マルチモーダル情報に基づく聞き手の理解度推定

令和5年度　修士論文

日本大学大学院　総合基礎科学研究科
地球情報数理科学専攻　宮田研究室

6122M15　木下　峻一

概要

日常生活や社会活動における対話において，対話相手が自分の話を理解しているか否かを推測することが重要である．これは，人と人の対話だけでなく，人とエージェントの対話においても重要である．対話相手の理解度を自動的に予測することができれば，対話エージェントはユーザの理解度に応じたコミュニケーションを行うことが可能となる．ユーザの内部情報を推定する対話エージェントを開発するために，対話参加者のマルチモーダル情報から対話相手の内部情報を推定する研究事例は数多く存在する．しかし，ユーザのマルチモーダル情報から対話における聞き手の理解度を推測することができるのかは明らかではない．そのため，ユーザの理解度を考慮してコミュニケーションを行う対話エージェントを開発することが困難であると考えられる．本研究では，理解度を聞き手が話し手の発話を理解している様子の度合と定義し，対話参加者のマルチモーダル情報から聞き手の理解度を推測できるのか明らかにする取り組みを行う．具体的には，はじめに対面の2者対話データと第三者による聞き手の理解度の評価が記録されている対話コーパスから対話参加者の視覚的，韻律的，言語的特徴量を抽出する．次に対話コーパスから抽出した特徴量を用いて聞き手の発話時の理解度を推定する機械学習モデルを構築した．このとき，聞き手の発話時の理解度を推定する二種類のモデルを作成した．1つ目は，聞き手の視覚的，韻律的，言語的特徴量から，聞き手の発話時の理解度を回帰タスクで推定する機械学習モデルである．2つ目は，聞き手と話し手の双方の視覚的，韻律的，言語的特徴から，聞き手の発話を理解をしている発話，中立の発話，理解をしていない発話に分類する機械学習モデルである．回帰モデル構築の結果，聞き手の視覚的，韻律的，言語的特徴が聞き手の発話時の理解度の推定に有効であることが確認された．クラス分類モデル構築の結果，聞き手と話し手の視覚的，韻律的，言語的特徴が聞き手の発話を理解をしている発話，中立の発話，理解をしていない発話の分類に有効であることが確認された．さらに，聞き手の視覚的，韻律的，言語的情報と，話し手の情報を組み合わせることが推定性能を向上するためには重要であることが示唆された．

本研究の貢献は下記の通りである．

- 聞き手のマルチモーダル情報が聞き手の発話時の理解度の推定に有効であること．
- 聞き手と話し手のマルチモーダル情報が聞き手の発話を理解をしている発話，中立の発話，理解をしていない発話の分類に有効であること．

目 次

図 目 次

表 目 次

第1章　序論

1.1 研究の背景

日常生活や社会活動での対話において対話相手（聞き手）が自分（話し手）の話を理解しているか否かを推測することが重要である．対話では，話し手が聞き手の理解度を判断し，次に行うアクションを決めている．話し手が自分の話を聞き手が理解していると推測した場合，話し手は次の話に移る，関連する話題に発展させるなどのアクションを取る．一方，話し手が自分の話を聞き手が理解していないと推測した場合，話し手は同じ話題を繰り返す，発話内容を単純化するなどのアクションを取る．

聞き手の理解を推測するという行為は，人間と人間のインタラクションだけでなく，人間とエージェントのインタラクションでも重要であると考える．近年，対話エージェントとコミュニケーションを行う機会は増加している．対話エージェントは，買い物シーン，学習シーン，医療現場といった，さまざまな日常生活のシーンで導入されている [1]．対話エージェントが人間のように対話相手であるユーザの理解度を推定することができるならば，より自然なコミュニケーションを取れるようになることが期待される．反対に，対話エージェントがユーザの理解度を推定することができないことで，人と様々なミスコミュニケーションが生じてしまう可能性がある．

1.2 研究の目的

理解度を考慮したコミュニケーションを行う対話エージェントの開発を実現するためには，対話における聞き手の理解度を推定できるのか明らかにする必要がある．現在，教育現場，ポスター発表などの様々なシーンにおける理解度を推定する取り組みが行われている [2, 3, 4, 5]．しかし，対話における理解度を推定する研究は行われておらず，対話における聞き手の理解度を推定できるのかは明らかになっていない．現在，対話参加者の表情，発話音声，発話内容などのマルチモーダル情報を用いて，人間の発話意欲や興味度といった内部情報を推定する様々な取り組みが行われている [6, 7, 8, 9]．上記の取り組みから，対話参加者のマルチモーダル情報から，内部情報の一つである聞き手の理解度も推定できるのではないかと考えられる．

本研究では，**理解度**を**聞き手が話し手の発話を理解している様子の度合**と定義し，対話参加者のマルチモーダル情報から対話における聞き手の理解度を推定することができるのかを明らかにすることを研究目的とする．具体的には，対面の 2 者対話データと第三者による聞き手の理解度の評価が記録されている対話コーパスから聞き手と話し手の双方の視覚的，韻律的，言語的特徴を抽出する．次に，抽出した聞き手と話し手の双方の視覚的，韻律的，言語的特徴から聞き手の発話時の理解度を推定する機械学習モデルを構築し，対話参加者のマルチモーダル情報と聞き手の理解度の関係を明らかにする取り組みを行う．

1.3 本論文の構成

本論文の構成は次のとおりである．

2 章では，理解する行為に関連する研究事例と内部情報推定に関連する研究事例について述べる．
3 章では，本論文における問題の定義と研究課題について述べる．
4 章では，本研究で用いた対話コーパスについて述べる．
5 章では，聞き手の理解度の推定について述べる．
最後に 6 章にて，本論文の結論を述べる．

第2章　関連研究

本研究では，対話における話し手と聞き手のマルチモーダル情報を用いて, 聞き手の理解度を推定する取り組みを行う．本章では，この取り組みを行うにあたり，理解する行為に関連する研究事例と内部情報推定に関連する研究事例について述べる．本研究が属する研究領域を表 2.1 に示す．2.1 節では，理解する行為に関連する研究事例について紹介する．これらは，人間の理解に着目しているという点で本研究と関連している．2.2 節では，対話における人間の内部情報を推定する研究事例について紹介する．これらは，人間の言語，非言語行動を用いて，対話中の人間の内部情報を推定するという点で本研究と関連している．

表 2.1: 本研究が属する研究領域

研究分野	研究事例
理解する行為に関する研究	・講義や e-Learning などを受講している 生徒の理解度を推定する研究
内部情報推定に関する研究	・人間の言語・非言語行動を用いて 対話における人間の内部情報を推定する研究

2.1　理解する行為に関連する研究事例

本節では，理解する行為に関連する研究事例 [10, 2, 3, 11, 4, 5] について述べる．これらの研究事例では，教育現場，プレゼンテーションなどの様々なシーンにおける理解度を推定する取り組みを行っている．

Sathik ら [10] は仮想教室において，生徒の感情を表現する表情と生徒の理解度の関係を明らかにする研究を行っている．感情を伝える表情には眉を上げる，額に皺を寄せる，目を丸くする，唇を丸めるなどの筋肉の動きが含まれる [12, 13]．この事例では，生徒の感情を表現する表情と生徒の理解度の関係を明らかにするために，学生へのインタビュー調査が行われた．その結果，学生が講義を理解している，講義に満足している，講師の考え方を理解することができている，講義に対する反応を積極的に反応させたい時などに，ポジティブな感情表現を学生が行うことが明らかになった．ポジティブな感情は，目を大きく開き，眉を上げることで表現される．反対に，学生が講義を理解できない，講師にもう一度言ってもらいたい，講師に助けを求めようとする，講師のスピードについていけない，理解できず混乱している時などに，ネガティブな感情表現を学生が行うことが明らかになった．ネガティブな感情は，眉を下げて目を細める，額にしわを寄せる，眉を上げて目を大きくする，唇を丸めるといった動きによって表現される．

Holmes ら [2, 3] は，e-Learning を受講する学習者の理解度の推定を行っている．彼らは COMPASS と名付けられたリアルタイム理解度評価モデルを構築した．COMPASS は学習者の頭の動きや回転，視線の方向，瞬きの回数，肌の色の変化，性別や民族を含む 5 つのメタデータなどの 37 の非言語行動から, 1 秒ごとの理解度を推定する機械学習モデルで

ある．COMPASS のモデルは選択問題に回答する学習者の録画映像と学習者が選択した答えの正解/不正解が記録されたデータセットを元にトレーニングされており，構築したモデルの分類正解率は 75.8%を達成している．彼らは実際に COMPASS を用いた会話型インテリジェント個別指導システムを作成し, 実際の e-Learning の現場で理解度が推定できるのか実験を行った．実験の結果，答えを簡単に推測できないような難しい問題が出題された際は学習者が考える際の非言語行動が顕著に出現するため，モデルトレーニング時と同等の精度で理解度を推定できていた．それに対し，答えを簡単に推測できるような問題では，難しい問題での理解度の推定と比較し分類の正確性が低かった．

Curtis ら [11] は，プレゼンテーションにおいて聴衆の理解度を推定する機械学習モデルの構築を行っている．彼らは，International Speech Conference Multi-modal Corpus [14] と呼ばれるコーパスを使用している．このコーパスには 31 回の学術発表の講演者と聴衆の録画データが含まれている．講演者のデータにはプレゼンテーションのスライドを録画したデータも含まれている．聴衆の潜在的な理解度の評価は第 3 者のアノテータ 3 名によって行われた．アノテータ間の一致度の評価するために，検者間信頼性の信頼性を確認する指標となる級内相関係数（ICC:intraclass correlation coefficients）を求めた．その結果，$ICC(1,1) = 0.6034$ であり，タスクの特殊性を考慮もふまえると，理解度の評価は良好の一致度であったといえる．彼らは，上記のデータセットから，プレゼンテーションスライドの内容，講演者の視覚的，韻律的特徴と感情特徴，聴衆の視覚的特徴を抽出した．その次に，抽出した特徴量を用いて，聴衆の理解度を推定するクラス分類モデルを構築した．各特徴量を組み合わせたモデルの性能を比較すると，7 クラス分類モデルでは，講演者の視覚，韻律的特徴を組み合わせたモデルの正解率が 52.9%と，最も高い結果となった．2 クラス分類モデルでは，全ての特徴を用いたモデルの正答率が 85.4%と，最も高い結果となった．また，2 つの分類モデルにおいて講演者の韻律的特徴を用いたモデルの性能が高い結果となっていた．この結果から，講演者の韻律的特徴が聴衆の理解度を推定するのに最も重量な特徴であることが示唆されている．

Buckingham ら [4] は，インフォームド・コンセントのプロセスにおける参加者の表情から参加者の理解度の推定を行っている．データセットはフィールド調査に基づいて作成された．フィールド調査では，80 人の女性参加者が，インフォームド・コンセントのプロセスにおける理解しやすい話題と理解しにくい話題についての学習課題に関する説明を聞き，インタビューに答えるタスクが行われた．データセットには，インフォームド・コンセントのプロセスにおける理解しやすい話題に関するインタビューを受けている参加者の様子と，理解しにくい話題に関するインタビューを受けている参加者の様子が記録されている．彼らは，このデータセットから参加者の表情を視覚的特徴量として抽出した．その次に，インタビューを行っている間の表情を用いて，参加者がインフォームド・コンセントのプロセスを理解しているか否かを推定する機械学習モデルを構築した．機械学習モデルは心理プロファイリングシステムである Silent Talker を応用した Neural network（NN）を用いて構築された．Silent Talker[15] は，人間の表情から真実または欺瞞なのかを検出する NN ベースの心理プロファイリングシステムである．構築したモデルは，参加者がインフォームド・コンセントのプロセスを理解しているか否かを 85%を超える正解率

で推定できることが確認された．さらに，参加者がインフォームド・コンセントのプロセスを理解しているか否かを推定するタスクでは，表情に理解と非理解を検出可能な特徴が存在することが示唆された．

Kawahara ら [5] は，ポスター発表における聴衆の発話時の理解度を推定している．彼らは多数のポスター発表の会話を記録したデータセットを用いている [16, 17]．収録されているポスター発表では，1 人の発表者が自分の学術研究に関するポスターを用いて 2 人の聴衆に対し発表するタスクが行われている．上記のデータセットには，全ての参加者の視線情報と音声データが記録されている．ポスター参加者全員に各発表の理解度を尋ねるのは困難なため，観察が容易な聴衆の質問の種類と理解度の関係に着目した．質問の種類と聴衆の理解度の関係を分析し，理解度が低い場合に説明内容を確認するための質問（確認質問）が多く行われていることが確認された．彼らは，上記のデータセットから，特徴量として質問が行われた話題内の相槌の出現頻度と聴衆の視線情報が抽出した．そのつぎに，抽出した特徴量を用いて理解度の低い質問か予想する機械学習モデルを構築した．モデル構築の結果，全ての特徴量を用いたモデルで 75.7%の正解率が確認された．上記の結果より，ポスター発表における聴衆の理解度の推定には相槌や視線情報などの特徴を組み合わせることが効果的であることが示唆された．

2.2　内部情報推定に関連する研究事例

本節では，視覚的，韻律的，言語的情報などのマルチモーダル情報を用いて，対話における人間の内部情報を推定する研究事例 [6, 7, 8, 9] について述べる．これらの研究事例では，はじめに，人間の言語・非言語行動を記録した対話コーパスを利用もしくは作成し，対話コーパスから人間の視覚的，韻律的，言語的特徴の抽出を行なっている．そのつぎに，抽出した特徴量から対話における人間の内部情報を推定する機械学習モデルを構築している．さらに推定する際に重要な特徴量の分析を行っている．

