マルチモーダル情報に基づく
多様な相槌の予測

令和5年度　卒業論文

日本大学　文理学部　情報科学科　宮田研究室

東　直輝

概要

対話において聞き手の相槌は，重要な要素の一つである．適切な相槌を打つことで，対話を円滑に進めることが可能である．これより，対話型エージェントが適切な相槌を打つことができるようになると，ユーザとの円滑な対話が実現すると考えられる．近年，自然な相槌を打つ対話型エージェントを実現するための研究は多く行われている．これらの研究では，マルチモーダル情報に基づいて相槌の発生やタイミングの予測，数種類（反応，表現，笑いなど）の相槌の生成を行っている．しかし，話し手の発話時のマルチモーダル情報から聞き手の多様な相槌を生成できるのか明らかでない．そこで本稿では対話における話し手のマルチモーダル情報から聞き手の多様な相槌の生成ができるのか明らかにするための検討を行う．具体的には話し手の発話時のマルチモーダル情報と聞き手の相槌からなるコーパスを作成し，機械学習を用いて話し手の発話から聞き手の相槌を予測する．先行研究では聞き手の相槌が行われている場合のみに着目したが，本稿では聞き手の相槌が行われていない場合も考慮して相槌の予測を行った．その結果，特徴量としてモダリティを単体で用いるよりも，複数用いることで推定性能が向上することを確認できた．

目 次

第 1 章　序論 **1**
1.1　研究の背景 2
1.2　研究の目的 2
1.3　本論文の構成 2

第 2 章　対話中の振る舞いから相槌を予測・生成する研究事例 **3**
2.1　相槌の予測に関する研究事例 4
2.2　相槌の生成に関する研究事例 4

第 3 章　研究課題 **5**
3.1　問題の定義 6
3.2　研究課題の設定 6

第 4 章　対話コーパス **7**
4.1　2 者対話データについて 8
4.2　相槌ラベルについて 8

第 5 章　予測モデルの構築 **9**
5.1　特徴量抽出 10
5.1.1　視覚的特徴量 10
5.1.2　韻律的特徴量 10
5.1.3　言語的特徴量 11
5.2　モデル構築 11

第 6 章　構築した機械学習モデルの分析 **14**
6.1　相槌の推定結果について 15
6.2　考察 15
6.2.1　モデルの性能についての考察 15
6.2.2　相槌の有無の予測に関する考察 16

第 7 章　結論 **18**

謝辞 **20**

参考文献 **22**

付録 **25**

研究業績 **27**

図 目 次

5.1 目的変数の抽出区間 12
5.2 相槌予測モデルの簡略図 13

A.1 付録 26

表 目 次

5.1 Action Units の内容 10
5.2 音声特徴量の内容の内容 11
5.3 抽出した統計量 11
5.4 データセット中の各相槌ラベルの数 12

6.1 各モデルの推定性能の各指標の平均値（N=100） 15
6.2 Model G における相槌ラベル別の推定性能の各指標の平均値（N=100） . . 17
6.3 各モデルの相槌の有無の推定性能の各指標の平均値（N=100） 17

第1章　序論

1.1 研究の背景

聞き手が「うんうん」,「はい」,「すごい」などの相槌を打つことは，円滑な対話を実現するために重要である．多くの研究では対話システムを作るために様々な方法で相槌を予測・生成する取り組みを行っている [1]，[2]，[3]，[4]，[5]，[6]，[7]，[8]，[9]，[10]．日本語における相槌のための言葉は豊富であり [11]，機能面に着目して相槌を分類する研究 [12] や，言語的・機能的側面に着目して話者の発話意図・種類と聞き手の相槌の関係を分類する研究 [13] がある．この背景から人と自然な対話を行う対話エージェントを実現する場合においても，ユーザの発言に対して適切な相槌を打つ仕組みが重要であることに異論の余地は少ないだろう．近年の典型的な対話エージェントは，話し手の振る舞いの言語的特徴（例：発言内容）や非言語的特徴（例：表情，声使い）をマルチモーダル分析した結果に基づいて返答の生成を行う [14]．しかしながら，話し手のマルチモーダル情報から機能面を考慮した適切な相槌を生成しようとする試みは我々が調査した範囲では見つかっていない．そこで我々は，自然な相槌を打つ対話エージェントの実現を視野に入れ，対話中の話し手のマルチモーダル情報から機能面を考慮した適切な相槌を生成する取り組みを行う．本研究が属するマルチモーダルインタラクションの分野では，特定のタスクやシーンにおける話者の行動や能力を生成する取り組みを行う際に，事前検討や準備として話者の行動や能力を予測する取り組みを行うことが多い．そのため本研究でも機能面に関して適切な相槌を生成する事前検討として，機能面に関して適切な相槌を予測する取り組みを行う．

