聞き手の表情分析に基づいた相槌提示システムの検討

令和６年度　卒業論文

日本大学　文理学部　情報科学科　宮田研究室

鹿摩　大智

概要

対話において聞き手が相槌を打つことは，対話を円滑に進めるために重要な要素の一つである．適切な相槌を打つことでコミュニケーションを円滑に進めることができ，良好な人間関係の構築に役立つことが期待されるが，人とのコミュニケーションが苦手な人は対話の流れに合う適切な相槌を打てているのか自身で把握し，改善することは難しい．そこで，実際の対話データを用いて，相槌の機能に基づいた適切な相槌の言語・非言語行動を明らかにする．本稿では，相槌の機能に基づいた適切な相槌の言語・非言語行動を明らかにする取り組みとして，言語・非言語行動の中から顔部の振舞いに着目して分析を行った．分析結果から，感情の動きを表すような相槌は目の周りの動きと唇の動きが活発になることや，話し手の話題を先取りするような応答の際には顎に動きがあること，感情を含まない相槌を打つ際には唇の動きが活発になるという特徴が明らかになった．さらに話し手の言語・非言語行動から聞き手が打つ相槌の機能を予測する機械学習モデルを用いて，生成した表情画像を提示するシステムの検討を行った．

目 次

図 目 次

表 目 次

第1章　序論

1.1 研究の背景

対話において聞き手が相槌を打つことは，対話を円滑に進めるための重要な要素の一つである．日本語における相槌を表す言葉は豊富であり [1]，相槌は肯定的な応答や発言を繰り返す応答などの機能的な側面に分類することができる [2]．さらに，相槌の機能的な側面に着目して，話し手の発話意図や種類と聞き手の相槌の関係を分類することができることから，聞き手の相槌は話し手の対話行為と関連があると考えられている [3]．適切な相槌を打つことでコミュニケーションを円滑に進めることができ，良好な人間関係の構築に役立つことが期待される．しかし，人とのコミュニケーションが苦手な人は，対話の流れに合う適切な相槌を打てているのか自身で把握し，改善することは難しい．相槌に関する研究事例は一定数存在するが，多くは相槌を表現する言葉や相槌に含まれる機能や意味を考慮する研究は少ない．すなわち，相槌の機能に基づいて適切な相槌を打つために言語・非言語行動をどのように用いると良いかは明らかにされていない．

1.2 研究の目的

1.1 節で述べた研究の背景より，相槌の機能に基づいて適切な相槌を打つために言語・非言語行動をどのように用いると良いか明らかにする必要がある．そこで本研究では，対話の流れに合う適切な相槌を打つ際には，どのような言語・非言語行動を用いるのかを明らかにする．具体的には，対話における相槌の機能と人間の振舞いの関係を明らかにするために，話者（話し手，聞き手）の振舞いに着目し，相槌の機能に基づいた言語・非言語行動の特徴を分析する．これまで，対話中の話し手の言語・非言語行動から聞き手の相槌の機能を予測する取り組みを行ってきた [4–7]．これらの研究では，聞き手の相槌の機能を予測する取り組みにとどまっており，適切な相槌を打つ際の言語・非言語行動を明らかにする分析や，適切な相槌の振舞いをフィードバックとして提示する取り組みは行っていない．そこで本稿では，相槌の機能に基づいて聞き手が相槌を打つ際の表情を分析する．さらにその分析結果を踏まえ，適切な相槌を提示するシステムを実現するために，話し手の言語・非言語行動から聞き手が打つべき相槌の機能を予測し，予測結果に基づいた表情画像を提示するシステムの検討を行う．初期検討として我々は，肯定的な応答，否定的もしくは悩んでいる応答，感情を含まない（ニュートラルな）応答を行う際の表情を提示するシステムを提案した [8]．上記のシステムでは，既存研究 [3] で提案されている相槌の機能を扱うことや，話し手の言語・非言語行動から聞き手が打つべき相槌の機能を予測し，フィードバックを提示することは行なっていない．本稿では，提案したシステム [8] を発展させるために，既存研究 [3] で提案されている相槌の機能を取り入れる．さらに，ユーザに対して適切な相槌を打つ際の表情をフィードバックとして提示するようなシステムの検討を行う．

