の継続時間
  - 発話の平均継続時間
  - **Speaking，Listening，Idling の各シチュエーションに移ってからの継続時間**
  - **最後に発話を行った人物**
- 韻律 (P) : 参加者が発話を行っている間の韻律情報を Praat によって収集した．センター参加者以外の参加者から次の 12 個の特徴量合計 3 x 12 = 36 個を抽出した．
  - 現在のピッチ
  - 議論開始時からのピッチの標準偏差
  - 過去 t 秒のピッチの標準偏差 (t はウィンドウサイズ)
  - 議論開始時から発話を行っていた時間の割合
  - 過去 t 秒の間のピッチの平均値 (t はウィンドウサイズ)
  - 議論開始時からの平均ピッチと現在のピッチとの差 (t はウィンドウサイズ)
  - 過去 t 秒の間のピッチの平均値と現在のピッチとの差 (t はウィンドウサイズ)
  - 現在のインテンシティ
  - 議論開始時からのインテンシティの標準偏差
  - 過去 t 秒のインテンシティの標準偏差 (t はウィンドウサイズ)
  - 議論開始時から発話を行っていた時間の割合
  - 過去 t 秒の間のインテンシティの平均値 (t はウィンドウサイズ)
  - 議論開始時からの平均インテンシティと現在のインテンシティとの差
  - 過去 t 秒の間のインテンシティの平均値と現在のインテンシティとの差 (t はウィンドウサイズ)
- 動作 (B) : 各参加者の頭に取り付けられた 3 軸加速度センサーから頭部運動量に関する情報を収集し，センター参加者以外の参加者から 6 個の特徴量合計 3 x 6 = 18 個を抽出した．
  - 最新の 0.1 秒の間で計測された活動量
  - 議論開始時から活動量の標準偏差
  - 過去 t 秒の間の活動量の標準偏差 (t はウィンドウサイズ)
  - 議論開始時から活動量の平均値と最新の 0.1 秒の間で計測された活動量との差
  - 過去 t 秒の間の活動量の平均値 (t はウィンドウサイズ)
  - 過去 t 秒の間の活動量の平均値と最新の 0.1 秒の間で計測されたの活動量との差 (t はウィンドウサイズ)

5.3 予測モデル

非線形 SVM（Support Vector Machine）を用いて 2 つのモデル 3 つの状況の予測モデルを生成した． SVM のコストパラメータ C は，0.5，1,5，10，15 の値の中から，RBF カーネルのパラメータ γ は，0.001,0.01，0.1，1 の値の中からパラメータのすべての組み合わせでテストを行い， その結果 C = 10.0 および γ = 0.01 を最良のパラメータとして今回のモデル生成に使用する．アテンション対象モデルでは *table* クラスおよび *front* クラスのインスタンス数が多かったため，すべてのクラスのインスタンスの数が最もインスタンス数が少ないクラスと同数となるよう調整した．また，頷き区間モデルにおいても *nomal* クラスのインスタンス数が圧倒的多いため，*nod* クラスのインスタンス数と同数になるよう調整した．評価には Leave-one-person-out 方を用いた． つまり，39 人分参加者のデータはトレーニングに使用され，1 人の参加者のデータはテストデータとして使用する． この手順を 40 回繰り返し，すべての参加者のデータでテストを行い，その結果の合計を

モデルの精度として評価する．
図 3，図 4，図 5 は，アテンション対象モデルのそれぞれ 1 秒から 10 秒で Speaking，Listening，および Idling のシチュエーションでウィンドウサイズを変化させたときの各モダリティの分類精度を示している．結果から，ウィンドウサイズによってパフォーマンスに大きな差は現れなかったが， すべてのモダリティを使用するモデルのパフォーマンスは，単一のモダリティのみを使用するよりも常に優れていることがわかる． 動作特徴量（B）は常に最も低い精度となった．一方図 6，図 7，図 8 では頷き区間モデルのそれぞれ 1 秒から 10 秒で Speaking，Listening，および Idling のシチュエーションでウィンドウサイズを変化させたときの各モダリティの分類精度を示している．結果から，ウィンドウサイズによってパフォーマンスに大きな差がないことはアテンション対象モデルと同様だが，シチュエーションによって最も精度が高いモダリティあるいは最も精度が低いモダリティが異なっている．