Hirano ら [6] は，音声対話システムにおけるユーザの感情に基づいた適応メカニズムを実装するために，対話エージェントを使用するユーザの視覚的，韻律的，言語的情報から興味度，感情レベル，話題継続度を推定する機械学習モデルを構築している．データセットは，Osaka University Multimodal Dialogue Corpus （Hazumi）[18] を用いている．このデータセットは，WoZ 法によって操作されたエージェントと人との対話が収録されたマルチモーダルコーパスである．エージェントの発話開始時刻から，エージェントの次の発話開始時刻までを一つのターンとし，ターン毎にユーザの内部情報が評価されている．評価されているユーザの内部情報は，対話の話題に対する興味度，エージェントの発話に対して肯定的な感情，または否定的な感情を持っているのかといった感情レベル，エージェントが現在の話題を継続するべきかといった話題継続度などである．彼らは，上記のデータセットから，エージェントと対話する人の視覚的，韻律的，言語的特徴を抽出した．そのつぎに，抽出した特徴量を用いて，興味度，感情レベル，話題継続度を予測するクラス分類モデルを構築している．モデル構築に用いた機械学習アルゴリズムは，Support-vector machine（SVM），Single-Task Deep Neural Network（ST-DNN），

Multitask Deep Neural Network（MT-DNN）である．構築したモデル同士の性能を比較した結果，興味度推定タスクでは，視覚的，韻律的，言語的特徴の全てモダリティに関する特徴量を説明変数として利用し，MT-DNN で構築したモデルが最も分類精度が高く，感情レベル，話題継続度推定タスクでは，韻律的特徴と言語的特徴を説明変数として利用し，MT-DNN で構築したモデルが最も分類精度が高いことが確認された．

Ishii ら [7] は，2 者対話における話し手と聞き手の視覚的，韻律的，言語的情報を用いて，話し手と聞き手の双方のターンマネジメントの意欲と対話の話者交代タイミングの予測する機械学習モデルを構築している．モデル構築に用いた対話データセットには，初対面同士の対面対話データと話し手と聞き手のターンマネジメントの意欲に関するアノテーションが記録されている．彼らは，上記のデータセットから，話し手，聞き手それぞれの視覚的，韻律的，言語的特徴を抽出した．そのつぎに，抽出した特徴量を用いて，話し手の発話意欲と傾聴意欲，聞き手の発話意欲と傾聴意欲，といった 4 つのターンマネジメントの意欲の推測と，話者交代が行われるか否かの予測を行うマルチタスクモデルを構築した．機械学習アルゴリズムは，Deep neural network を用いている．構築したモデルの性能を比較したところ，話し手と聞き手の全てのマルチモーダル情報を使用したモデルが，ターンマネジメントの意欲と話者交代を最も正確に予測できることが確認された．また，話者交代タイミングの予測を個別で推定するシングルタスクモデルの性能に比べ，マルチタスクモデルの性能がより正確であることが確認された．この結果は，人間の行動を予測する際に，人間の心理状態に関連する情報を推定するタスクを組み合わせることで，人間の行動をより正確に予測できる可能性を示唆している．

Ohba ら [8] は，面接者のコミュニケーションスキルと自己効力間の関係を明らかにすることを目標とし，エージェントとの面接での対話におけるユーザの視線，韻律，言語的特徴と生体情報を用いて，面接を行う人のコミュニケーションスキルと自己効力感の推定する機械学習モデルを構築した．彼らは面接時のデータセットを作成するために，独自の面接システムを開発した．面接システムは，ユーザが HMD を使用し，VR 空間上の 3 体の面接官エージェントからの質問に答える面接体験を行う仕様である．データセットには，面接システムを用いて面接体験を行うユーザの音声，視線情報，生体信号と，ユーザの回答ごとに，ユーザ自身によって評価された自己効力感データと，第 3 者の専門の面接トレーナによってアノテーションされたコミュニケーションスキルのデータが記録されている．また，自己効力感とコミュニケーションスキルの差によって求められるギャップデータも記録されている．彼らは，上記のデータセットから，ユーザの視線，韻律，言語的特徴と生体情報を抽出した．そのつぎに，抽出した特徴量を用いて，面接を行う人の自己効力感，コミュニケーションスキル，ギャップデータを推定するモデルを構築した．機械学習モデルは，線形回帰モデルと Sequential モデルの二つのモデルを構築している．構築したモデル同士の性能の比較を行った．モデル構築の結果，それぞれの推定タスクによって有効な特徴が異なることが確認された．自己効力感の推定の場合，生体特徴のみを用いて推定する Sequential モデルの精度が最も高かった．コミュニケーションスキルの推定の場合，視線，韻律，言語的特徴を用いて推定する Sequential モデルの精度が最も高かった．ギャップデータの推定の場合，韻律，言語的特徴を用いて推定する線形回帰モデルの精度

が最も高かった．この結果から，第三者に評価されるコミュニケーションスキルは，韻律や言語的情報といった観察可能な特徴を用いて推定することができ，ユーザ自身によって評価される自己効力感は，外部観察が不可能な生体情報を用いて推定することができる可能性が明らかになった．

Pellet-Rostaing ら [9] は，2 者対話における話し手の視覚的，韻律的，言語的情報を用いて，話し手のエンゲージメントの推定する機械学習モデルを構築している．彼らは，モデル構築には Paco-Cheese コーパス [19, 20] を使用している．このデータセットには，フランス語での 2 者対面対話のデータが収録されている．対話参加者には，対話前に短い物語を読み，その後，自由に対話を行うよう指示がされている．アノテーションは 2 人の第三者の専門家によって行われ，参加者のエンゲージメントの度合いが 5 段階で評価されている．彼らは，上記のデータセットから，話し手の視覚的，韻律的，言語的特徴を抽出した．韻律的特徴は，発話音声に関する特徴と発話時間に関する特徴を抽出している．そのつぎに，抽出した特徴量を用いて，話し手のエンゲージメントを推定する機械学習モデルを構築した．機械学習モデルは，Logistic regression，SVM，K-nearest neighbors（KNN）, Ada Boost，Naive Bayes, Random forest, Multilayer perception の異なる 7 つのアルゴリズムを用いて構築され，それぞれのモデルの推定性能の比較を行っている．モデル構築の結果，SVM を用いたモデル，Logistic regression を用いたモデルの F 値が 0.58 であり，ベースラインや他のアルゴリズムを用いたモデルの性能を上回り，他のモデルと比較し最良の性能を確認することができた．話し手のエンゲージメント推定に重要な特徴を明らかにするため，ユニモーダル情報を説明変数としたモデルとマルチモーダル情報を説明変数としたモデルの性能比較を行った．性能比較の結果，発話音声に関する特徴量，発話時間に関する特徴量，視覚的特徴量を組み合わせたモデルが他のモデルと比較し，最も高い F 値を確認することができた．また，言語的特徴は，他のモダリティと組み合わせても，有意な性能の向上にはならないことが確認された．

第3章　研究課題

本章では，本研究における問題の定義と研究課題について述べる．

3.1　問題の定義

日常生活や社会活動での対話において対話相手（聞き手）が自分（話し手）の話を理解しているか否かを推測することは重要である．上記の点は，人と人のインタラクションだけでなく，人と対話エージェントのインタラクションにおいても重要であると考える．対話エージェントが人間と同じようにユーザの理解度を推測できるようになれば，ユーザの理解に応じた適切なコミュニケーションを行うことが期待できる．理解度を考慮して円滑にコミュニケーションを行う対話エージェントの開発を実現するためには，対話シーンにおける人間の理解度を推定することが可能なのかを明らかにする必要がある．

2.1 節で述べたように，医療現場，ポスター発表などの様々なシーンにおける理解度を推定する取り組みが行われている．Buckingham ら [4] は，インフォームド・コンセントのプロセスにおける参加者の表情から参加者の理解度を推定しており，参加者の表情から，参加者がインフォームド・コンセントのプロセスを理解しているのか否かを推定できることを明らかにしている．Kawahara ら [5] は，ポスター発表における聴衆の理解度を推定しており，聴衆の相槌や視線情報から，聴衆の発話時の理解度を推定できることを明らかにしている．これらの研究より，表情や音声といった人間の言語・非言語行動が人間の理解を推定することに役立つことが示されている．しかし，上記の研究では，医療行為の説明やポスター発表といったフォーマルなシーンに限定した理解度を推定しており，対話シーンにおける理解度を推定する研究は行われておらず，対話における聞き手の理解度を推定できるのかは明らかになっていない．対話シーンにおける理解度が推定されていないことにより，ユーザの理解を考慮して円滑にコミュニケーションを行う対話エージェントを開発することが困難であると考えられる．上記をふまえ，本研究における問題は，対話における対話参加者のマルチモーダル情報から聞き手の理解を推定することができるのか明らかではないことであると定義できる．

3.2　研究課題の設定

3.1 節で述べたように，対話における対話参加者のマルチモーダル情報から聞き手の理解を推定することができるのか明らかではないという問題があり，ユーザの理解を考慮してコミュニケーションを行う対話エージェントを開発することが困難であると考えられる．この問題を解決するためには，対話における聞き手の理解を推定することができるのか明らかにする必要がある．2.2 節で述べたように，マルチモーダル情報を用いて，対話における人間の内部情報を推定する取り組みが行われている．Hirano ら [6] は，音声対話システムを使用するユーザの興味度，感情レベル，話題継続度を推定しており，ユーザの言語・非言語情報から，ユーザの興味度，感情レベル，話題継続度を推定できることを明らかにしている．Ishii ら [7] は，2 者対話における話し手と聞き手の双方のターンマネジメ

ントの意欲を推定しており，話し手，聞き手の視覚的，韻律的，言語的特徴から，双方のターンマネジメントの意欲を推定できることを明らかにしている．Pellet-Rostaing ら [9] は，2 者対話における話し手のエンゲージメントを推定しており，話し手の視覚，韻律的特徴から，話し手の発話時のエンゲージメントを推定できることを明らかにしている，これらの既存研究から，対話参加者のマルチモーダル情報から，内部情報の一つである聞き手の理解度も推定できるのではないかと考えられる．そこで本研究では，**聞き手が話し手の発話を理解している様子の度合**を**理解度**と定義し，対話参加者のマルチモーダル情報と聞き手の理解度の関係を分析する取り組みを行う．2 者対話は，話し手と聞き手のインタラクションで成り立っているため，聞き手自身だけでなく，話し手の振舞いも聞き手の理解度に影響があることも考えられる．さらに，マルチモーダルインタラクションの研究分野では，対話における話者のスキルや内部情報を推定する際に，対話相手の情報も取り入れる研究も数多く存在する [7, 21, 22]．話し手の振舞いが聞き手の理解度に与える影響が明らかになることで，よりユーザが理解しやすいアクションを取る対話エージェントの設計が期待できる．そこで本研究では，聞き手のマルチモーダル情報だけではなく，話し手のマルチモーダル情報も分析の対象とする．上記の取り組みを行うことにより，対話参加者のマルチモーダル情報から聞き手の理解度を推定することが明らかになり，人間のようにユーザの理解を考慮して円滑にコミュニケーションを行う対話エージェントの開発が期待される．

本研究では，対話参加者のマルチモーダル情報と聞き手の理解度の関係を分析する取り組みとして，聞き手と話し手の双方のマルチモーダル情報から，聞き手の発話時の理解度を推定する機械学習モデルを構築する．具体的には，対面の 2 者対話データと第三者による聞き手の理解度の評価が記録されている対話コーパスから聞き手と話し手の双方の視覚的，韻律的，言語的特徴量を抽出する．抽出した特徴量を用いて聞き手の発話時の理解度を推定する機械学習モデルを構築する．構築する機械学習モデルは聞き手の発話時の理解度を推定する回帰モデルと，聞き手の発話を理解をしている発話，中立の発話，理解をしていない発話に分類する 3 クラス分類モデルである．さらに，理解度を推定するために重要なマルチモーダル情報を明らかにするため，重要な特徴量の算出や，説明変数として用いる特徴を組み合わせた複数のモデルを構築し，構築したモデル同士の性能の比較を行う．

上記をふまえ，本研究では，**聞き手と話し手の双方の視覚的，韻律的，言語的特徴から聞き手の理解度を推定することができるのかを明らかにすること**を研究課題として設定する．

第4章　対話コーパス

本章では，本研究で用いた対話コーパスについて述べる．対話コーパスには，対面の 2 者対話データと第三者による聞き手の理解度の評価が記録されている．

4.1 2 者対話データ

本節では，本研究で用いた対話コーパスに含まれる対面の 2 者対話データについて述べる．本研究では，聞き手と話し手の双方のマルチモーダル情報から，聞き手の理解度を推定する機械学習モデルを構築するために対面における 2 者対話データを利用する．2 者対話データは，既存の対話データセット [23] と，上記のデータセットを拡張した対話データセットを使用する．それぞれの対話データセットの情報を表に示す．

表 4.1: それぞれの対話データセットの情報

データセット	収録対話数	聞き手の IPU	話し手の IPU	合計 IPU
既存の対話データセット [23]	26 対話	2,865 件	4,940 件	7,805 件
拡張した対話データセット	94 対話	15,892 件	24,687 件	40,579 件

既存の対話データセットには 26 件の対話が記録されているのに対し，拡張された対話データセットには 94 件 (既存対話 26 件，新規対話 68 件) の対話が記録されている．

2 者対話の参加者は，初対面の 20 代から 50 代の日本人男女である．対話データの収録には，ストーリテリングと呼ばれる手法が用いられている．ストーリーテリングは，対話前に話し手がアニメーションなどを鑑賞し，その後，聞き手に鑑賞した内容を自由に説明する手法である．この手法は，話し手の発話に伴う聞き手の反応に関するデータを多く収集することができ，対話データを用いて機械学習モデルを構築する類似研究でも用いられている [24, 25, 26]．対話参加者は対話の前に，アニメーション「トムとジェリー」を鑑賞した．その後，一方の参加者（話し手）が他方の参加者（聞き手）に，対話の前に鑑賞したアニメ内容の説明を行っている．1 つの対話は 10 分程度で行われている．聞き手は話し手の発話に対し自由に質問することが許可されている．話し手と聞き手の対話例を表 4.2 示す．対話データは，話し手，聞き手の正面に設置されたビデオカメラによって撮影された動画データと話し手，聞き手の音声データ聴き手と話し手の発話内容から構成されている．