1.2 研究の目的

1.1 節より，本稿では，機能面を考慮した適切な相槌を予測する取り組みを行う具体的には，既存の対面での 2 者対話コーパスを利用し，話し手の言語・非言語から聞き手の相槌を予測する機械学習モデルの構築を行う．これにより，対面における話し手の振る舞いから聞き手の相槌を予測できるようになると考えられる

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，XXX に関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，本論文における提案手法を述べる．
5 章では，XXX に関する実装について述べる．
6 章では，XXX に関する評価実験・考察について述べる．
最後に 7 章にて，本論文の結論を述べる．

第2章　対話中の振る舞いから相槌を予測・生成する研究事例

本章では，対話中の振る舞いから相槌を予測・生成する研究事例について述べる．これらは，言語的・非言語的行動を利用して．対話中の聞き手の相槌を予測・生成するという点で本研究と関係している．2.2 節では，相槌の予測に関する研究事例について紹介する．??節では，相槌の生成に関する研究事例について紹介する．

2.1 相槌の予測に関する研究事例

対話中の話し手の言語・非言語情報を利用して聞き手の相槌を予測する研究が多く行われている．これらの研究には，韻律的情報を用いる研究 [7]，[5]，[4]，[10]，[3]，言語的情報を用いる研究 [10]，[3]，視覚的情報を用いる研究 [4]，[3] がある．Ward は韻律的情報に基づき，ルールベースで相槌の予測を行うモデルを構築した [7]．Ward は，日本語対話において話者の低音域が相槌の予測の際に重要な手がかりになることを示唆していた．Mueller らは韻律的情報に基づき，ニューラルネットワークを利用して相槌の予測を行うモデルを構築した [5]．Mueller らは予測モデルにおけるネットワークの層を増やすことで性能が高くなることを示唆していた．上記は単体のモダリティの情報を用いた研究であるが，複数のモダリティの情報を用いた研究も多く存在する．Morency らは韻律的，視覚的情報を用いて相槌を予測する確率モデルを構築した [4]．Ruede らは韻律的，言語的情報を用いて相槌を予測している [10]．Ruede らの論文では韻律的情報に加えて言語的情報を用いることでモデルの性能が向上したことが報告されていた．Ishii らは，マルチモーダル情報（韻律的，言語的，視覚的情報）を用いて，話者交代・話者交代の管理意欲の予測を相槌の予測と同時に行うモデルを構築した [3]．

2.2 相槌の生成に関する研究事例

聞き手の相槌の生成を試みようとする研究も多く存在する [15]，[16]，[17]．dermouche らは Intaraction Loop LSTM モデルを用いてエージェントとユーザの両方のマルチモーダル行動（頭の回転、視線、笑顔）と対話の会話状態（誰が話しているか）を入力として，エージェントの頭の回転、視線、笑顔を生成するモデルを構築した [15] Mori らは対話の会話状態（誰が話しているか）と感情の状態を入力として，話者の個性と感情の状態を考慮した笑いの生成モデルを構築した [16]．Lala らは韻律的情報を用いて相槌が行われるタイミングと 3 つのカテゴリに分類された相槌を予測するモデルをトレーニングし，リアルタイムで動作する相槌の生成モデルを構築した [17]．