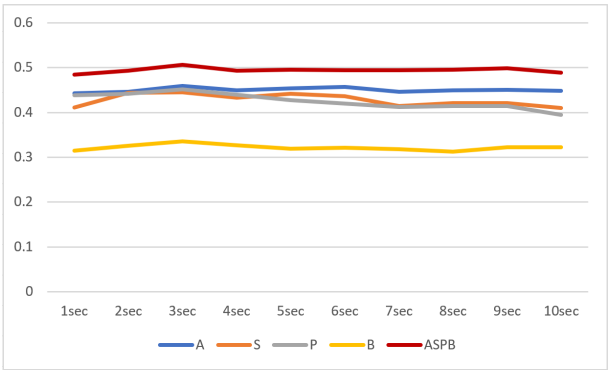

図 3: アテンション対象モデルにおける Idling の場合にウィンドウサイズを 1〜10 秒で変化させたときのモダリティごとの F-measure(縦軸)

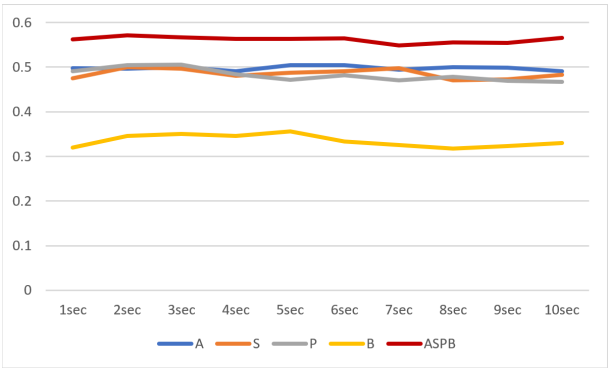

図 4: アテンション対象モデルにおける Listening の場合にウィンドウサイズを 1〜10 秒で変化させたときのモダリティごとの F-measure(縦軸)

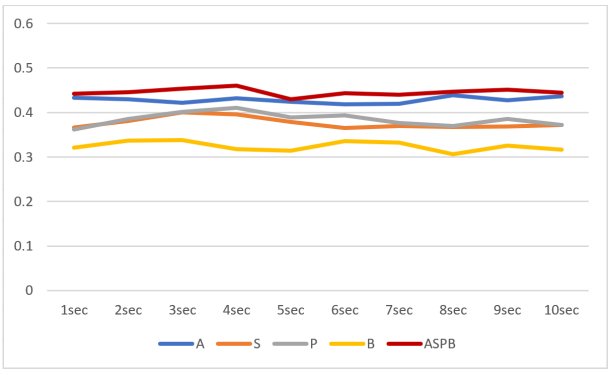

図 5: アテンション対象モデルにおける Speaking の場合にウィンドウサイズを 1〜10 秒で変化させたときのモダリティごとの F-measure(縦軸)

特に Idling の場合アテンション対象モデルでは常に最も精度が低かった動作特徴量 (B) が常に最も精度が高いモダリティとなっている．これによって生成するモデルの目的によって重要なモダリティが変わることが示唆された．表 3 は，それぞれのモデルと状況によって最も F-Measure が高かった最適なウィンドウサイズを示している．
これらの特徴量モダリティごとの最適なウィンドウサイズを使用し，それぞれのモデル精度への貢献度を検証するために， 15 種類の組み合わせで特徴量セットを作成しモデルの生成を行った．図 9，図 10 にその結果を示す．この結果から特徴量のモダリティが豊富であれば必ずしも F-Measure が高くなるというわけではないことがわかる．特にアテンション対象モデルでは A と S の貢献度が高く，P と B の貢献度が低い．これは，エージェントの適切なアテンション対象を決定する際には，他の参加者のアテンションと発話ターンが重要な役割を果たすことを意味する．また，全体的にアテンション対象をモデルの F-Measure はおよそ 0.4〜0.6 で頷き区間モデルの F-Measure はおよそ 0.5〜0.6 となっており頷き区間モデルの方が高い数値が結果として出ている．しかし，アテンション対象モデルは 4 クラスの分類問題である一方で頷き区間モデルは 2 クラスの分類問題なため，必ずしも後者の方が性能がよいとは言えない．このことからアテンション対象モデルとは反対にエージェントの適切な頷き区間を決定する際には他の参加者の非言語行動はあまり重要ではない可能性が高いことが示された． 表 4，表 5 には，それぞれのモデルにおける最適なウィンドウサイズですべての特徴量モダリティを使用して生成したモデルの，Precision，Recall および F-measure の詳細な数値である．アテンション対象モデルでは 3 つのシチュエーション全てにおいて *table* と *front* クラスの分類精度が他の 2 つのクラスに比べて低かった．これは参加者が *table* と *front* に向いている時間が長いためその他 2 つの方向よりも特徴的