発話の単位は Inter-pausal units（IPU）[27] を用いており，沈黙時間が 200ms 未満の連続した音声区間を 1 つの発話としている．本研究と同様に日本語の対話コーパスを用いる研究事例でも発話を区切る時間として 200ms を閾値とし設定している研究が数多く存在する [28, 29, 30, 7]．既存の対話データには，合計 7,805 件（話し手：4,940 件，聞き手：2,865 件）の IPU が記録されている．拡張した対話には，合計 40,579 件（話し手：24,687 件，聞き手：15,892 件）の IPU が記録されている．

表 4.2: 話し手と聞き手の対話例

発話開始時	発話終了時間	話し手	聞き手
01:58.104	01:59.900	ネズミのほうはー次はちゃんと	
02:00.087	02:00.400		うん
02:00.822	02:03.620	ネズミをー袋に閉じてーぐるぐるに巻いて	
02.02.636	02.02.900		うん
02:03.857	02:05.160	で動けない状態にして	
02:05.184	02:06.200		かわいそううん
02:06.040	02:06.754	でバーッて	
02:06.971	02:07.640	遠くに投げて	
02:07.011	02:07.320		うん
02:07.753	02:08.080		うん
02:08.313	02:08.900	でめちゃ	
02:09.232	02:11.660	怖い魚がいんのでかくってワニみたいな	
02:10.605	02:12.340		うんうんうんうんうんうん
02:13.092	02:14.988	そいつに食べられそうになりつつも	
02:15.120	02:14.420		うん
02:15.838	02:16.280	なんとか	
02:16.640	02:17.586	逃げられるみたいな	

4.2 理解度の判定

本節では，聞き手の理解度の評価方法について述べる．本研究では，第三者が聞き手の理解度を評価する方法を採用した．理解度の評価は聞き手自身に評価してもらう方法も考えられる．例えば，対話を行っている最中に聞き手自身にリアルタイムで理解度を評価する方法である．しかしこの方法では，聞き手が自身の理解度をリアルタイムで評価することに集中してしまい，自然な対話が成立しない可能性が生じる．また，対話後に聞き手自身に対話時の理解度を評価してもらう方法も考えられる．しかしこの方法では，聞き手はすでに対話内容を知ってしまっているため，対話後に対話中の各発話時点などの任意のタイミングに対して聞き手自身が正確な理解の評価を行うことは難しい．したがって，聞き手自身に対話時の理解度を評価することは困難である．そこで本研究では，第三者が参加者の理解度を評価する方法を採用した．内部情報を第三者が評価する方法は，多くの内部情報推定の類似研究で用いられている [7, 31, 32, 33, 34]．また，第三者による客観的な観察に基づく評価と参加者本人による主観的な評価には，一定の相関関係があることも報告されている [35]．

理解度の評価は，1 つの対話に対して第三者のアノテータ 3 名によって行われている．4.1 節にあるように，使用した対話データは，既存の対話データセットを拡張した対話データセットを用いている．拡張した対話データセットに含まれる新規収録した対話データセットでは理解度の評価がアノテータ 2 名によって行われていた．そこで，本研究ではアノテータの人数を 3 名に合わせるため，新規収録した対話データに含まれる対話に対して，追加でアノテーションを行った．

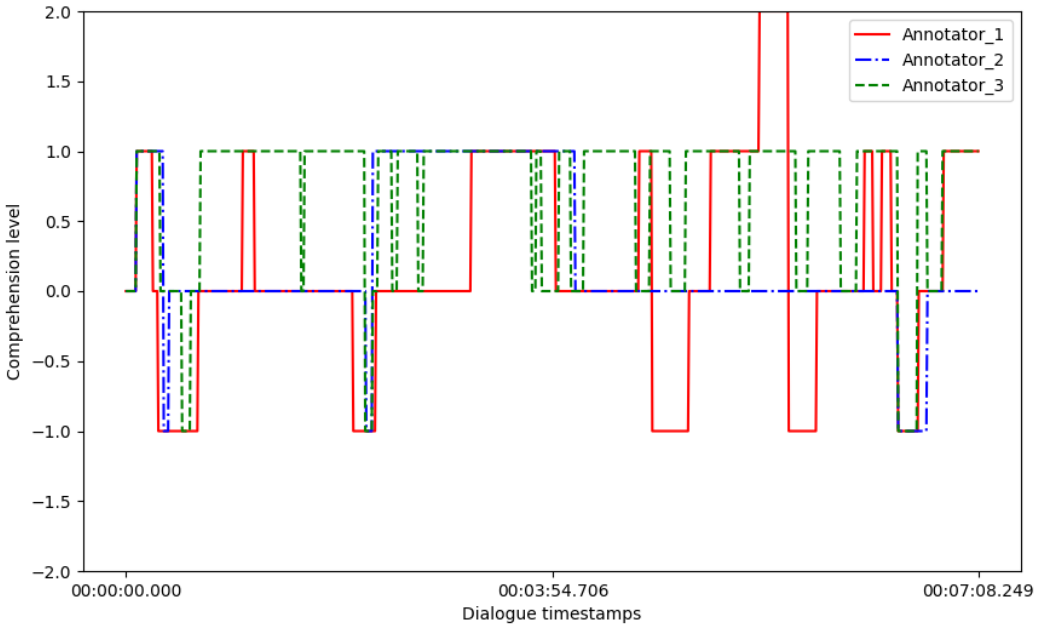

図 4.1: 3 名のアノテータによる理解度の評価の一例

アノテータによる理解度の判定は「相手の話をどれだけ理解できているのか」という観点で評価を行った．アノテータは，聞き手の発話毎に理解度の評価を行っているのではなく，対話を通して聞き手の理解している様子が変化したという任意のタイミングで評価を変更している．図 4.1 は一つの対話における 3 名のアノテータの理解度の評価の一例を可視化したものである．図 4.1 から，各アノテータの理解度の評価が変更されるタイミングは，アノテータ同士で一致していないことがわかる．理解度の評価は，理解の度合いを-2（全く理解していない），-1（あまり理解していない），0（中立の状態），+1（少し理解している），+2（とても理解している）の 5 段階で行なわれた．

第5章　理解度の推定

本章では，聞き手の理解度の推定について述べる．本研究では，対話参加者のマルチモーダル情報から聞き手の理解度の関係を分析する取り組みを行う．そのため，聞き手と話し手の双方の視覚的，韻律的，言語的特徴から，聞き手の理解度を推定する機械学習モデルを構築し，理解度が推定できるのか検証する必要がある．具体的には，話し手と聞き手の発話時の視覚的，韻律的，言語的特徴を抽出し，聞き手の発話時の理解度を推定する機械学習モデルを構築する．本研究では聞き手の発話時の理解度を推定する二種類のモデルを作成した．1 つ目は，聞き手の視覚的，韻律的，言語的特徴から，理解度を回帰タスクで推定するモデルである．2 つ目は，聞き手と話し手の双方の視覚的，韻律的，言語的特徴から理解度を 3 クラスに分けクラス分類タスクで推定するモデルである．さらに，本章では，構築したモデルの結果から，対話参加者のマルチモーダル情報と聞き手の理解度の関係に関する考察について述べる．理解度を推定するために重要なマルチモーダル情報を明らかにするため，重要な特徴量の算出や，説明変数として用いる特徴を組み合わせた複数のモデルを構築し，構築したモデル同士の性能の比較を行う．

5.1　発話時の理解度

本節では，機械学習モデルの目的変数として用いる聞き手の発話時の理解度について述べる．聞き手と話し手の双方のマルチモーダル情報から，聞き手の理解度を推定する機械学習モデルを構築するために，4.2 節で判定した理解度を目的変数として用いる．理解度

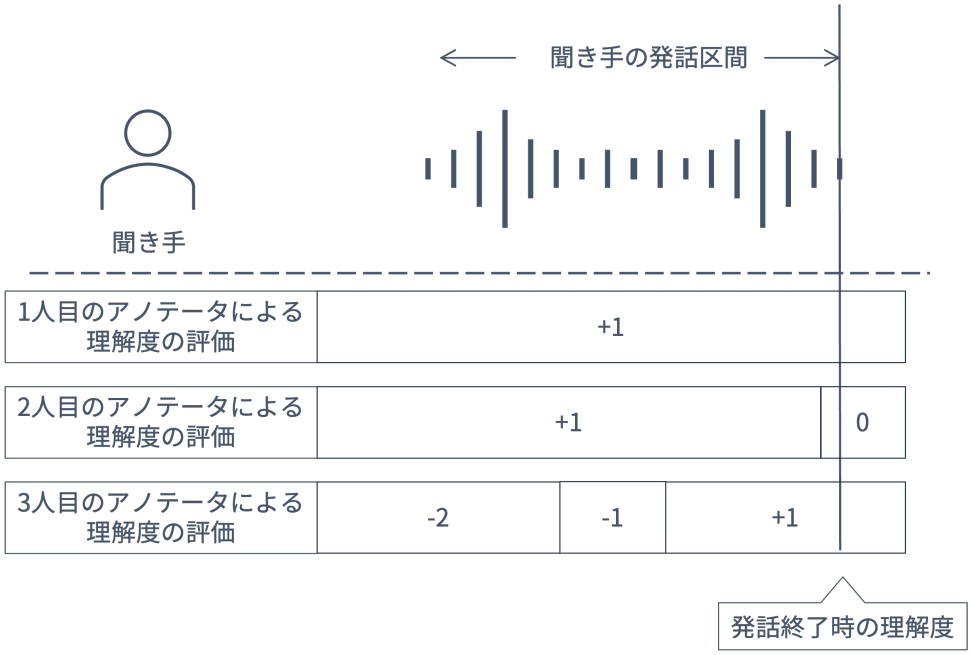

図 5.1: 目的変数に用いる聞き手の理解度

のアノテーションは図 4.1 のように対話全体を通して評価されている．アノテータの評価が対話全体を通して行われているため，発話内でアノテータが評価を変更している発話も

存在する．本研究では，発話区間中の行動が発話終了時点での理解度に貢献すると考え，発話終了時点でのアノテータが評価した理解度（図 5.1）を発話時の理解度する．

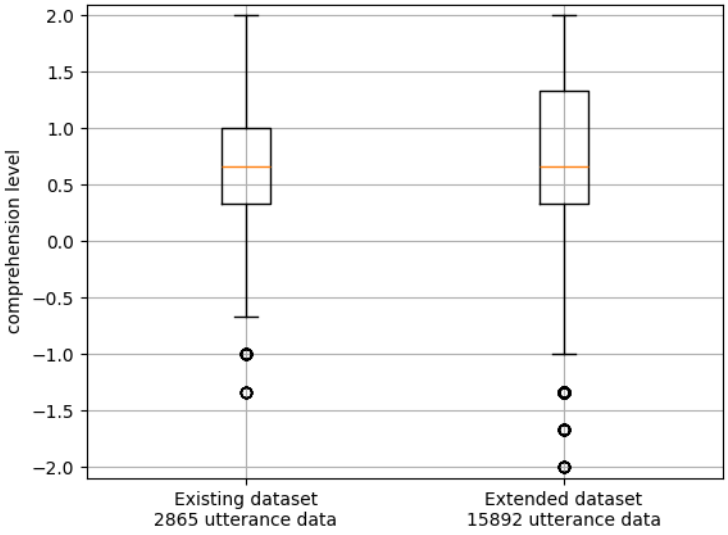

図 5.2: 聞き手の発話時の理解度の分布

既存のデータセットと拡張したデータセットでアノテータ 3 名が評価した発話時の理解度の分布を図 5.2 に示す．聞き手の発話時におけるアノテータ間の理解度の一致率を評価するために，検者間信頼性の信頼性を確認する指標となる級内相関係数（ICC）[36] を算出した．聞き手の発話時における 3 名のアノテータが評価した理解度の検者間信頼性は既存のデータセットでは $ICC(2,k) = 0.546$，拡張したデータセットでは $ICC(2,k) = 0.640$ であった．それぞれの評価，本研究と同じように第三者が評価している研究 [14, 37] の検者間信頼性を用いた評価と同程度であり，既存のデータセットと拡張したデータセットの理解度の評価データは，どちらも信頼性の高いデータであることが示唆された．

5.2　特徴量の抽出

本節では，話し手と聞き手の発話時の視覚的，韻律的，言語的特徴の抽出について述べる．これらの特徴量はマルチモーダルインタラクションの分野でよく使われている特徴量である [38, 39, 6, 7, 8, 9, 26]．

5.2.1　視覚的特徴量の抽出

本項では，話し手と聞き手の視覚的特徴の抽出について述べる．本研究では，話し手と聞き手の双方の頭部・顔部の振舞いに関連する特徴量を，顔画像処理ツールである

OpenFace[40] を用いて抽出した．具体的には，参加者の正面に設置したカメラで撮影した映像データから，頭部の向き，視線の角度，筋肉群の基本的な行動の単位を表す Action Units（AU）[41] に関する特徴量の抽出を行った．頭部の向きは，カメラ側から顔を見て，左から右の方向を x 軸，下から上の方向を y 軸，手前から奥の方向を z 軸とした．聞き手の発話時における頭部の x 軸，y 軸，z 軸回りの回転角度（pose_Rx，pose_Ry， pose_Rz）の分散（_var），中央値（_med），10 パーセンタイル値（_p10），90 パーセンタイル値（_p90）を用いた．視線の角度は，カメラ側から顔を見て，左から右の方向を x 軸，下から上の方向を y 軸とした．聞き手の発話時における視線の向きの x 軸，y 軸回りの回転角度（gaze_Ax，gaze_Ay）の分散，中央値，10 パーセンタイル値，90 パーセンタイル値を用いた．Action Units は，聞き手の発話時における各 Action Units（表 5.1）の強度の分散，中央値，10 パーセンタイル値，90 パーセンタイル値を用いた．