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

聞き手が「うんうん」,「はい」,「すごい」などの相槌を打つことは，円滑な対話を実現するために重要である．相槌に関する研究では，相槌を機能的側面から分類する研究 [11] や，言語的・機能的側面に着目して話者の発話意図・種類と聞き手の相槌の関係を分類する研究 [13] がある．そして 2 章で述べたように，相槌の予測・生成を行う研究が数多く行われている．しかし，これらの研究の中で相槌の機能面に着目して予測・生成している取り組みは我々が調査した範囲では見つかっていない．これより，話し手の発話時の情報から聞き手の機能面を考慮した適切な相槌を予測・生成できるのかについては明らかにされていないという問題がある．

3.2　研究課題の設定

3.1 節で述べたように，話し手の発話時の情報から聞き手の機能面を考慮した適切な相槌を予測・生成できるのかについては明らかにされていないという問題がある．本研究が属するマルチモーダルインタラクションの分野では，特定のタスクやシーンにおける話者の行動や能力を生成する取り組みを行う際に，事前検討や準備として話者の行動や能力を予測する取り組みを行うことが多い．そのため本研究でも機能的に適切な相槌を生成する事前検討として，機能的に適切な相槌の種類を予測する取り組みを行う．これより本研究では話し手の発話時の情報から聞き手の機能面を考慮した適切な相槌を予測できるのか明らかにすることを研究課題として設定する．具体的には既存の 2 者対話コーパスを利用し，機能的な側面から分類した相槌の種類を予測する機械学習モデルを構築する．

第4章　対話コーパス

本章では，本論文における対話コーパスについて述べる．本研究では既存の 2 者対話コーパス [13] を利用した．この対話コーパスは，2 者対話データ [18] と相槌の種類を表す相槌ラベルが含まれている．

4.1　2 者対話データについて

2 者対話の参加者は，合計で 26 名（異なるペアを 13 組）であり，初対面の日本人男女である．参加者はお互いに向き合って座り，対話を行った．発話を含んだ相槌のデータをより多く収集するため，一方の参加者（話し手）が他方の参加者（聞き手）にアニメ「トムとジェリー」の内容を説明するタスクを行っている．発話の単位は Inter-pausal units (IPU)[19] 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．この対話データでは，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．

4.2　相槌ラベルについて

対話に参加していない第三者のアノテータ 3 名が，聞き手の発話ごとに下記に示す相槌ラベルを付与している．なお，アノテータは 1 つの発話に対し複数の相槌ラベルを付与することが許可されている．

- N（Neutral word）:「うん」,「はい」,「おお」など話し手への感情を含まない応答.
- P（Positive word）:「うんうん」,「そうそう」,「それいい」,「なるほど」,「たしかに」など話し手への肯定的な応答.
- NP（Non-positive word）:「うーん」,「ふーん」,「はーん」,「あー」,「へー」,「んー」など話し手への否定的または悩んでいるような応答.
- E（Emotional word）:「すごい」,「ふふ」,「あぁ」,「へぇ」，その他短い感嘆詞など感情の動きを表しているような応答.
- A（Anticipation）：話し手の話題を先取りしている応答.
- C（Confirmation）:「えっ」,「はっ」,「あっ」,「なんで」など確認を促す，質問するような応答.
- R（Repetition of speaker’s utterance）：話し手の発言を繰り返す応答.
- S（Summary of speaker’s utterance）：話し手の発話の要約，および言い換えをしているような応答.
- O（Other）：聞き手の感想や独り言など，他に該当するラベルがない応答.

第5章　予測モデルの構築

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		

本章では，予測モデルの構築について述べる．予測モデルの推論の一連の流れを図 5.2 に示す．はじめに視覚的，韻律的，言語的特徴量を抽出する．次に，相槌ラベルを予測する機械学習モデルの構築を行う．

5.1 特徴量抽出

5.1.1 視覚的特徴量

対話データの映像から顔画像処理ツールである OpenFace[20] を用いて話し手の頭部，視線，Action Units[21] に関する特徴量を抽出した．頭部に関する特徴量としては，話者を正面から撮影した映像データにおいて，カメラ側から見て左から右方向を x 軸，下から上方向を y 軸，手前から奥方向を z 軸として頭部の x 軸，y 軸，z 軸周りの回転角度（pose_Rx，pose_Ry，pose_Rz）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．視線に関する特徴量としては，話者を正面から撮影した映像データにおいて，カメラ側から見て左から右方向を x 軸，下から上方向を y 軸，として視線の x 軸，y 軸方向の角度（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 に関する特徴量としては，OpenFace で用いられている各 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．視覚的特徴量は 88 次元となった．