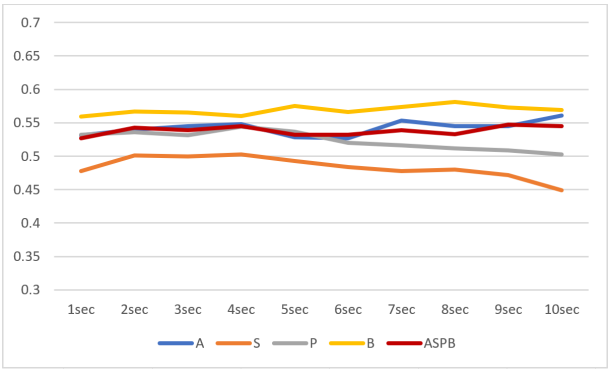

図 6: 頷き区間モデルにおける Idling の場合にウィンドウサイズを 1〜10 秒で変化させたときのモダリティごとの F-measure(縦軸)

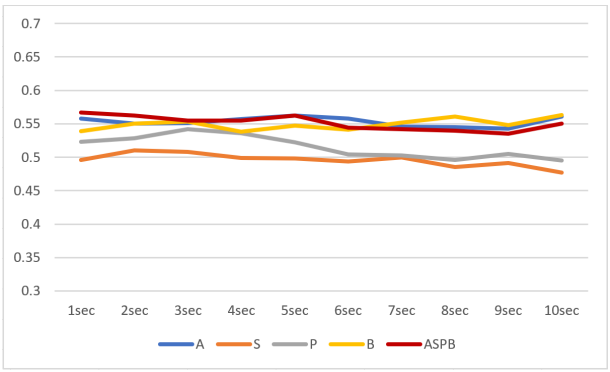

図 7: 頷き区間モデルにおける Listening の場合にウィンドウサイズを 1〜10 秒で変化させたときのモダリティごとの F-measure(縦軸)

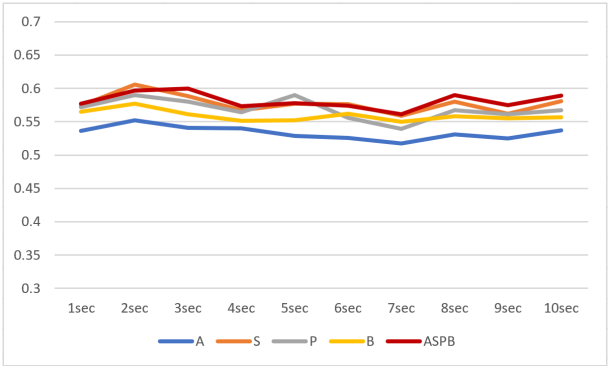

図 8: 頷き区間モデルにおける Speaking の場合にウィンドウサイズを 1〜10 秒で変化させたときのモダリティごとの F-measure(縦軸)

表 3: 各モデルおけるモダリティごとの最適なウィンドウサイズ (数字の単位は秒)

モデル	シチュエーション	A	S	P	B
アテンション対象	Idling	3	3	3	3
	Listening	6	2	3	5
	Speaking	8	3	4	3
頷き区間	Idling	10	6	4	8
	Listening	6	2	3	10
	Speaking	2	2	2	2

表 4: アテンション対象モデルにおける最適なウィンドウサイズですべての特徴量モダリティを使用して生成したモデルの分類結果

	クラス	Precision	Recall	F-measure
Idling	table	0.413	0.572	0.480
	front	0.501	0.393	0.441
	right	0.553	0.502	0.526
	left	0.583	0.536	0.559
	平均	0.512	0.501	0.507
Listening	table	0.422	0.432	0.427
	front	0.566	0.521	0.542
	right	0.621	0.661	0.641
	left	0.655	0.649	0.652
	平均	0.566	0.566	0.566
Speaking	table	0.412	0.588	0.484
	front	0.401	0.301	0.344
	right	0.462	0.419	0.439
	left	0.538	0.494	0.515
	平均	0.453	0.450	0.452

でないことをが要因としてあげられる．そしてアテンション対象モデル，Listening>Idling>Speaking の順で精度が高いことに対し，頷き区間モデルでは Speaking の場合が最も精度が高くなった． これは，センター参加者が発話をしているとき，アテンション対象は他の参加者の非言語行動に影響されにくく，頷きは他のシチュエーションに対して比較的影響されやすいということが示された．