表 5.1: Action Units の内容

項目	内容	項目	内容
AU01	眉の内側を上げる	AU14	笑窪を作る
AU02	眉の外側を上げる	AU15	唇の両端を下げる
AU04	眉を下げる	AU17	顎を上げる
AU05	上瞼を上げる	AU20	唇の両端を横に引く
AU06	頬を持ち上げる	AU23	唇を固く閉じる
AU07	瞳を緊張させる	AU25	顎を下げずに唇を開く
AU09	鼻に皺を寄せる	AU26	顎を下げて唇を開く
AU10	上唇を上げる	AU45	瞬きをする
AU12	唇の両端を引き上げる		

5.2.2 韻律的特徴量の抽出

本項では，話し手と聞き手の韻律的特徴の抽出について述べる．本研究では，話し手，聞き手の双方の音声データから，音声の韻律の代表的な特徴量，発話に関する代表的な特徴量を，音声情報処理ツールである openSMILE[42] を用いて抽出を行った．表 5.2 に示した音声の韻律の代表的な特徴量につき，表 5.3 で示した統計量を算出し，336 次元の特徴量を抽出した．さらに，音声データに含まれる 1 つの発話の継続時間も特徴量として用いる．

5.2.3 言語的特徴量の抽出

本項では，話し手と聞き手の言語的特徴の抽出について述べる．話し手と聞き手の双方の発話内容から，発話単語数と品詞に関する言語的特徴と発話内容をベクトル表現に変換した言語的特徴を用いた．

表 5.2: 韻律的特徴量の内容

特徴量	内容
pcm_intensity_sma	正規化された強度の値
pcm_loudness_sma	正規化された強度に 0.3 乗した値
mfcc_sma[1]–[12]	1〜12 次のメル周波数ケプストラム係数
lspFreq_sma[0]–[7]	8 つの LPC 係数から計算される周波数
pcm_zcr_sma	ゼロ交差率
voiceProb_sma	声である確率
F0_sma	基本周波数
F0env_sma	基本周波数のエンペローブ

表 5.3: 抽出した統計量

統計量	内容	統計量	内容
max	最大値	linregc1	線形近似の勾配
min	最小値	linregc2	線形近似のオフセット
range	最大値と最小値の差	linregerrA	線形近似と実際の値の誤差の差
maxPos	最大値の絶対位置	linregerrQ	線形近似の二乗誤差
minPos	最小値の絶対値位置	skewness	歪度
amean	平均値	kurtosis	尖度
stddev	標準偏差		

品詞の出現頻度と単語数に関する言語的特徴

品詞の出現頻度と単語数に関する言語的特徴は，日本語形態素解析システムである MeCab[43] を用いて抽出を行った．各品詞の出現頻度と単語数を特徴量とした．単語数は，1 つの発話内容に含まれる単語数である．各品詞の出現頻度は．発話内容から抽出した各単語を表 5.3 で示した 35 種類の品詞に分類し，品詞ごとに利用された数を単語数で除算することで求めた．

高次元ベクトル表現に変換した言語的特徴

話し手と聞き手の双方の発話から，自然言語処理モデルである BERT[44] を用いて言語的特徴量を抽出した．具体的には，日本語事前学習済みの BERT モデルを利用し，発話を 768 次元のベクトル表現に変換し，言語的特徴量とした．

5.3　機械学習モデルの構築

本節では，理解度を推定する機械学習モデルの構築について述べる．本研究では，対話参加者の視覚的，韻律的，言語的特徴から聞き手の理解度を推定することができるのか明らかにするために理解度を推定する機械学習モデルを構築する．このとき，聞き手の発話

表 5.4: 抽出した品詞

大分類	小分類
フィラー	フィラー
感動詞	感動詞
形容詞	形容詞
助詞	助詞（格助詞），助詞（副助詞），助詞（連体化）， 助詞（接続助詞），助詞（特殊），助詞（副詞化）， 助詞（副助詞），助詞（副助詞／並立助詞／終助詞）， 助詞（並立助詞），助詞（連体化），
助動詞	助動詞
接続詞	接続詞
動詞	動詞（自立），動詞（接尾），動詞（非自立）
名詞	名詞（サ変接続），名詞（ナイ形容詞語幹），名詞（一般）， 名詞（引用文字列），名詞（形容動詞語幹），名詞（固有名詞）， 名詞（数），名詞（接続詞的），名詞（接尾），名詞（代名詞）， 名詞（動詞非自立的），名詞（特殊），名詞（非自立）， 名詞（副詞可能）
連体詞	連体詞
副詞	副詞
接頭詞	接頭詞

時の理解度を推定する二種類のモデルを作成した．1 つ目は，聞き手の視覚的，韻律的，言語的特徴から，聞き手の発話時の理解度を推定する回帰モデルである．2 つ目は，聞き手と話し手の双方の視覚的，韻律的，言語的特徴から聞き手の発話を理解をしている発話，中立の発話，理解をしていない発話にクラス分類するモデルである．

回帰モデルは，聞き手の視覚的，韻律的，言語的特徴量から聞き手の発話時の理解度を推定するモデルである．このモデルは，本研究の初期検討 [45] として構築したため，4 章で述べた，既存のデータセットのみを用いている．また，説明変数として 5.2 節で抽出した聞き手のマルチモーダル情報のみを用いている．クラス分類モデルでは，聞き手と話し手の双方の視覚的，韻律的，言語的特徴から聞き手の発話を理解をしている発話，中立の発話，理解をしていない発話に分類するクラス分類モデルである．聞き手の発話時の理解度の分布 5.2 を確認すると，本研究の理解度データは全体的に高めに評価されていることがわかる．そのため，回帰モデルでは理解していない発話を推測することが難しいと考えられる．理解していない発話を推定するために，理解度を 3 群に分け，クラス分類モデルを構築する．回帰モデルでは，本研究の初期検討という理由で聞き手の振舞いのみに着目していたが，聞き手だけでなく，話し手の振舞いも聞き手の理解度に影響があることも考えられるため，クラス分類モデルでは話し手の振舞いも説明変数として用いる．さらに，構築した各モデルの結果から，対話参加者のマルチモーダル情報と聞き手の理解度の関係

に関する考察についても述べる．理解度を推定するために重要なマルチモーダル情報を明らかにするため，重要な特徴量の算出や，説明変数として用いる特徴を組み合わせた複数のモデルを構築し，構築したモデル同士の性能の比較を行う．

5.4　回帰モデル

5.4.1　回帰モデルの構築

本項では，聞き手の理解度を推定する回帰モデルの構築について述べる．回帰モデルは，聞き手の発話時の視覚的，韻律的，言語的特徴量から聞き手の発話時の理解度を推定するモデルである．モデルの概要を図 5.3 に示す．このモデルは，聞き手のマルチモーダ

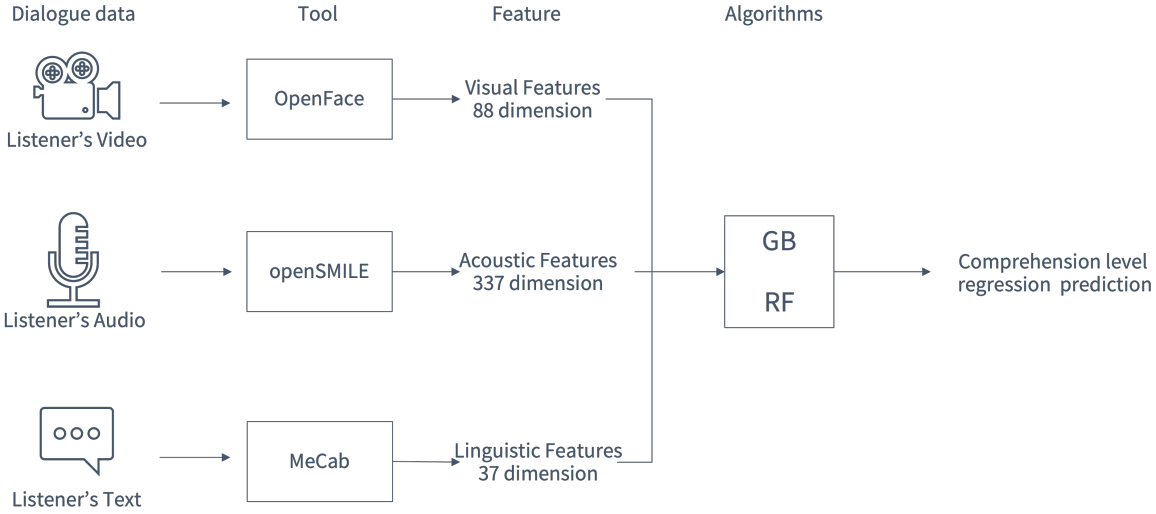

図 5.3: 回帰モデルの概要

ル情報を用いて聞き手の理解度を推定できるのか明らかにする取り組みの初期検討 [45] として作成した．そのため，データセットは 4.1 節に示した 2 つのデータセットのうち，他の既存研究で用いられているデータセットのみを用いている．モデル構築の手法は，理解度を推定するためにどのような振舞いが重要なのかを分析するために，人間による解釈が容易である特徴量や特徴量の重要度を評価することができる機械学習アルゴリズムを採用した．

目的変数

目的変数は 5.1 節で述べた聞き手の発話終了時点 (図 5.1) でのアノテータ 3 名が評価した理解度の平均値である．

説明変数

説明変数は 5.2 節で抽出した聞き手の視覚的，韻律的，言語的特徴である．言語的特徴は人間による解釈が容易である Mecab を用いて抽出した品詞の出現頻度と単語数に関する特徴を用いた．

モデルの構築方法

モデル構築には重要な特徴量が把握しやすいアルゴリズムである Gradient Boosting（GB）[46]，Random Forests （RF）[47] の 2 つを採用した．それぞれのアルゴリズムは，scikit-learn [48] を用いて実装した．それぞれのモデルの決定木の数や深さ，学習率などのハイパーパラメータは，Hyperopt[49] を用いて最適化を行った．

5.4.2　回帰モデルの推定結果

本項では，5.4.1 項で構築した聞き手の視覚的，韻律的，言語的特徴量から聞き手の発話時の理解度を推定する回帰モデルの推定結果について述べる．モデルの推定結果を分析するため，次のタスクを 100 回ずつ行った．

(1) 特徴量が含まれているデータセットを訓練データとテストデータを 9 : 1 に無作為に分割する．

(2) 訓練データで構築したモデルを構築したモデルを用いてテストデータにおける目的変数を推定するタスクを行う．

本研究の回帰モデルで用いる Baseline モデルは，シーン全体の発話終了時理解度の平均値（0.570）を推定結果として出力するモデルである．上記で述べたタスクを 100 回行った結果を表 5.5 に示す．

表 5.5: 回帰モデルの推定結果の平均値（N = 100）

Indicators	Baseline	GB	RF
MSE	0.257	0.197	0.198
R^2	-0.003	**0.243***	**0.231***

* $p < 0.01$

各アルゴリムで構築したモデルの決定係数と Baseline モデルの決定係数で，1%水準で Holm 法による補正を用いて対応のない t 検定を行った．その結果，本研究で作成したすべての回帰モデルで Baseline モデルと有意差を認められることが確認できた．このことから，聞き手のマルチモーダル情報が聞き手の発話時の理解度の推定に有効であることが示された．

5.4.3　回帰モデルに関する考察

本項では，5.4.2 項で述べた回帰モデルの結果から聞き手のマルチモーダル情報と聞き手の理解度の関係に関する考察を述べる．具体的には，聞き手の発話時のマルチモーダル情報から聞き手の発話時の理解度を推定するモデルでの特徴量の重要度を算出する．算出した特徴量の重要度を分析し，聞き手の理解度を推定する際に重要な聞き手の振舞いを明

らかにする．GB，RF を用いて実装したそれぞれのモデルでの特徴量の重要度を Gain*を用いて算出した．各特徴量の重要度を図 5.4 と図 5.5 に示す．

どちらのモデルでも，重要度の高さが上位 3 件の特徴量は，Pose_Rx_VAR，AU06_r_P90，AU12_r_P90 であった．これらは，頭部の x 軸回りの回転角度，頬を持ち上げる動き，唇の両端を引き上げる動きに該当する．理解度の高さによって，重要度の高かった 3 つの動きの振舞い方に違いがあるのかを明らかにするために，理解度の高さを理解度の分布（図 5.2）を元に 3 群に分けて特徴量を分析する．3 名のアノテータが評価した理解度の平均値が 1 以上のデータを理解度高群（938 件），平均値が-1 より大きく 0.333 未満のデータを理解度中群（1,335 件），平均値が 0.333 以下のデータを理解度低群（588 件）とする．理解度高群での特徴量の平均値と理解度低群での特徴量の平均値の間で 5%水準で対応のない t 検定を行い，得られた p 値を表 5.6 に示す．

表 5.6: 理解度高群と低群における重要度の高い特徴量の平均値

特徴量	理解度高群（N = 938）	理解度低群（N = 588）	p 値
pose_Rx_VAR	0.0042	0.0031	0.1398
AU06_r_P90	0.9673	0.6511	**2.42E-25***
AU12_r_P90	1.0770	0.6721	**2.50E-33***

* $p < 0.05$

理解度高群と低群を比較した結果，AU06_r_P90 と AU12_r_P90 の特徴量で有意差を確認することができた．このことから，理解度を推定するためには，頬を持ち上げる動き，唇の両端を引き上げる動きの 2 つの振舞いが特に重要であると考えられる．さらに，2 つの AU の振舞いは理解度高群の時に強度が高く，理解度低群の時に強度が低いことを確認することができた．この 2 つ AU は人間が笑顔を表す際に動かされる部分であり，人間の笑顔を調査している研究 [50] で用いられている振舞いである．そのため，理解度が高いとより笑顔の表情になることが示唆される．今回の対話コーパスでは，アニメーション「トムとジェリー」の内容を説明するタスクを用いて収録されている．アニメーション「トムとジェリー」は内容がコミカルという点が特徴である．聞き手は話し手からアニメーションのコミカルな内容を伝えられているため，話し手の話の内容が面白いという感情が生まれ，笑顔が表出していると考えられる．しかし，例えば，話し手からシリアスな説明をされているタスクでは，聞き手が面白いという感情にならず，さらには聞き手が笑顔を表出しにくい状況のため，違う振舞いが理解度の推定をするために重要となるかもしれない．