5.1.2 韻律的特徴量

話し手の発話から音声情報処理ツールである openSMILE[22] を用いて代表的な韻律的特徴量を抽出した．具体的には表 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 に示す統計量を算出した．これらの特徴量は標準セット [23] として提供されているものである．韻律的特徴量は 336 次元となった．

表 5.2: 音声特徴量の内容の内容

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		

5.1.3 言語的特徴量

話し手の発話から，自然言語処理モデルである BERT[24] 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，話し手の発話を 768 次元のベクトルに変換し，言語的特徴量とした．

5.2 モデル構築

話し手の発話終了前後における聞き手の相槌を予測するために機械学習モデルの構築を行った．モデルの構築には全結合型ニューラルネットワークを用いた．また，モデルのパラメータの調整には Hyperopt[25] を用いた．説明変数を 5.1 節で抽出した話し手の発話に関する特徴量とした．目的変数を話し手の発話終了時点の 0.5 秒前から 1 秒後の間の聞き手の各相槌ラベルの有無とした（図 5.1）．本稿では，3 名いるアノテータのうち，2 名以上のアノテータが付与している相槌ラベルのみを抽出した．各相槌ラベルの数は表 5.4 の通りとなった．加えて，今回のモデル構築においては，聞き手の相槌が行われなかった場合のデータも学習データ及びテストデータに含まれている．データは合計で 4940 件あり，聞き手の相槌が行われなかった場合のデータは全てで 2413 件存在した．そして上記