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，対話中の振舞いから相槌を予測・生成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，本論文で扱う対話コーパスについて述べる．
5 章では，対話コーパスにおける適切な相槌を打つ際の表情の分析を行った．
6 章では，分析結果を踏まえ，相槌の機能に基づいた適切な相槌の表情を提示するシステムの検討について述べる．
最後に 7 章にて，本論文の結論を述べる．

第2章　関連研究

本章では，対話中の振舞いから相槌を予測・生成する研究事例とマルチモダール情報に基づいたスキル推定とスキルトレーニングに関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，特定のタスクや対話中の行動や能力を分析するという点で本研究と関係している．2.1 節では，相槌の機能に関する研究事例について紹介する．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．2.3 節では，相槌の生成に関する研究事例について紹介する．2.4 節では，マルチモーダル情報を用いたスキル推定とスキルトレーニングに関する研究事例について紹介する．

2.1 相槌の機能に関する研究事例

相槌の機能に関する研究事例として [1,3,9] が挙げられる．

Maynard [1] は 12 組の日米語の日常会話が録画されたデータを基に談話上の相槌の機能を挙げ，日本語と英語の類似点や相違点を指摘している．相槌の機能として，話を続けてほしいことを示す表現，内容の理解を示す表現，話し手への共感を示す表現，感情を示す表現を挙げている．相違点については相槌の頻度を挙げており，日本語が英語よりも著しく相槌の回数が多いことを指摘している．森ら [9] は相槌の機能として，受容や承認を表す応答系感動詞，気づき，驚き，感心を表す表出系感動詞，理解や同意を表す語彙系応答を挙げている．

2.2 相槌の予測に関する研究事例

相槌の予測に関する研究事例として [9–16] が挙げられる．

Ward [10] はルールベースで相槌の機会を予測を行う機械学習モデルを構築し，日本語対話において話し手の低音域が相槌の予測で重要な要素になることを示唆している．

Morency ら [11] は確率を用いて相槌の機会を予測する機械学習モデルを構築している．彼らは話者の言語，韻律，視覚的特徴量を利用して，聞き手の相槌のタイミングを予測する取り組みを行っている．彼らは，視覚的な相槌の予測において，ルールベースを用いた従来の手法よりも統計的に有意に改善したことを示している．

森ら [9] は頭部運動である頷きが相槌と共起することが多いという特徴を挙げ，3 人会話データから韻律，言語的な情報に加えて参加者の視線情報から相槌と頷きが共起するか予測するモデルを構築している．加えて，共起すると予測された場合には，頷きのタイプや移動範囲などの物理的特徴を予測するモデルを構築し，多人数会話においても有効な予測ができることを示唆している．

Blache ら [12] らは，従来の手法では聞き手の関心を示すような相槌と発話内容に反応するような相槌を異なる処理で予測する必要があったが，彼らは 1 つのループ処理で予測することによってより自然で効率的な相槌の予測と生成する手法を提案している．彼らは，このアーキテクチャが感情や対話の微細なニュアンスに基づいてエージェントの反応

する能力を向上させることで，人間らしさを向上させることができる可能性があることを示唆している．

Muller ら [13] は，視覚的，韻律的特徴量を用いてグループディスカッションにおける相槌の検出を行うモデルの構築をしている．その結果，頭部の動きと身体姿勢の視覚的特徴量を用いたモデルが最も高く，次に頭部の動きのみを用いたモデルが高い精度を達成している．このことから彼らは，相槌と頷きには強い関連性があることを示唆している．

Jain [14] らは視覚的，韻律的特徴量を用いて聞き手の相槌の機会の予測を半教師あり学習手法を提案している．彼らの提案した手法では，少量のラベル付きデータを用いて自動的に相槌を認識し，相槌の機会を予測する機械学習モデルを構築している．この機械学習モデルは，全て手動でラベル付けを行う従来の手法と比較して 95%の精度を達成している．さらに彼らは，個人の性格が相槌の種類に与える影響を調査している．その結果から，エージェントが自然な対話を行うためには個々の性格に基づいた相槌の生成が重要であることを示している．