品詞頻度に関する特徴量が GB モデル，RF モデルの上位 20 件には含まれなかった．実際の対話データを確認したところ，聞き手の発話内容が「うんうん」,「はい」,「ああ」などの一つの品詞で構成されている．これらは，品詞分類を行うと感動詞に分類される．そのため，多くの発話が別々の発話内容だったとしても特徴量に変換すると同一の特徴量となってしまっていた可能性がある．このようなことが要因となり，品詞に関する特徴量の重要度が低くなってしまったと考えられる．BERT などを用いて発話内容を高次元ベクトルに変換した特徴量を説明変数として用いることで，言語的特徴量の重要度が上がり，理

*ある特徴量がある決定木の分岐により得た目的変数の改善幅の評価

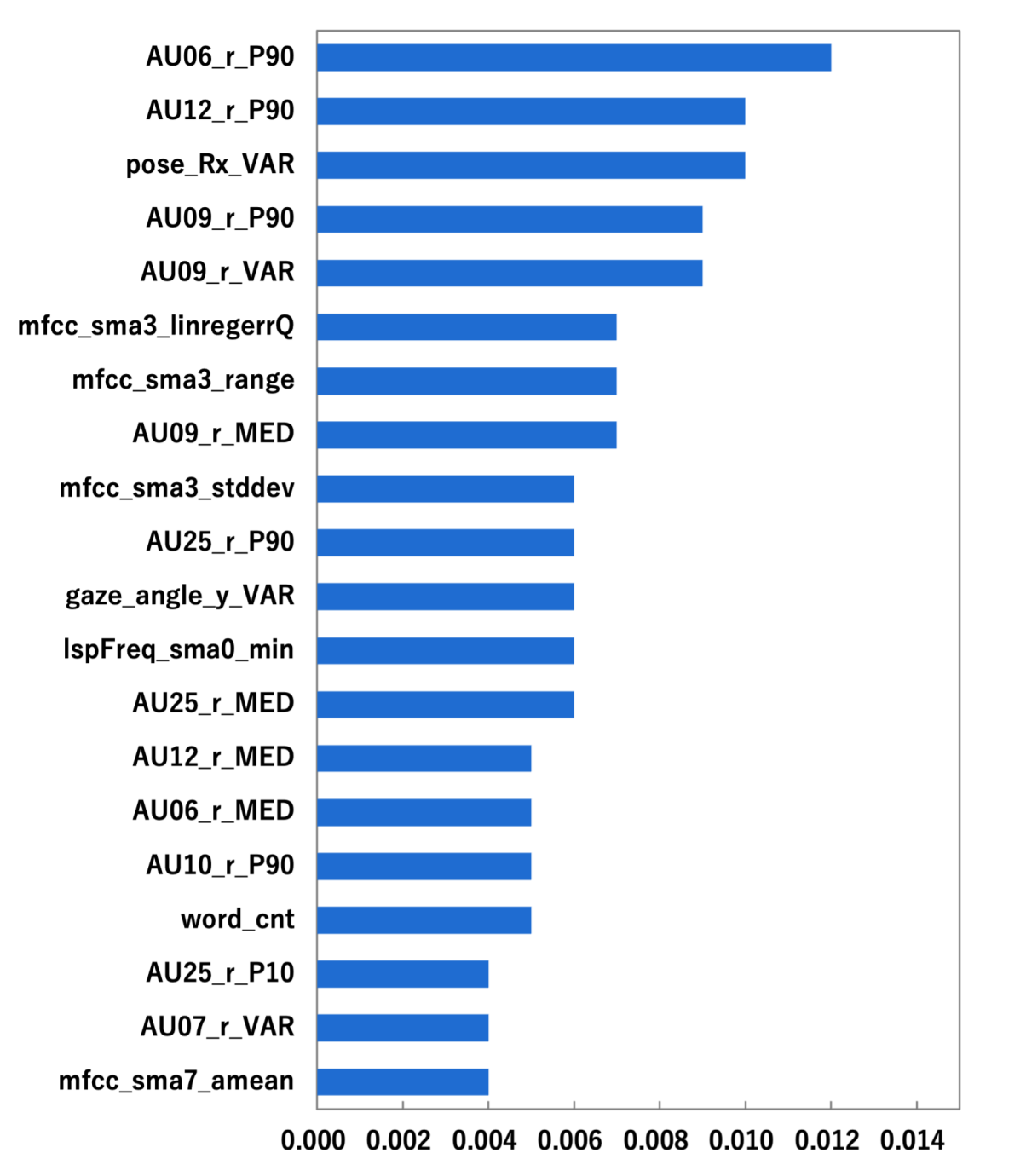

図 5.4: GB モデルにおける特徴量の重要度上位 20 件

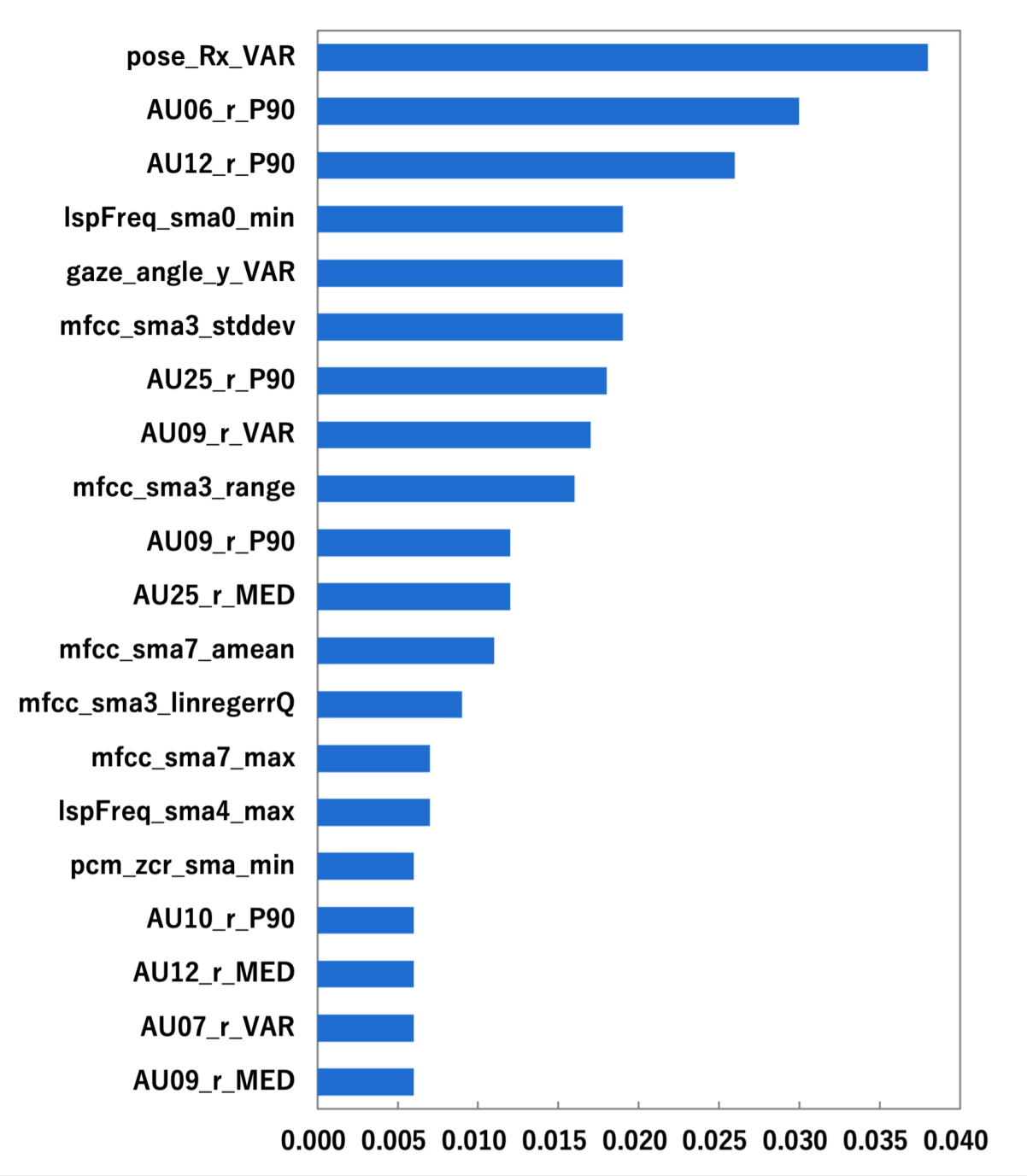

図 5.5: RF モデルにおける特徴量の重要度上位 20 件

解度の推定精度も向上するかもしれない．

5.5　クラス分類モデル

5.5.1　クラス分類モデルの構築

本項では，聞き手の発話を理解度を理解をしている発話，中立の発話，理解をしていない発話に分類するクラス分類モデルの構築について述べる．クラス分類モデルは，聞き手と話し手の双方の視覚的，韻律的，言語的特徴から聞き手の発話が理解をしている発話，中立の発話，理解をしていない発話なのかを分類するモデルである．クラス分類モデルの概要を図 5.6 に示す．聞き手の発話時の理解度の分布（図 5.2）を確認すると，本研究の理解度データは全体的に高めに評価されていることがわかる．そのため，回帰モデルでは理解していない発話を推測することが難しいと考えられる．理解していない発話を推定するために，理解度を 3 群に分け，クラス分類モデルを構築する．回帰モデルでは，初期検討として重要な特徴量を解釈することに重きを置き，言語特徴量として解釈しやすい品詞の出現頻度と単語数に関する特徴や，決定木ベースのアルゴリズムをモデル構築に用いていた．クラス分類モデルではより推定性能の高いモデルを構築するするため近年のマルチモーダルインタラクションの関連研究 [6, 7, 26] で多く用いられている手法を参考にした．言語特徴量として BERT を用いて抽出した高次元ベクトル表現に変換した特徴や，ニューラルネットワークベースの機械学習手法を用いる．

回帰モデルでは，聞き手の振舞いのみに着目していたが，聞き手の発話だけでなく，話し手の発話も用いる．そのため，推定する聞き手の発話に対応する話し手の発話の選定を行う．抽出する発話の選定は，話し手の発話開始時よりも聞き手の発話開始時が遅いという条件で行った．選定された発話は 3 つの発話パターン存在した (図 5.7)．**Pattern 1** は，聞き手の発話開始時が話し手の発話終了時よりも遅い発話パターンである．**Pattern 2** は，聞き手の発話開始時が話し手の発話途中であり，聞き手の発話終了時が話し手の発話終了時よりも遅い発話パターンである．**Pattern 3** は，聞き手の発話開始時と発話終了時が話し手の発話途中である発話パターンである．

本研究では，5.1 節に記載したように，聞き手の発話終了時点の理解度を推定する．そのため，聞き手の発話終了より前の話し手のマルチモーダル情報のみを用いることが適切だと考えた．本研究では聞き手の発話終了時よりも前に話し手の発話が終了する Pattern 1 と Pattern 2 の発話パターンを分析対象とする．発話の選定を行った結果，本研究で用いる聞き手の IPU は 11133 件となった．

目的変数

聞き手の発話を理解をしている発話，中立の発話，理解をしていない発話に分類するモデルを構築するために，5.1 節で述べた聞き手の発話時の理解度をもとに，理解クラス，非理解クラス，中立クラスの 3 つのクラスに分割した．3 クラスに分類した条件とそれぞれの群のデータ数を表 5.7 に示す．この 3 つのクラスを目的変数とする．