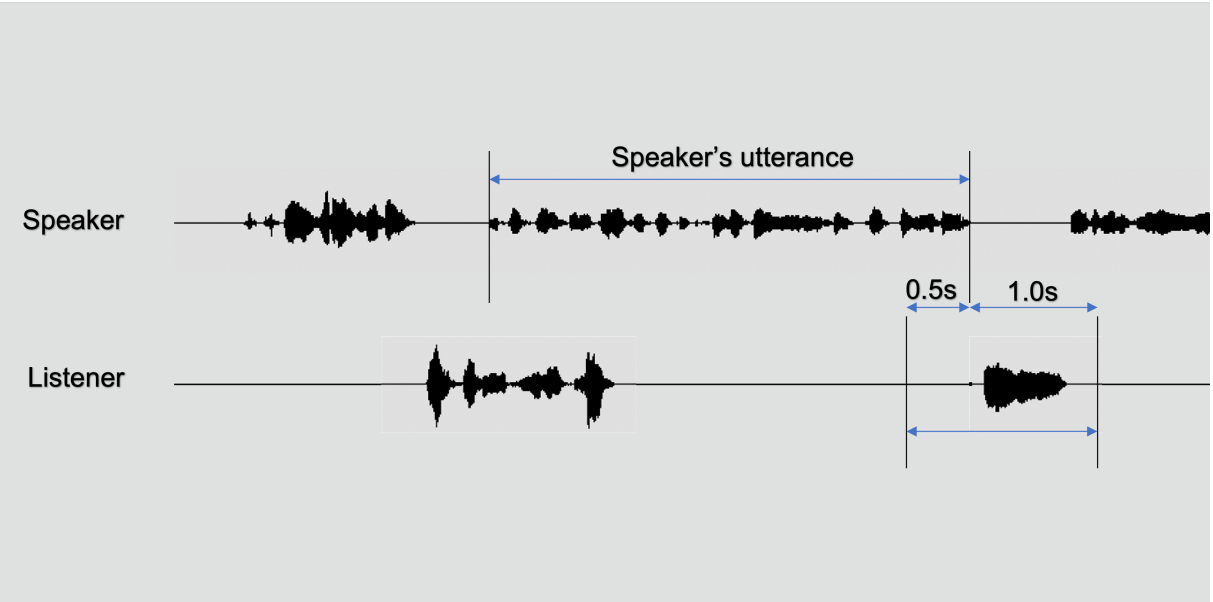

図 5.1: 目的変数の抽出区間

における説明変数から目的変数を推定する機械学習モデルを構築した（図 5.2）．本稿では下記のタスクを 100 回行い，モデルの推定性能の各指標を算出した．

(1) データセットをホールドアウト法により学習データとテストデータに 9:1 の割合で無作為に分ける．

(2) 相槌を予測する機械学習モデルを構築する．

加えて，各モダリティ（視覚・韻律・言語）がどの程度影響しているのかを調べるために，モダリティの組み合わせごとにモデルを構築した．

表 5.4: データセット中の各相槌ラベルの数

ラベル	データ数
N	1577
P	689
NP	85
E	1171
A	540
C	379
R	153
S	259
O	41

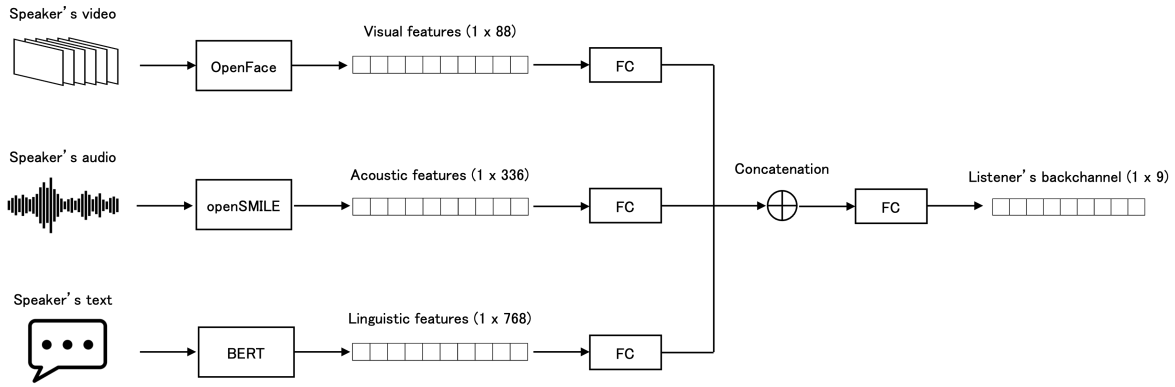

図 5.2: 相槌予測モデルの簡略図

第6章　構築した機械学習モデルの分析

本章では，構築した推定モデルの分析・考察を述べる．

6.1　相槌の推定結果について

5.2 節で構築した機械学習モデルにおける推定性能の各指標の平均値を表 6.1 に示す．ベースラインでは学習データにおけるそれぞれの相槌ラベルの出現割合によって各ラベルを出力するようにした．ベースラインにおける F 値は 0.170 であった．本稿で構築したモデルにおいて最も F 値が高かったのは Model G であった．Model G では視覚・韻律・言語のモダリティを特徴量として使用している．Model G は，ベースラインよりも推定性能が高いことが確認できた．また，単体のモダリティを特徴量として用いている Model A，Model B，Model C に比べて，複数のモダリティを用いている Model D，Model E，Model G の方が推定性能が高いことがわかった．このことから，単体のモダリティを特徴量として使用するよりも，複数のモダリティを組み合わせて使用する方が推定性能が上がる可能性が示された．

表 6.1: 各モデルの推定性能の各指標の平均値（N=100）

	視覚	韻律	言語	適合率	再現率	F 値
Baseline				0.171	0.171	0.170
Model A	✓			0.569	0.026	0.048
Model B		✓		0.342	0.120	0.172
Model C			✓	0.270	0.167	0.202
Model D	✓	✓		0.367	0.209	0.259
Model E	✓		✓	0.351	0.236	0.277
Model F		✓	✓	0.353	0.148	0.200
Model G	✓	✓	✓	0.342	0.242	**0.280**

6.2　考察

6.2.1　モデルの性能についての考察

はじめに，全てのモダリティを用いた Model G における相槌ラベル別の推定性能の各指標の平均値を表 6.2 に示す．表 6.2 においてラベル NP，R，O の性能が著しく低く，ラベル C，S の性能が低い結果となっている．これらのラベルはデータ数が少ないことから，データ数の不足が推定性能に影響している可能性が考えられる．そして，データセットにおける各ラベルのデータが不均衡であるため，データ数の少ないラベルは推定ができていない可能性も考えられる．