5.4 考察

5 章では 2 つのモデルで 3 つのシチュエーションに分け，ウィンドウサイズと特徴量セットの両面から性能の評価をした．結果的にアテンション対象モデルにおいてはチャンスレベル（25 %）よりも大幅に優れており，グループディスカッションの参加者のアテンション対象には傾向があることを示された一方で，頷き区間モデルでは，チャンスレベル（50 %）とあまり変わ

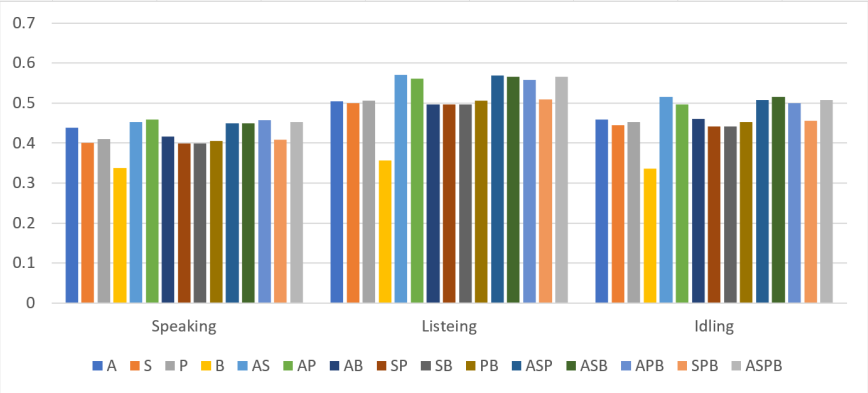

図 9: アテンション対象モデルにおけるモダリティごとに最適なウィンドウサイズを使用した特徴量セットごとの F-measure（縦軸）

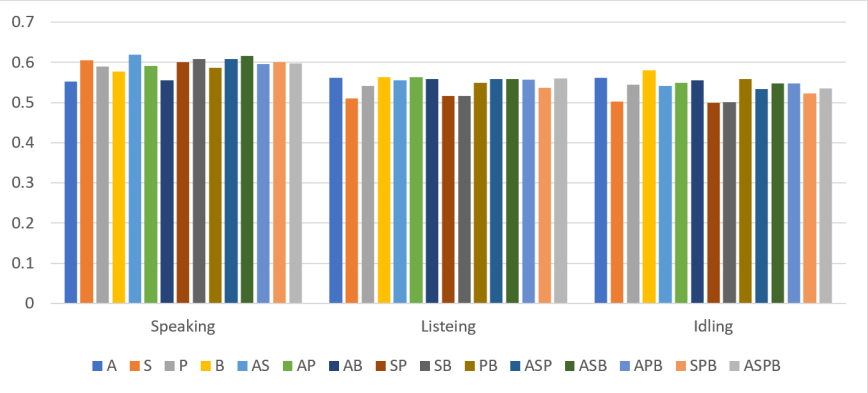

図 10: 頷き区間モデルにおけるモダリティごとに最適なウィンドウサイズを使用した特徴量セットごとの F-measure（縦軸）

らないという結果となった．しかし，特徴量を充実させることで，両モデル共に精度が向上する可能性がある．特に今回の研究では言語行動に関する情報が使用されていなかったため，それらを特徴量に組み込むことで，性能を改善することができると考えられる．

表 5: 頷き区間モデルにおける最適なウィンドウサイズですべての特徴量モダリティを使用して生成したモデルの分類結果

	クラス	Precision	Recall	F-measure
Idling	nomal	0.531	0.622	0.572
	nod	0.543	0.450	0.492
	平均	0.537	0.536	0.536
Listening	nomal	0.548	0.666	0.601
	nod	0.574	0.451	0.505
	平均	0.566	0.566	0.566
Speaking	nomal	0.585	0.666	0.623
	nod	0.612	0.527	0.566
	平均	0.598	0.596	0.597

6 終わりに

本論文では，グループディスカッションに参加できるエージェントの実現に向けて，アテンション対象モデルと頷き区間モデルの検討をした．対話収集実験からグループディスカッション参加者の非言語行動を，アテンション対象，発話ターン，韻律，動作，の 4 つモダリティで特徴量を抽出し，「エージェントが発話をしている場合」「他の参加者のが発話をしている場合」「参加者全員が発話をしていない場合」の 3 種類のシチュエーションに分け SVM を用いてモデルの生成を行った．結果はアテンション対象と頷き区間モデルの F-Measure がそれぞれ 0.4〜0.6 程度と 0.5〜0.6 程度となり，参加者のアテンション対象についてはグループディスカッション参加者のアテンション対象は他の参加者の非言語行動による影響があることが示された一方で，頷きについては他の参加者の非言語行動による影響は小さいことが示された．今後は，MFCC のような詳細な韻律情報，あるいは言語行動に関する特徴量などを追加し，両モデルの性能向上を図りたいと考えている．モデル