Lala ら [15] は韻律的特徴量を用いて応答的感動詞，表現的感動詞，笑いの 3 つに分類された相槌の種類と相槌が打たれるタイミングを予測するモデルをトレーニングし，相槌の生成を行うモデルを構築している．彼らは，他のベースラインモデルと比較を行うための実験を行い，構築したモデルが優れていることを示している．しかし，応答や笑いの予測は行うことができているものの，感情を表現するような応答の予測の性能が不十分であり，パフォーマンスの低下を引き起こしていることを示している．さらに彼らは，言語的特徴量を含めることでモデルの性能が向上することを示唆している．

Dermouche ら [16] はユーザの視覚的特徴量として笑顔の強度，頭部の動き，視線の向きを用いて対話エージェントの応答として同じく笑顔の強度等の表情的特徴量を予測するモデルを構築している．彼らは LSTM ネットワークを用いることで時間の経過による行動の変化を考慮することのできる予測モデルの構築を行っている．

2.3 相槌の生成に関する研究事例

相槌の生成に関する研究事例として [16–19] が挙げられる．

Haddad ら [19] らは，自然な視聴覚情報の生成のために，笑顔と笑いの生成手法について提案している．彼らは，1 名の被験者が録画された映像を閲覧するという内容のデータの中から，笑いが起こった部分にアノテーションを行っている．笑顔については強弱の 2 種類をラベル付けしている．笑いについては高・中・低の 3 つの覚醒の度合いに加えて，覚醒の度合いの高と低には強弱という 5 種のラベル付けを行っている．彼らはまず，笑顔と笑いを生成するためのシステムの構築を行っている．エージェントの表情の生成については，元の映像データに対してデータの複製と切り取りを繰り返す作業と，フレーム間に対して線形補完を適用することによって，滑らかで自然な表情の生成を行う手法を提案している．音声の生成については，ラベルを考慮してデータ中から適切な音声データをサンプルとして扱っている．このシステムの評価実験については，元データと生成されたデータの間に有意差が無かったため，自然な生成が肯定されている．彼らは次に実際の対

話データから笑顔と笑いを予測し，生成するシステムの構築を行っている．視覚的・韻律的特徴量から条件付き確率場のモデルを構築している．評価実験では比較のために,「生成されたエージェント」,「元のデータに基づいたエージェント」,「笑顔のみのエージェント」,「中立的なエージェント」の 4 つのクラスに分類している．結果として，生成されたエージェントは他のクラスのエージェントよりも理解しているように認識され，元のデータに基づいたエージェントよりも自然であると認識されている．彼らは，生成されたエージェントの笑顔と笑いの数値は元のデータに基づいたエージェントよりもばらつきが大きかったため，自然であると認識されたことを示唆している．

Dermouche ら [16] は韻律的特徴量としてユーザの笑顔の強度，頭部の動き，視線の向きを用いて，視覚的特徴量を予測する機械学習モデルの構築と予測された視覚的特徴量を適用したエージェントを生成する対話システムの実装と評価を行っている．実験は，エージェントがガイドの役割を果たし，博物館の訪問者にビデオゲームに関する展示を紹介するというシナリオで評価を行っている．彼らはやりとりの満足度に関して，エージェントがユーザの動作に対して笑顔を適応させた場合のみに有意であったことを示している．視線の向きと頭部の動きについては，インタラクション全体を通じてお互いを見つめ合うシナリオの適性上，あまり動きがなかった．

Jonell ら [17] は視覚的，韻律的特徴量を用いて対話におけるエージェントの表情と頭部の動きを生成するモデルを構築している．構築した機械学習モデルを基に構築したシステムを用いた評価実験から，話し手を考慮しない不適合なジェスチャーよりも，話し手に応答したジェスチャーの方が有意に好まれることが分かった．このことから，彼らのアプローチがよりマルチモーダル情報を十分に活用できることを示している．

Bucci ら [18] は言語的特徴量を用いることによって予測された文章の感情から，予測された感情に応じた表情を生成する 2 つのモジュールで構成されたアーキテクチャを提案している．1 つ目のモジュールは，与えられた文章から感情値と覚醒度を推定する機械学習モデルで構成されている．2 つ目のモジュールは，推定された 2 つの値から感情を表現するために必要な顔の筋肉の動きを数値として推定する機械学習モデルで構築されている．提案したアプローチについて，最先端のアプローチと比較して高い性能を有することを示している．