次に，モダリティ単体でのモデル（Model A，Model B，Model C）の性能について考察する．表 6.1 より，モダリティ単体のモデルでは Model C（言語のみ），Model B（韻律のみ），Model A（視覚のみ）の順に性能が高い結果となった．**??**章で紹介した関連研究では韻律的特徴が相槌の予測に重要であることが述べられている [7]．そのため韻律的特徴に着目して相槌の予測を行う事例が多い．今回の結果においては韻律的特徴量のみを使用したモデル（Model B）の性能はベースラインをわずかに越えたが，大幅に高い結果とはならなかった．一方で，相槌が行われる場合のみに着目した先行研究 [26] では，韻律的特徴量を用いたモデルはベースラインの性能を大幅に超える結果となった．先行研究と本稿との差分として，相槌が行われない場合を考慮したことが挙げられる．そこで，相槌が行われない場合を含めることで相槌の推定にどのくらい影響を与えるのか確認するため，本稿で構築したモデルを用いて相槌の有無を推定した．相槌の有無を推定するため，構築したモデルが相槌ラベルを一つでも行われると推定したものを相槌有りとみなした．下記の表 6.3 は各モデルの相槌の有無の推定性能を示したものである．表 6.3 から，ベースラインの性能を上回ったのは全てのモダリティを用いた Model G のみであることが確認できる．つまり，Model G 以外のその他のモデルでは相槌の有無を推定できていない可能性が考えられる．上記のことから，本稿で使用した韻律的特徴量は，相槌の有無の予測を行うための特徴をうまく捉えられていなかった可能性があることが挙げられる．このことに関する一つの原因として，本稿で使用した韻律的特徴量は統計処理を行った値であることが挙げられる．先ほども紹介した ward の研究では，話者の低音域が 110ms 続いた後に聞き手の相槌が発生する傾向があることがわかっている [7]．つまり，今回の話者のデータでは話者の発話の末尾に重要な手がかりがある可能性がある．しかし，本稿では韻律的特徴量として韻律の特徴量の統計量を使用しているため，話者の末尾にある韻律的に重要な手がかりの情報が統計処理により失われている可能性が考えられる．

次にモダリティ単体でのモデル（Model A，Model B，Model C）と複数のモダリティを用いたモデル（Model D，Model E，Model F，Model G）について比較する．複数のモダリティを用いたモデル（Model D，Model E，Model F，Model G）に関して，Model F を除き，単体のモダリティを用いたモデル（Model A，Model B，Model C）よりも推定性能が高い結果となった．このことから，相槌を予測する際には複数のモダリティを用いることで推定性能が向上することが示唆された．**??**章で紹介した関連研究では複数のモダリティを用いることで性能が向上することが述べられている [10]．今回の結果においても複数のモダリティを用いたモデル（Model D，Model E，Model G）の性能が高いことから，関連研究と一致していることが確認できた．

6.2.2　相槌の有無の予測に関する考察

先行研究 [26] では, 相槌がある場合のみに着目して相槌ラベルの予測を行った．その結果，モデルの推定性能（F 値）が 0.456 と高い性能を示している．しかし，本稿で構築したモデルの推定性能（F 値）は 0.280 であり，先行研究よりも大幅に下回っている．このことに関して，相槌が行われる場合と行われない場合のデータ数のばらつきや時系列情報

を考慮していないことが大きく影響している可能性が考えられる．本稿で使用したデータセットは，相槌が行われたデータは 2527 件あり，相槌が行われなかったデータは 2413 件であった．相槌ラベルの中で最も多いデータ数は 1577 件であるため，相槌が行われなかったデータに比べて数が足りていないのは明らかである．加えて，関連研究では主に時系列を考慮したモデルを構築している [3]．しかし，本稿で構築したモデルでは時系列が考慮できていない．そのため，性能が大きく向上しなかったことが考えれる．

表 6.2: Model G における相槌ラベル別の推定性能の各指標の平均値（N=100）

	データ数	適合率	再現率	F 値
N	1577	0.377	0.304	0.335
P	689	0.378	0.283	0.322
NP	85	0.182	0.067	0.095
E	1171	0.389	0.283	0.326
A	540	0.394	0.270	0.318
C	379	0.270	0.166	0.204
R	153	0.178	0.064	0.088
S	259	0.194	0.091	0.121
O	41	0.130	0.043	0.061