説明変数

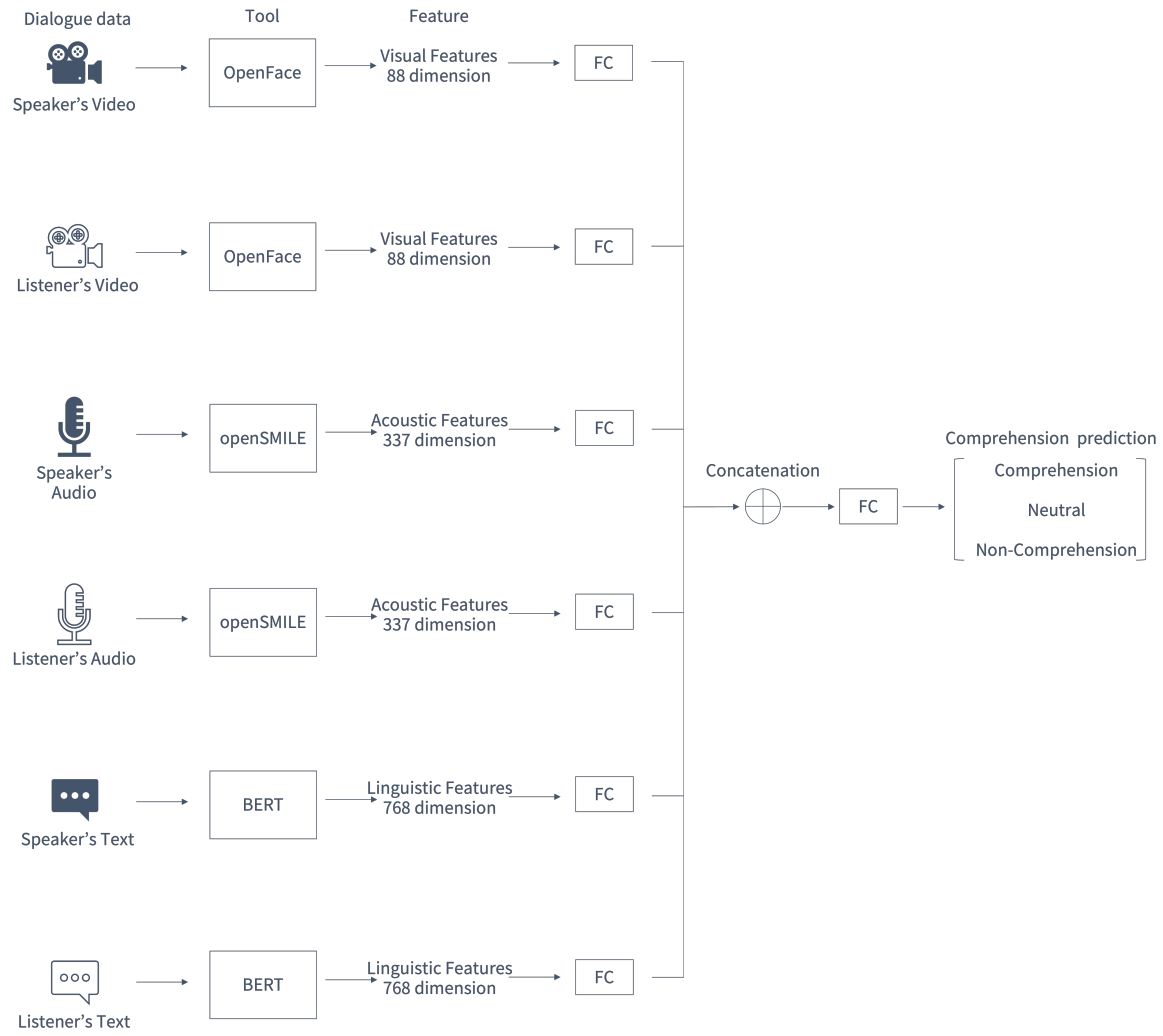

図 5.6: クラス分類モデルの概要

説明変数は 5.2 節で抽出した聞き手の視覚的，韻律的，言語的特徴である．回帰モデルでは，言語的特徴として品詞の発話頻度に関する特徴量を用いていた．しかし，5.4.3 項で述べたように，品詞の発話頻度に関する特徴量は理解度の推定に適さない可能性もある．そこでクラス分類モデルでは，より推定性能の高いモデルを構築するするため近年のマルチモーダルインタラクションの研究事例でも用いられている手法を参考にし，BERT を用いて抽出した高次元ベクトル表現に変換した特徴を用いた．

モデルの構築方法

モデルの構築は近年のマルチモーダルインタラクションの研究事例でも用いられている手法を参考にし．全結合型のニューラルネットワークで行った．全結合型のニューラルネットワークでは，まず，各特徴を全結合（FC）層を通じて 32 次元の埋め込みベクトルに変換する．次に，それぞれの変換された 32 次元のベクトルを結合し，192 次元のベクトルに融合する．最後に FC 層を通じて 3 種類の発話時の理解度のクラスを出力する．表

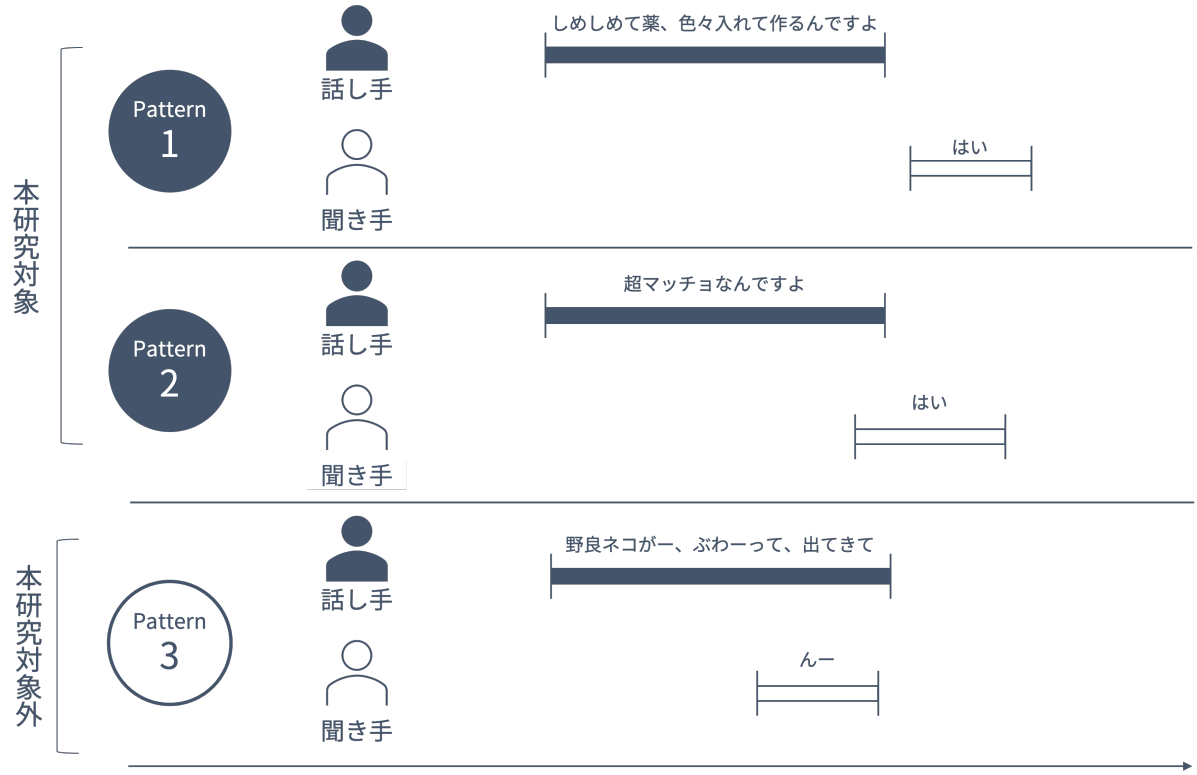

図 5.7: 発話パターン

5.7 を確認すると，理解度の各クラスのデータ数は不均衡である．そのため，各クラスに対し各クラス数に反比例した値を元にした重みを付与した．本研究では，対話参加者のマルチモーダル情報と聞き手の理解度の関係を分析し，理解度を推定するために重要なマルチモーダル情報を明らかにする．そのため，聞き手と話し手の双方のそれぞれのモダリティを組み合わせた 63 種類の機械学習モデルを構築した．

表 5.7: クラスの分類条件と件数

クラス	分割条件	件数	割合
理解クラス	アノテータ 2 名以上が 理解している (+1 以上の評価) と評価された発話	6,684 件	60.04%
非理解クラス	アノテータ 2 名以上が 理解していない (-1 以下の評価) と評価された発話	460 件	4.13%
中立クラス	理解群，非理解クラスの どちらの条件にも該当しない発話	3,989 件	35.83%

5.5.2　クラス分類モデルの推定結果

本項では，5.5.1 項で構築した話し手と聞き手の双方のマルチモーダル情報から聞き手の発話時の理解度を 3 クラスに分類するモデルの推定結果について述べる．モデルの推定結果を分析するため，次のタスクを 100 回ずつ行った．

(1) 特徴量が含まれているデータセットを訓練データとテストデータを 9：1 に無作為に分割する．

(2) 訓練データで構築したモデルを構築したモデルを用いてテストデータにおける目的変数を推定するタスクを行う．

本研究のクラス分類モデルで用いる Baseline モデルは，表 5.7 で示した各クラスの割合に応じてランダムにクラスを予測するモデルである．上記で述べたタスクを 100 回行った結果を表 5.8 に示す．

Baseline モデルの F 値とそれぞれのモダリティを用いたモデルの F 値で，1%水準で Holm 法による補正を用いて対応のない t 検定を行った．その結果，本研究で作成したすべてのクラス分類モデルで Baseline モデルと有意差を認められることが確認できた．このことから，本研究で作成した全てのクラス分類モデルが，Baseline モデルよりも性能を向上させることができたと判断できる．この結果から，聞き手と話し手のマルチモーダル情報が聞き手の発話を理解をしている発話，中立の発話，理解をしていない発話の分類に有効であることが示された．

5.5.3　クラス分類モデルに関する考察

本項では，5.5.2 項で述べたクラス分類モデルの結果から話し手と聞き手の双方のマルチモーダル情報と聞き手の理解度の関係に関する考察を述べる．具体的には，聞き手と話し手の双方のそれぞれのモダリティを組み合わせた 63 種類のクラス分類モデルを構築しそれぞれの F 値の比較を行う．F 値の比較は，単一のモダリティに関する情報を用いたユニモーダルモデル，聞き手のマルチモーダル情報を用いたモデル，F 値が高かった上位 5 件のモデルに着目する．

聞き手と話し手のそれぞれの単一のモダリティに関する情報を用いたユニモーダルモデルの結果の比較を行う．ユニモーダルモデルの結果を表 5.9 に示す．全てのモデルの F 値が Baseline モデルの F 値を上回る結果が確認できた．このことから，聞き手と話し手の双方の単一のモダリティに関する情報だけを用いても，聞き手の発話を理解をしている発話，中立の発話，理解をしていない発話に分類することができることが示唆された．聞き手の単一のモダリティに関する情報を用いたモデル同士の F 値を比較したところ，聞き手の韻律的特徴を用いたモデル（M02）の F 値が最も高いことが確認できた．この結果から，聞き手の発話を理解をしている発話，中立の発話，理解をしていない発話に分類するタスクにおいて，聞き手の韻律的特徴が重要であることが示唆された．話し手の単一のモダリティに関する情報を用いたモデルに着目すると，話し手の韻律的特徴を用いたモデル

表 5.8: クラス分類モデルの推定結果の平均値（N = 100）

	Listener			Speaker					
Model	Visual	Acoustic	Linguistic	Visual	Acoustic	Linguistic	precision	recall	f1-score
Baseline							0.461	0.520	0.452
M01	✓						0.583	0.444	0.491
M02		✓					0.606	0.536	0.563
M03			✓				0.577	0.513	0.531
M04				✓			0.540	0.414	0.459
M05					✓		0.544	0.476	0.503
M06						✓	0.519	0.465	0.487
M07	✓	✓					0.615	0.550	0.574
M08	✓		✓				0.597	0.550	0.567
M09	✓			✓			0.596	0.482	0.522
M10	✓				✓		0.577	0.517	0.541
M11	✓					✓	0.558	0.512	0.531
M12		✓	✓				0.606	0.564	0.580
M13		✓		✓			0.604	0.540	0.564
M14		✓			✓		0.601	0.556	0.573
M15		✓				✓	0.584	0.548	0.563
M16			✓	✓			0.578	0.530	0.546
M17			✓		✓		0.579	0.539	0.554
M18			✓			✓	0.563	0.529	0.542
M19				✓	✓		0.552	0.494	0.516
M20				✓		✓	0.532	0.485	0.504
M21					✓	✓	0.535	0.493	0.510
M22	✓	✓	✓				0.621	0.584	0.598
M23	✓	✓		✓			0.618	0.560	0.582
M24	✓	✓			✓		0.615	0.573	0.589
M25	✓	✓				✓	0.597	0.565	0.578
M26	✓		✓	✓			0.600	0.556	0.572
M27	✓		✓		✓		0.602	0.563	0.578
M28	✓		✓			✓	0.586	0.560	0.571
M29	✓			✓	✓		0.581	0.530	0.550
M30	✓			✓		✓	0.560	0.517	0.534
M31	✓				✓	✓	0.567	0.531	0.546
M32		✓	✓	✓			0.615	0.578	0.593
M33		✓	✓		✓		0.612	0.581	0.593
M34		✓	✓			✓	0.602	0.577	0.587
M35		✓		✓	✓		0.605	0.563	0.579
M36		✓		✓		✓	0.589	0.556	0.569
M37		✓			✓	✓	0.587	0.558	0.570
M38			✓	✓	✓		0.582	0.544	0.559
M39			✓	✓		✓	0.574	0.544	0.555
M40			✓		✓	✓	0.570	0.545	0.555
M41				✓	✓	✓	0.544	0.506	0.522
M42	✓	✓	✓	✓			0.619	0.585	0.598
M43	✓	✓	✓		✓		0.622	0.593	0.604
M44	✓	✓	✓			✓	0.612	0.589	0.598
M45	✓	✓		✓	✓		0.614	0.575	0.590
M46	✓	✓		✓		✓	0.604	0.572	0.585
M47	✓	✓			✓	✓	0.599	0.573	0.584
M48	✓		✓	✓	✓		0.606	0.570	0.584
M49	✓		✓	✓		✓	0.593	0.568	0.578
M50	✓		✓		✓	✓	0.593	0.570	0.579
M51	✓			✓	✓	✓	0.570	0.535	0.549
M52		✓	✓	✓	✓		0.617	0.587	0.599
M53		✓	✓	✓		✓	0.608	0.585	0.594
M54		✓	✓		✓	✓	0.603	0.583	0.591
M55		✓		✓	✓	✓	0.588	0.562	0.573
M56			✓	✓	✓	✓	0.579	0.554	0.564
M57	✓	✓	✓	✓	✓		0.624	0.598	0.608
M58	✓	✓	✓	✓		✓	0.614	0.595	0.602
M59	✓	✓	✓		✓	✓	0.613	0.593	0.601
M60	✓	✓		✓	✓	✓	0.601	0.575	0.585
M61	✓		✓	✓	✓	✓	0.597	0.577	0.585
M62		✓	✓	✓	✓	✓	0.610	0.589	0.597
M63	✓	✓	✓	✓	✓	✓	0.615	0.596	0.604

表 5.9: ユニモーダルモデルの結果の平均値（N = 100）

	Listener			Speaker					
Model	Visual	Acoustic	Linguistic	Visual	Acoustic	Linguistic	precision	recall	f1-score
Baseline							0.461	0.520	0.452
M01	✓						0.583	0.444	0.491
M02		✓					0.606	0.536	**0.563**
M03			✓				0.577	0.513	0.531
M04				✓			0.540	0.414	0.459
M05					✓		0.544	0.476	0.503
M06						✓	0.519	0.465	0.487