の分類精度向上に加えてコーパス自体，特に参加者の議論中の役割についても調査をする予定である．エージェントがグループディスカッションに参加した場合，ディスカッションへの貢献度や進捗に影響を及ぼす様々な役割を果たすことができるはずである．したがってこれについて調査し，議論中の役割にも関連づけた頭部動作モデルの開発をしたいと考えてる．最後に，コミュニケーションロボットや VR 環境の仮想エージェントにモデルを実装し， そのエージェントを用いた実験についても進めていく予定である．

参考文献

[1] Catharine Oertel, Jose Lopes, Yu Yu, Kenneth A. Funes Mora, Joakim Gustafson, Alan W. Black, and Jean-Marc Odobez ： Towards building an attentive artificial listener:on the perception of attentiveness in audio-visual feedback tokens,In 18th ACM International Conference on Multimodal Interaction, pp. 21-28(2016)

[2] 小山大幾, 水本武志, 中村圭佑, 中臺一博, 今井倫太：複数人会話における振り向き動作と発話動作解析, HAI シンポジウム 2015, pp. 256-261 (2015)

[3] Tobias Baur, Ionut Damian, Patrick Gebhard, Kaska Porayska-Pomsta, and Elisabeth Andre : A Job Interview Simulation: Social Cue-based Interaction with a Virtual Character, In 2013 International Conference on Social Computing, pp. 220-227 (2013)

[4] Iolanda Leite, Marissa McCoy, Monika Lohani, Daniel Ullman, Nicole Salomons, Charlene Stokes, Susan Rivers, and Brian Scassellati:Emotional Storytelling in the Classroom: Individual versus Group Interaction between Children and Robots. In 10th ACM/IEEE International Conference on Human-Robot Interaction, pp. 75-82 (2015)

[5] Mathieu Chollet, Magalie Ochs, and Catherine Pelachaud:Mining a Multimodal Corpus for Non-Verbal Signals Sequences Conveying Attitudes. In The 9th edition of the Language Resources and Evaluation Conference, pp. 3417-3424 (2014)

[6] Hazael Jones, Mathieu Chollet, Magalie Ochs, Nicolas Sabouret, and Catherine Pelachaud : Expressing social attitudes in virtual agents for social coaching, Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems, pp. 1409-1410 (2014)

[7] Marynel Vazquez, Elizabeth J. Carter, Braden McDorman, Jodi Forlizzi Aaron Steinfeld, and Scott E. Hudson : Towards Robot Autonomy in Group Conversations: Understanding the Effects of Body Orientation and Gaze, Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, pp. 42-52 (2017)

[8] Shogo Okada, Yukiko Nakano, Yuki Hayashi, Yutaka Takase, and Katsumi Nitta : Estimating Communication Skills using Dialogue Acts and Nonverbal Features in Multiple Discussion Datasets, In 18th ACM International Conference on Multimodal Interaction, pp169-176. (2016)

[9] Gianluca Schiavo, Alessandro Cappelletti, Eleonora Mencarini, Oliviero Stock, and Massimo Zancanaro ： Overt or Subtle? Supporting Group Conversations with Automatically Targeted Directives, In Proceedings of the 19th international conference on Intelligent User Interfaces, pp. 225-234 (2014)

[10] Catharine Oertel, Jose Lopes, Yu Yu, Kenneth A. Funes Mora, Joakim Gustafson, Alan W. Black, and Jean-Marc Odobez. 2016. Towards building an attentive artificial listener: on the perception of attentiveness in audio-visual feedback tokens. In 18th ACM International Conference on Multimodal Interaction, pp. 21-28 (2016)

[11] Fumio Nihei, Yukiko I. Nakano, Yuki Hayashi, Hung-Hsuan Huang, and Shogo Okada : Predicting Influential Statements in Group Discussions using Speech and Head Motion Information. In Proceedings of the 16th International Conference on Multimodal Interaction, pp. 136-143 (2014)

[12] Hedda Lausberg and Han Sloetjes, Coding gestural behavior with the NEUROGES-ELAN system, Behavior Research Methods 41, pp. 841-849 (2009)

[13] Osamu Morikawa and Takanori Maesako, HyperMirror : Toward Pleasant-to-use Video Mediated Communication System, In Proceedings of the 1998 ACM conference on Computer supported cooperative work, pp. 149-158 (1998)