表 6.3: 各モデルの相槌の有無の推定性能の各指標の平均値（N=100）

	視覚	韻律	言語	適合率	再現率	F 値
Baseline				0.676	0.507	0.579
Model A	✓			0.058	0.802	0.108
Model B		✓		0.311	0.690	0.428
Model C			✓	0.466	0.613	0.529
Model D	✓	✓		0.459	0.693	0.551
Model E	✓		✓	0.505	0.670	0.575
Model F		✓	✓	0.371	0.684	0.480
Model G	✓	✓	✓	0.536	0.662	0.592

第7章　結論

本稿では言語的・機能的に適切な相槌を生成するための初期検討を行った．初期検討として，対話中の話し手のマルチモーダル情報から言語的・機能的に適切な相槌の種類を予測する試みを行った．具体的には話し手の視覚的，韻律的，言語的特徴量を用いて，相槌がない場合を含めた聞き手の言語的・機能的に適切な相槌の種類を予測する機械学習モデルを構築した．その結果，本稿で提案したモデルはベースラインモデルを上回る性能であることが確認された．さらに，モダリティ単体のモデルでは，言語，韻律，視覚の順で性能が高いことが確認された．また，特徴量としてモダリティを複数用いることで推定性能が向上することが示唆された．

本稿では上記の結果が得られたが，いくつかの制約がある．1 つ目に，推定性能が低かった相槌ラベルのデータに関してはデータ数が足りていないこと，データセット中の各相槌ラベルのデータに関して不均衡であったこと，データセット中の各相槌ラベルのデータと相槌が行われなかった場合のデータが不均衡であったことが性能の向上を妨げたのではないかという可能性がある．今後の課題として，データ数を増やすことやデータの不均衡をなくすことが挙げられる．2 つ目に，本稿で使用したモデルに関して時間情報を考慮していないことが挙げられる．対話における分析では話の流れなどを理解する必要があり，時間情報が重要なのは明らかである．そのため，モデルの構築方法に関して時間情報を考慮した LSTM[27] などを用いることで性能の向上を図る予定である．

最後に，本稿では対話中における聞き手の適切な相槌の種類を予測できるかについて焦点を当てたが，自然な相槌を打つ対話エージェントを実現するためには適切な相槌の種類を予測するだけでなく，予測された相槌の種類からエージェントの具体的な言語的・非言語的な行動を生成する必要がある．今後さらに研究を進めて，自然な相槌を打つ対話エージェントを実現できるかどうか検討していきたい．

参考文献

[1] Shinya Fujie, Kenta Fukushima, and Tetsunori Kobayashi. Back-channel feedback generation using linguistic and nonlinguistic information and its application to spoken dialogue system. In *9th European Conference on Speech Communication and Technology (INTERSPEECH '05)*, pp. 889–892, 2005.

[2] Kohei Hara, Koji Inoue, Katsuya Takanashi, and Tatsuya Kawahara. Prediction of turn-taking using multitask learning with prediction of backchannels and fillers. In *Proc. 19th Annual Conference of the International Speech Communication Association (INTERSPEECH '18)*, pp. 991–995, 2018.

[3] Ryo Ishii, Xutong Ren, Michal Muszynski, and Louis-Philippe Morency. Multimodal and multitask approach to listener ' s backchannel prediction: Can prediction of turn-changing and turn-management willingness improve backchannel modeling? In *Proc. 21st ACM International Conference on Intelligent Virtual Agents (IVA '21)*, pp. 131–138, 2021.

[4] Louis-Philippe Morency, Iwan De Kok, and Jonathan Gratch. Predicting listener backchannels: A probabilistic multimodal approach. In *International Workshop on Intelligent Virtual Agents (IVA '08)*, pp. 176–190, 2008.

[5] Markus Mueller, David Leuschner, Lars Briem, Maria Schmidt, Kevin Kilgour, Sebastian Stueker, and Alex Waibel. Using neural networks for data-driven backchannel prediction: A survey on input features and training techniques. In *International Conference on Human-Computer Interaction (HCI '15)*, pp. 329–340, 2015.