（M05）の F 値が話し手の視覚，言語的特徴を用いたモデル（M04，M06）よりも高いことが確認できる．2.1 節で述べたプレゼンテーションにおける聴衆の理解度を推定する研究 [11] では，聴衆の理解度推定には講演者の韻律的特徴が重要である結果が示されている．上記の研究では，本研究と同じように，説明を聞いている聴衆 (聞き手) の理解度の推定を行っている．そのため，話し手から説明を受けるシーンにおいて聞き手の理解度を推定する際には，対話参加者の韻律的特徴が重要である可能性が示唆される．この知見は他者から説明を受ける対話シーンに限定されているため，ディベートや雑談といった他の対話シーンでも理解度の推定に韻律的特徴が重要であるのかは明らかではない．

5.4.3 項で重要だと示した，笑顔の表情などの特徴を含む，聞き手の視覚的特徴を用いたモデル（M01）は，聞き手の韻律，言語的特徴を用いたモデル（M02，M03）よりも F 値が低い結果となった．回帰モデルの目的変数で用いる理解度は，聞き手の理解度のアノテータ 3 名の平均値を使用している．それに対し，クラス分類モデルでは，5.5.1 項で述べたように，目的変数で用いている理解度は，アノテータの評価の方向に注目した多数決で決めており，+1（少し理解している）と+2（とても理解している）や，-1（あまり理解していない）と-2（全く理解していない）の評価が統合されてしまっている．そのため，クラス分類モデルでは，どの程度理解しているのか推定していない．したがって，5.4.3 項で着目した笑顔の振舞いは，どの程度理解しているのか推定するタスクにおいて特に重要であるが，聞き手の発話を理解をしている発話，中立の発話，理解をしていない発話なのかを分類するタスクでは重要ではない可能性が示唆される．

聞き手のマルチモーダル情報のみを用いたモデルの結果の比較を行う．聞き手のマルチモーダル情報のみを用いたモデルの結果を表 5.10 に示す．聞き手のモダリティに関する情報を組み合わせたマルチモーダルモデル（M07，M08，M12，M22）の F 値と聞き手の単一のモダリティに関する情報のみを用いているユニモーダルモデル（M01，M02，M03）の F 値を比較したところ，全てのマルチモーダルモデルがユニモーダルモデルよりも F 値が高いことが確認された．対話における興味度や発話意欲やエンゲージメントを推定する研究事例 [6, 7, 9] においても，複数のモダリティを組み合わせたモデルの推定性能が単一のモダリティを用いたモデルの推定性能よりも向上していることを示している．このことから，聞き手の複数のモーダル情報を組み合わせることで聞き手の発話を理解をしている発話，中立の発話，理解をしていない発話に分類するモデルの性能が向上することが示

表 5.10: 聞き手のマルチモーダル情報を用いたモデルの結果の平均値（N = 100）

	Listener			Speaker					
Model	Visual	Acoustic	Linguistic	Visual	Acoustic	Linguistic	precision	recall	f1-score
Baseline							0.461	0.520	0.452
M22	✓	✓	✓				0.621	0.584	**0.598**
M12		✓	✓				0.606	0.564	0.580
M07	✓	✓					0.615	0.550	0.574
M08	✓		✓				0.597	0.550	0.567
M02		✓					0.606	0.536	0.563
M03			✓				0.577	0.513	0.531
M01	✓						0.583	0.444	0.491

唆される．

表 5.11: F 値が高かった上位 5 件のモデルの結果の平均値（N = 100）

	Listener			Speaker					
Model	Visual	Acoustic	Linguistic	Visual	Acoustic	Linguistic	precision	recall	f1-score
Baseline							0.461	0.520	0.452
M57	✓	✓	✓	✓	✓		0.624	0.598	0.608
M43	✓	✓	✓		✓		0.622	0.593	0.604
M63	✓	✓	✓	✓	✓	✓	0.615	0.596	0.604
M58	✓	✓	✓	✓		✓	0.614	0.595	0.602
M59	✓	✓	✓		✓	✓	0.613	0.593	0.601

F 値が高かった上位 5 件のモデルに使用されているマルチモーダル情報に着目する．F 値が高かった上位 5 件のモデルの結果を表 5.11 に示す．全てのモデルに共通する点は，聞き手の視覚的，韻律的，言語的特徴を用いていること，話し手のモダリティに関する情報を用いていることである．この結果から，聞き手の視覚的，韻律的，言語的情報の全ての情報を用いることで聞き手の発話を理解をしている発話，中立の発話，理解をしていない発話に分類するタスクの分類精度が向上することが示唆される．さらに，対話におけるターンマネジメントの意欲を推定する研究 [7] や対話における重要発言を推測する研究 [22] においても，本研究と同様に，複数参加者のモダリティを用いたモデルの推定性能が参加者一人のモダリティのみを用いたモデルの推定性能よりも向上していることを示している．そのため，聞き手の視覚的，韻律的，言語的情報だけではなく，話し手のマルチモーダル情報を組み合わせることにより聞き手の発話を理解をしている発話，中立の発話，理解をしていない発話に分類するタスクの分類精度が向上することが示唆される．

本研究で構築したモデルの各クラス毎の推定性能を確認する．構築したモデルの中で F 値が最も高かったモデル（M57）の Confusion Matrix を表 5.12 に示す．正解データが理解発話の推測結果に着目すると，M57 のモデルは 65%程度の確率で正しく理解発話を推定することができており，中立発話と誤って推定する傾向があることを確認できる．正解データが中立発話の推測結果に着目すると，M57 のモデルは 55%程度の確率で正しく中

表 5.12: F 値が最も高かったモデル（M57）の Confusion Matrix の平均値（N=100）

		Predicted		
		理解発話	中立発話	非理解発話
Actual	理解発話	306.51	139.62	25.66
	中立発話	105.11	166.89	27.93
	非理解発話	9.31	17.04	8.93

立発話を推定することができていることが確認できる．上記の結果から，M57 のモデルは理解発話と中立発話を 55%以上の確率で分類できていると確認できる．本研究で構築したモデルは，ユーザが理解できない発話を検出し，ユーザの理解度に応じて適切なコミュニケーションを行う対話エージェントの開発に貢献することが期待される．しかし，正解データが非理解発話の推測結果に着目すると，M57 のモデルは非理解発話を他のクラスと同程度の精度で推定することができていないことが確認できる．非理解発話の推定性能が低い理由として，理解度の各クラスのデータ数が不均衡であることが考えられる．クラス分類モデルで使用した対話コーパスは，表 5.7 で示したように，非理解発話のデータが他のクラスデータよりも少ない．そのため，非理解発話を他クラスと同程度の精度で推定することができないモデルとなってしまったと考えられる．非理解発話も他クラスと同程度の精度で推定できるモデルを構築するためには，各クラスのデータ数の調整や各クラスに付与する重みの再調整を行う必要があると考えられる．

5.6 本研究で得られた知見

本節では，本研究で得られた知見を述べる．

- 聞き手のマルチモーダル情報が聞き手の発話時の理解度の推定に有効であること．
- 話し手が聞き手にコミカルなアニメーションの内容を説明する対話において，聞き手の理解度を推定するためには，聞き手の笑顔に関する Action Units が重要であり，聞き手の理解度が高いとき，聞き手はより笑顔になることが示唆されたこと．
- 聞き手と話し手のマルチモーダル情報が聞き手の発話を理解をしている発話，中立の発話，理解をしていない発話の分類に有効であること．
- 聞き手の視覚的，韻律的，言語的特徴と話し手の視覚的，韻律的，言語的特徴のいずれか一つの情報から，聞き手の発話が理解をしている発話，中立の発話，理解をしていない発話なのか推定できること．
- 聞き手の発話を理解をしている発話，中立の発話，理解をしていない発話に分類するタスクでは対話参加者の韻律的特徴が重要であることが示唆されたこと．

- 聞き手のモダリティに関する情報を組み合わせることにより聞き手の発話を理解をしている発話，中立の発話，理解をしていない発話に分類するタスクの分類精度が向上することが示唆されたこと．
- 聞き手の視覚的，韻律的，言語的情報の全ての情報を用いることで聞き手の発話を理解をしている発話，中立の発話，理解をしていない発話に分類するタスクの分類精度が向上することが示唆されたこと．
- 聞き手の視覚的，韻律的，言語的情報と，話し手のマルチモーダル情報を組み合わせることが聞き手の発話を理解をしている発話，中立の発話，理解をしていない発話に分類する際に重要であることが示唆されたこと．

第6章　結論

本研究では，対話における聞き手と話し手の双方のマルチモーダル情報から聞き手の理解度を推定できるのか明らかにする取り組みを行った．具体的には，対話コーパスから聞き手と話し手の双方の視覚的，韻律的，言語的特徴と発話時の理解度を抽出し，聞き手の視覚的，韻律的，言語的特徴から聞き手の発話時の理解度を推定する回帰モデルと，聞き手と話し手の双方の視覚的，韻律的，言語的特徴から聞き手の発話を理解が話が理解している発話，理解していない発話，中立した発話なのかを推定するクラス分類モデルを構築した．回帰モデル構築の結果，聞き手のマルチモーダル情報が聞き手の発話時の理解度の推定に有効であることが示された．さらに，聞き手の理解度を推定するためには，聞き手の笑顔に関する Action Units が重要であり，聞き手の理解度が高いとき，聞き手はより笑顔になることが示唆された．クラス分類モデル構築の結果，聞き手と話し手のマルチモーダル情報が聞き手の発話を理解をしている発話，中立の発話，理解をしていない発話の分類に有効であることが示された．さらに，聞き手の視覚的，韻律的，言語的情報と，話し手の情報を組み合わせることが推定性能を向上するためには重要であることが示唆された．

本研究では，対話参加者のマルチモーダル情報から対話相手の理解度を推定できる知見が得られたが，いくつかの制約がある．1 つ目は，理解度の判断を聞き手本人が評価していないことである．聞き手が自認している理解度を正確に取得するためには，聞き手自身が理解度を評価を行う必要があると考えられる．そのため，本研究で得られた知見は，聞き手が話し手の発話を理解している様子を推定することに限定される．上記の知見でも，人間の理解度を考慮した対話エージェントの開発に期待できる．しかし，実際の聞き手自身が自認している理解度を推定できるのかは明らかになっていない．この問題を解決するには，聞き手自身が理解度の評価を行った対話コーパスを構築し，実際の聞き手自身が自認している理解度を推定できるのかは明らかにする必要がある，2 つ目は，今回の研究が対話エージェントの開発を行っていないことである．今回の研究では，理解度の推定する機械学習モデルの構築を行った．しかし，構築したモデルを搭載した対話エージェントの開発は行っていない．そのため，構築したモデルを搭載した対話エージェントがユーザにどのような影響を与えるのかは明らかになっていない．今後は，理解度を推定しながら対話を行うエージェントがユーザにどのような影響を与えるのか明らかにする必要がある．3 つ目は，聞き手に対し話し手が説明を行う対話シーンに限定されていることである．本研究で用いた対話データは参加者がステーリーテリングと呼ばれる物語の内容を説明するタスクを行っている．ストーリーテリングの性質上，対話データは話し手が聞き手へ説明を行っている様子となっている．そのため，本研究で得られた知見は他者から説明を受ける対話シーンに限定されており，ディベートや雑談など他の対話シーンにおいて有効であるかは明らかではない．今後は，様々な対話シーンにおいて本研究で明らかになった知見が有効なのか分析する必要がある．

謝辞

本研究は，日本電信電話株式会社 NTT 人間情報研究所との共同研究の成果である．

本研究における主査を務めてくださいました宮田章裕教授に感謝申し上げます．まず，イレギュラーなタイミングでの研究室への参加を受け入れてくださり，ありがとうございました．約 2 年間という他の人よりも短い期間の研究活動でしたが，先生のアドバイスのおかげで，研究への取り組み方を学ぶことができ，ドイツでの国際学会への参加といったとても濃厚な修士生活を送ることができました．ご指導いただき，ありがとうございます．

本研究における副査を務めてくださいました北原鉄朗教授，大澤正彦准教授に感謝申し上げます．ご多忙であるにも関わらず，研究の細部に至る所まで丁寧なご指導をいただき，ありがとうございます．また，大澤先生から本学の大学院への進学を提案していただいたことに感謝申し上げます．大澤先生のおかげで研究というものに触れることができました．ご指導いただき，ありがとうございます．

日本電信電話株式会社 NTT 人間情報研究所の石井亮様に感謝申し上げます．石井様から研究テーマの提案や研究データの提供をしていただき，また，とても丁寧な研究指導をしていただいたおかげで，2 年間研究を進めることができました．ありがとうございます．

宮田研究室で同じ研究チームであった大西俊輝さん，大串旭君，丸山葉君，東直輝君，岡哲平君，鹿摩大智君に感謝申し上げます．研究に対する悩みや研究の進め方などを気軽に相談させていただいたおかげで，様々なことにチャレンジしながら研究を進めることができました．また，研究以外の雑談や相談のおかげで，とても楽しみながらこの 2 年間過ごすことができました．ありがとうございます．

宮田研究室の全員に感謝申し上げます．私の研究室への参加がイレギュラーなタイミングでしたが，暖かく迎え入れてくださり，ありがとうございました．様々な研究議論や研究に関係のないコミュニケーションなどのおかげでとても有意義な修士生活を送ることができました．ありがとうございます．

参考文献

[1] Daniel McDuff and Mary Czerwinski. Designing emotionally sentient agents. *Communications of the ACM*, Vol. 61, No. 12, pp. 74–83, 2018.

[2] Mike Holmes, Annabel Latham, Keeley Crockett, and James D. O'Shea. Modelling e-learner comprehension within a conversational intelligent tutoring system. In *11th IFIP TC 3 World Conference on Computers in Education*, pp. 251–260, 2017.