2.4 マルチモーダル情報に基づいたスキル推定とスキルトレーニングに関する研究事例

マルチモーダル情報を用いたスキル推定とスキルトレーニングに関する研究事例として [20, 21] が挙げられる．

Ohba ら [20] は視覚的，韻律的，言語的特徴量，さらに生理学的特徴量を用いて，面接訓練システムのためのユーザのスキル推定を行う機械学習モデルの構築と，ユーザのスキルと自己評価の関係の調査を行っている．彼らが構築したモデルの予測結果として，コミュニケーションスキルの推定には視覚的，韻律的，言語的特徴量を用いたモデルが最も

高い精度を達成している．さらに自己効力感の推定には生理学的特徴料を用いたモデルが最も高い精度を達成している．加えて，コミュニケーションスキルと自己効力感のギャップの推定には韻律的，言語的特徴量を用いたモデルが最も高い精度を達成している．このことから，コミュニケーションスキルの推定のためには韻律的特徴量などのような外部から観測が可能な特徴量，自己申告による自己効力感の推定のためには生理学的特徴量が重要であることを示している．さらに彼らは，推定結果を用いたフィードバックはスキル向上のためのシステムの開発に有用であることを示唆している．

Ito ら [21] は視覚的，韻律的，言語的特徴量を用いて，グループディスカッションにおける説得力の推定を行う機械学習モデルの構築と，推定結果を用いた説得力の可視化を行うビデオ会議システムの実装と，実装したシステムを用いた評価実験を行っている．その結果から，長く複雑な議論の内容を把握するために，説得力を可視化する彼らのシステムはユーザにとって利用価値が高いことを示している．彼らはさらにスキルトレーニングを行うシステムの実装のために，有用な特徴の調査を行っている．その結果から，説得力の高い話者は声が大きく，強く，高い傾向を示している．さらに 1 発話あたりの形態素数が多く発話が長いことを示している．

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

対話において聞き手が相槌を打つことは，対話を円滑に進めるための重要な要素の一つである．2.2 節で述べたように，話し手の言語・非言語情報を用いて相槌の機会やどのような相槌を打つのかを予測する研究が多く行われている．Morency ら [11] は確率を用いて相槌の機会を予測する機械学習モデルを構築した．森ら [9] は頭部運動である頷きが相槌と共起することが多いという特徴を挙げ，3 人会話データから韻律，言語な情報に加えて参加者の視線情報から相槌と頷きが共起するか予測するモデルを構築した．Dermouche ら [16] は視覚特徴量を用いて対話エージェントの応答として笑顔の強度，頭部の動き，視線の向きを予測するモデルの構築と，予測された視覚的特徴量を用いてエージェントを生成するシステムの実装，加えて実装したシステムの評価実験を行った．さらに 2.3 節で述べたように，エージェントの表情の生成に関する研究は多く行われている．Jonell ら [17] は視覚的，韻律的特徴量を用いて対話におけるエージェントの表情と頭部の動きを生成するモデルを構築した．Bucci ら [18] は言語的特徴量を用いることによって予測された文章の感情から，予測された感情に応じた表情を生成する 2 つのモジュールで構成されたアーキテクチャを提案している．2.1 節で述べたように，品詞等を用いて聞き手の相槌を機能的な側面に基づいて分類を行う研究が行われているが，このような相槌の予測やエージェントの表情の生成に関する研究事例の多くは相槌に含まれる機能や意味を考慮していなため，対話の流れや文脈に沿った相槌を予測・生成することができない可能性がある．相槌に含まれる機能や意味を考慮することで，対話の流れや文脈に沿った相槌を予測・生成できるかは明らかになっていない．