[6] Khiet P. Truong, Ronald Poppe, and Dirk Heylen. Arule-based backchannel prediction model using pitch and pause information. In *11th Annual Conference of the International Speech Communication Association (INTERSPEECH '10)*, pp. 3058–3061, 2010.

[7] Nigel Ward. Using prosodic clues to decide when to produce back-channel utterances. In *Proc. 4th International Conference on Spoken Language Processing (ICSLP '96)*, Vol. 3, pp. 1728–1731, 1996.

[8] Nigel Ward and Wataru Tsukahara. Prosodic features which cue back-channel responses in english and japanese. *Journal of Pragmatics*, Vol. 32, No. 8, pp. 1177–1207, 2000.

[9] Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch. Parasocial consensus sampling: Combining multiple perspectives to learn virtual human behavior. In *Proc. 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS '10)*, pp. 1265–1272, 2010.

[10] Robin Ruede, Markus Müller, Sebastian Stüker, and Alex Waibel. Yeah, right, uh-huh: A deep learning backchannel predictor. *Advanced Social Interaction with Agents*, Vol. 510, pp. 247–258, 2019.

[11] Senko K. Maynard. On back-channel behavior in japanese and english casual conversation. *Linguistics*, Vol. 24, No. 6, pp. 1079–1108, 1986.

[12] Chiharu Mukai. The use of back-channels by advanced learners of japanese: Its qualitative and quantitative aspects. *Japanese language education around the globe*, Vol. 9, pp. 197–219, 1999.

[13] Akira Morikawa, Ryo Ishii, Hajime Noto, Atsushi Fukayama, and Takao Nakamura. Determining most suitable listener backchannel type for speaker's utterance. In *Proc. 22nd ACM International Conference on Intelligent Virtual Agents (IVA '22)*, pp. 1–3, 2022.

[14] Tatsuya Kawahara. Spoken dialogue system for a human-like conversational robot erica. In *Proc. 10th International Workshop on Spoken Dialogue Systems (IWSDS '19)*, pp. 65–75, 2019.

[15] Soumia Dermouche and Catherine Pelachaud. Generative model of agent ' s behaviors in human-agent interaction. In *2019 International Conference on Multimodal Interaction*, pp. 375–384, 2019.

[16] Hiroki Mori and Shunya Kimura. A Generative Framework for Conversational Laughter: Its 'Language Model' and Laughter Sound Synthesis. In *Proc. INTERSPEECH 2023*, pp. 3372–3376, 2023.

[17] Divesh Lala, Koji Inoue, Tatsuya Kawahara, and Kei Sawada. Backchannel generation model for a third party listener agent. Proc. 10th International Conference on Human-Agent Interaction (HAI ' 22), pp. 114–122, 2022.

[18] Ryo Ishii, Ryuichiro Higashinaka, and Junji Tomita. Predicting nods by using dialogue acts in dialogue. In *Proc. 11th International Conference on Language Resources and Evaluation (LREC '18)*, pp. 2940–2944, 2018.

[19] 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.

[20] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *13th IEEE international conference on automatic face and gesture recognition (FG '18)*, pp. 59–66, 2018.

[21] P. Ekman and W.V. Friesen. Manual for the facial action coding system. *Palo Alto: Consulting Psychologists Press*, 1977.

[22] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proc. 21st ACM international conference on Multimedia (MM '13)*, pp. 835–838, 2013.

[23] Björn Schuller, Stefan Steidl, and Anton Batliner. The interspeech 2009 emotion challenge. 2009.

[24] 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.

[25] J. Bergstra, D. Yamins, and D.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.

[26] 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), 第 2023-GN-119 巻, pp. 1–6, 2023.

[27] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

研究業績

査読付き国際会議

(1) 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 (IVA2023) (2023 年 9 月).

(2) 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 (IVA2023) (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, 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 (DICOMO2023), pp.352–358 (2023 年 7 月).

(2) 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).

(3) 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).

(4) 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).

受賞

(1) マルチメディア、分散、協調とモバイル (DICOMO2023) 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 7 月).