[3] Mike Holmes, Annabel Latham, Keeley Crockett, and James D. O'Shea. Near real-time comprehension classification with artificial neural networks: Decoding e-learner non-verbal behavior. *IEEE Transactions on Learning Technologies*, Vol. 11, pp. 5–12, 2018.

[4] Fiona J. Buckingham, Keeley A. Crockett, Zuhair A. Bandar, James D. O'Shea, Kathleen. M. MacQueen, and Mario Chen. Measuring human comprehension from nonverbal behaviour using artificial neural networks. In *The 2012 International Joint Conference on Neural Networks*, pp. 1–8, 2012.

[5] Tatsuya Kawahara, Soichiro Hayashi, and Katsuya Takanashi. Estimation of interest and comprehension level of audience through multi-modal behaviors in poster conversations. In *Proceedings of Interspeech 2013*, pp. 1882–1885, 2013.

[6] Yuki Hirano, Shogo Okada, Haruto Nishimoto, and Kazunori Komatani. Multitask prediction of exchange-level annotations for multimodal dialogue systems. In *Proceedings of the 2019 International Conference on Multimodal Interaction(ICMI '19)*, pp. 85–94, 2019.

[7] Ryo Ishii, Xutong Ren, Michal Muszynski, and Louis-Philippe Morency. Trimodal prediction of speaking and listening willingness to help improve turn-changing modeling. *Frontiers in Psychology*, Vol. 13, pp. 1–17, 2022.

[8] Tomoya Ohba, Candy Olivia Mawalim, Shun Katada, Haruki Kuroki, and Shogo Okada. Multimodal analysis for communication skill and self-efficacy level estimation in job interview scenario. In *Proceedings of the 21st International Conference on Mobile and Ubiquitous Multimedia(MUM '22)*, pp. 110–120, 2022.

[9] Arthur Pellet-Rostaing, Roxane Bertrand, Auriane Boudin, Stéphane Rauzy, and Philippe Blache. A multimodal approach for modeling engagement in conversation. *Frontiers in Computer Science*, Vol. 5, pp. 1–14, 2023.

[10] Mohamed Sathik and Sofia G. Jonathan. Effect of facial expressions on student's comprehension recognition in virtual educational environments. *SpringerPlus*, Vol. 2, pp. 1–9, 2013.

[11] Keith Curtis, Gareth J.F. Jones, and Nick Campbell. Speaker impact on audience comprehension for academic presentations. In *Proceedings of the 2016 ACM International Conference on Multimodal Interaction(ICMI '16)*, pp. 129–136, 2016.

[12] Resmana Lim and MJT Reinders. Facial landmark detection using a gabor filter representation and a genetic search algorithm. In *Annual conference of the Advanced School for Computing and Imaging*, pp. 72–78, 2000.

[13] Bruce Perry. Can some people read minds. *Science World*, 2004.

[14] Keith Curtis, Gareth J.F. Jones, and Nick Campbell. Effects of good speaking techniques on audience engagement. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction(ICMI '15)*, pp. 35–42, 2015.

[15] Janet Rothwell, Zuhair Bandar, James O'Shea, and David McLean. Silent talker: a new computer-based system for the analysis of facial cues to deception. *Applied Cognitive Psychology*, Vol. 20, pp. 757–777, 2006.

[16] Tatsuya Kawahara, Hisao Setoguchi, Katsuya Takanashi, Kentaro Ishizuka, and Shoko Araki. Multi-modal recording, analysis and indexing of poster sessions. In *Proceedings of Interspeech 2008*, pp. 1622–1625, 2008.

[17] Tatsuya Kawahara. Multi-modal Sensing and analysis of poster conversations: Toward smart posterboard. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 1–9, 2012.

[18] Kazunori Komatani, Shogo Okada, Haruto Nishimoto, Masahiro Araki, and Mikio Nakano. Multimodal dialogue data collection and analysis of annotation disagreement. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*, pp. 201–213, 2021.

[19] Mary Amoyal, Béatrice Priego-Valverde, and Stéphane Rauzy. PACO: a corpus to analyze the impact of common ground in spontaneous face-to-face interaction. In *Language Resources and Evaluation Conference*, pp. 628–633, 2020.

[20] Béatrice Priego-Valverde, Brigitte Bigi, and Mary Amoyal. “cheese!”: a corpus of face-to-face French interactions. a case study for analyzing smiling and conversational humor. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 467–475, 2020.

[21] Fumio Nihei, Ryo Ishii, Yukiko I. Nakano, Atsushi Fukayama, and Takao Nakamura. Whether contribution of features differ between video-mediated and in-person meetings in important utterance estimation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2023.

[22] Fumio Nihei, Ryo Ishii, Yukiko Nakano, Kyosuke Nishida, Ryo Masumura, Atsushi Fukayama, and Takao Nakamura. Dialogue Acts Aided Important Utterance Detection Based on Multiparty and Multimodal Information. In *Proceedings of Interspeech 2022*, pp. 1086–1090, 2022.

[23] Ryo Ishii, Ryuichiro Higashinaka, and Junji Tomita. Predicting nods by using dialogue acts in dialogue. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC ’18)*, pp. 2940–2944, 2018.

[24] Shogo Okada, Mayumi Bono, Katsuya Takanashi, Yasuyuki Sumi, and Katsumi Nitta. Context-based conversational hand gesture classification in narrative interaction. In *Proceedings of the 2023 ACM on International conference on multimodal interaction(ICMI ’13)*, pp. 303–310, 2013.

[25] Akira Morikawa, Ryo Ishii, Hajime Noto, Atsushi Fukayama, and Takao Nakamura. Determining most suitable listener backchannel type for speaker’s utterance. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, pp. 1–3, 2022.

[26] Toshiki Onishi, Naoki Azuma, Shunichi Kinoshita, Ryo Ishii, Atsushi Fukayama, Takao Nakamura, and Akihiro Miyata. Prediction of various backchannel utterances based on multimodal information. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*, pp. 1–4, 2023.

[27] Hanae Koiso, Yasuo Horiuchi, Syun Tutiya, Akira Ichikawa, and Yasuharu Den. An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs. *Language and Speech*, Vol. 41, pp. 295–321, 1998.

[28] 山口貴史, 井上昂治, 吉野幸一郎, 高梨克也, Nigel G. Ward, 河原達也. 傾聴対話システムのための言語情報と韻律情報に基づく多様な形態の相槌の生成. 人工知能学会論文誌, Vol. 31, No. 4, pp. 1–10, 2016.

[29] 翠輝久, 水上悦雄, 志賀芳則, 川本真一, 河井恒, 中村哲. ユーザの相づち・うなずきを喚起する音声対話システム. 電子情報通信学会論文誌 A, Vol. 95, No. 1, pp. 16–26, 2012.

[30] Kikuo Maekawa. Corpus of spontaneous japanese: Its design and evaluation. In *Proceedings of The ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR '03)*, pp. 7–12, 2003.

[31] Laurence Devillers, Laurence Vidrascu, and Lori Lamel. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, Vol. 18, No. 4, pp. 407–422, 2005.

[32] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, Vol. 42, pp. 335–359, 2008.

[33] Dennis Reidsma and Rieks op den Akker. Exploiting 'subjective' annotations. In *Proceedings of the Workshop on Human Judgements in Computational Linguistics - HumanJudge '08*, pp. 8–16, 2008.

[34] Shiro Kumano, Kazuhiro Otsuka, Dan Mikami, Masafumi Matsuda, and Junji Yamato. Analyzing interpersonal empathy via collective impressions. *IEEE Transactions on Affective Computing*, Vol. 6, No. 4, pp. 324–336, 2015.

[35] Kazunori Komatani, Ryu Takeda, and Shogo Okada. Analyzing differences in subjective annotations by participants and third-party annotators in multimodal dialogue corpus. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL '23)*, pp. 104–113, 2023.

[36] Patrick E. Shrout and Joseph L. Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, Vol. 86, No. 2, pp. 420–428, 1979.

[37] Asahi Ogushi, Toshiki Onishi, Yohei Tahara, Ryo Ishii, Atsushi Fukayama, Takao Nakamura, and Akihiro Miyata. Analysis of praising skills focusing on utterance contents. In *Proceedings of Interspeech 2022*, pp. 2743–2747, 2022.

[38] Sowmya Rasipuram, Pooja Rao S. B., and Dinesh Babu Jayagopi. Asynchronous video interviews vs. face-to-face interviews for communication skill measurement: a systematic study. In *Proceedings of the 2016 ACM International Conference on Multimodal Interaction(ICMI '16)*, pp. 370–377, 2016.

[39] Shogo Okada, Yoshihiko Ohtake, Yukiko I. Nakano, Yuki Hayashi, Hung-Hsuan Huang, Yutaka Takase, and Katsumi Nitta. Estimating communication skills using dialogue acts and nonverbal features in multiple discussion datasets. In *Proceedings of the 2016 ACM International Conference on Multimodal Interaction(ICMI '16)*, pp. 169–176, 2016.

[40] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision*, pp. 1–10, 2016.

[41] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

[42] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, pp. 1459–1462, 2010.

[43] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 230–237, 2004.

[44] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT '19)*, pp. 4171–4186, 2019.

[45] Shunichi Kinoshita, Toshiki Onishi, Naoki Azuma, Ryo Ishii, Atsushi Fukayama, Takao Nakamura, and Akihiro Miyata. A study of prediction of listener's comprehension based on multimodal information. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*, pp. 1–4, 2023.

[46] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, Vol. 29, No. 5, pp. 1189–1232, 2001.

[47] Leo Breiman. Random forests. *Machine learning*, Vol. 45, pp. 5–32, 2001.

[48] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, Vol. 12, No. 85, pp. 2825–2830, 2011.

[49] James Bergstra, Daniel Yamins, and David D. Cox. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. In *Proceedings of the 12th Python in Science Conferences (SciPy '13)*, pp. 13–20, 2013.

[50] Seho Park, Kunyoung Lee, Jae-A Lim, Hyunwoong Ko, Taehoon Kim, Jung-In Lee, Hakrim Kim, Seong-Jae Han, Jeong-Shim Kim, Soowon Park, Jun-Young Lee, and Eui Chul Lee. Differences in facial expressions between spontaneous and posed smiles: Automated method by action units and three-dimensional facial landmarks. *Sensors*, Vol. 20, No. 4, 2020.

研究業績

査読付き国際会議

(1) Shunichi Kinoshita, Toshiki Onishi, Naoki Azuma, Ryo Ishii, Atsushi Fukayama, Takao Nakamura, and Akihiro Miyata: A Study of Prediction of Listener's Comprehension Based on Multimodal Information. Proc. the 23rd ACM International Conference on Intelligent Virtual Agents, Article No.30, pp.1–4 (IVA '23) (2023 年 9 月).

(2) Toshiki Onishi, Naoki Azuma, Shunichi Kinoshita, Ryo Ishii, Atsushi Fukayama, Takao Nakamura, and Akihiro Miyata: Prediction of Various Backchannel Utterances Based on Multimodal Information. Proc. the 23rd ACM International Conference on Intelligent Virtual Agents, Article No.47, pp.1–4 (IVA '23) (2023 年 9 月).

研究会・シンポジウム

(1) 東直輝, 大西俊輝, 木下峻一, 石井亮, 深山篤, 宮田章裕: マルチモーダル情報に基づく相槌種類予測の定性的評価. 情報処理学会研究報告コラボレーションとネットワークサービス (CN), Vol.2024-CN-121, No.29, pp.1–7 (2024 年 1 月)

(2) 丸山葉, 大西俊輝, 大串旭, 木下峻一, 石井亮, 深山篤, 大澤正彦, 宮田章裕: デジタルツインを用いた自己効力感向上システムの基礎検討. 情報処理学会コラボレーションとネットワークサービスワークショップ 2023 論文集, Vol.2023, pp.88–89 (2023 年 11 月).

(3) 東直輝, 大西俊輝, 木下峻一, 石井亮, 深山篤, 中村高雄, 宮田章裕: マルチモーダル情報に基づく多様な相槌の予測の検討. 情報処理学会シンポジウム論文集, マルチメディア、分散、協調とモバイル (DICOMO '23), pp.352–358 (2023 年 7 月).

(4) 木下峻一, 大西俊輝, 東直輝, 石井亮, 深山篤, 中村高雄, 大澤正彦, 宮田章裕: マルチモーダル情報に基づく聞き手の理解度推定の基礎検討. 情報処理学会研究報告グループウェアとネットワークサービス (GN), Vol.2023-GN-119, No.7, pp.1–6 (2023 年 3 月).

(5) 東直輝, 大西俊輝, 木下峻一, 石井亮, 深山篤, 中村高雄, 宮田章裕: マルチモーダル情報に基づく多様な相槌の生成の基礎検討. 情報処理学会研究報告グループウェアとネットワークサービス (GN), Vol.2023-GN-119, No.8, pp.1–6 (2023 年 3 月).

(6) 大西俊輝, 木下峻一, 東直輝, 石井亮, 深山篤, 中村高雄, 宮田章裕: マルチモーダル情報に基づく聞き手のバックチャネルの種類推定の基礎検討. 情報処理学会グループウェアとネットワークサービスワークショップ 2022 論文集, Vol.2022, pp.64–66 (2022 年 11 月).

(7) 木下峻一, 宮田章裕, 大澤正彦: ハイタッチが記憶に与える影響の調査. HAI シンポジウム 2022 予稿集, P-12 (2022 年 3 月).

受賞

(1) マルチメディア、分散、協調とモバイル (DICOMO '23) シンポジウム 優秀論文賞, マルチモーダル情報に基づく多様な相槌の予測の検討, 受賞者：東直輝, 大西俊輝, 木下峻一, 石井亮, 深山篤, 中村高雄, 宮田章裕 (2023 年 9 月).

(2) 情報処理学会第 119 回グループウェアとネットワークサービス研究会 優秀発表賞, マルチモーダル情報に基づく聞き手の理解度推定の基礎検討, 受賞者：木下峻一 (2023 年 3 月).