加えて，2.4 節で述べたように，マルチモーダル情報を用いたスキル推定やスキルトレーニングを行う研究は多く行われている．Ohba ら [20] は視覚的，韻律的，言語的特徴量を用いて，面接訓練システムのためのユーザのスキル推定を行う機械学習モデルの構築と，ユーザのスキルと自己評価の関係の調査を行った．Ito ら [21] は視覚的，韻律的，言語的特徴量を用いて，グループディスカッションにおける説得力の推定を行う機械学習モデルの構築と，推定結果を用いた説得力の可視化を行うビデオ会議システムを構築した．本研究が属する研究領域では，ユーザのマルチモーダル情報を基に対話において適切な相槌を打つスキルを向上させるような取り組みは行われておらず，コミュニケーションが苦手な人が自身で対話の流れに合う適切な相槌を打つことができているかを把握し，改善することができるかは明らかになっていない．

上記をふまえ，本研究における問題は，コミュニケーションが苦手な人が自身で対話の流れに合う適切な相槌を打つことができているかを把握し，改善することができるか明らかになってい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，自身で対話の流れに合う適切な相槌を打つことができているかを把握し，改善することができるのか明らかになっていないという問題があり，コミュニケーションが苦手な人はどのような相槌を打つべきであるのか分からないということ考えられる．この問題を解決するためには，対話の流れに合う適切な相槌を打つために言語・非言語行動をどのように用いると良いかを明らかにする必要がある．そこで，人間の言語・非言語行動を利用して，相槌の機能に基づいた分析を行う．これにより，対話において適切な相槌を打つためにはどのような行動が重要であるか明らかになり，適切な相槌を打つための能力の向上を目指すトレーニングシステムの開発が期待される．

本稿では，対話における相槌に含まれる機能や意味を予測・生成し，ユーザにフィードバックする取り組みの初期検討して，聞き手の言語・非言語行動から相槌の機能に着目した表情分析を行う．加えて，分析結果を踏まえた表情画像を生成し，話し手の言語・非言語行動から聞き手が打つべき相槌の機能を予測した上で，ユーザに表情画像を提示するようなシステムの検討を行う．上記をふまえ，本稿では，**聞き手の言語・非言語行動から相槌の機能に基づいた表情の分析を行うこと．さらに，話し手の言語・非言語行動から相槌に含まれる機能や意味を予測し，聞き手となるユーザに適切な相槌の表情を提示するシステムを検討すること**を研究課題として設定する．

第4章　対話コーパス

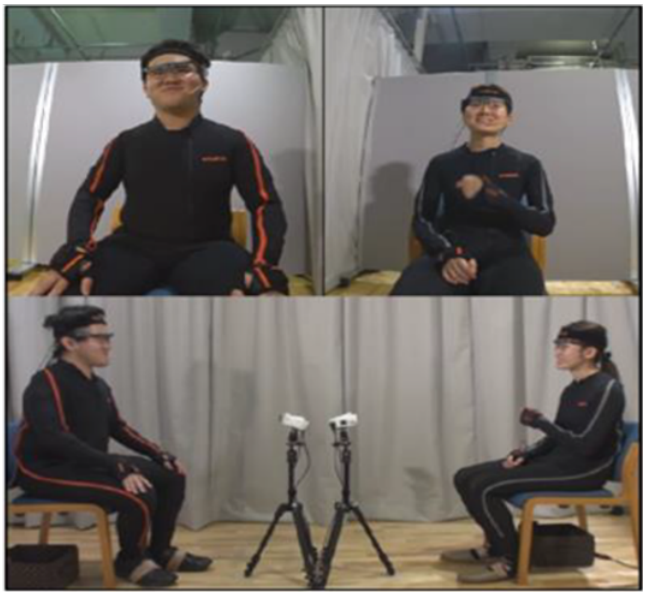

図 4.1: 2 者対話の様子

本研究では既存の 2 者対話コーパスを利用する．この対話コーパスには，2 者対話データ [22] と相槌ラベル [3] が記録されている．

4.1　2 者対話データについて

2 者対話の参加者は，初対面の日本人男女合計で 26 名（異なるペア 13 組）である．発話を含んだ相槌のデータをより多く収集するために，アニメ「トムとジェリー」を視聴した一方の参加者（話し手）が他方の参加者（聞き手）に内容を説明するタスクを行っている．発話の単位には Inter-pausal units（IPU）[23] を使用し，沈黙時間が 200ms 未満の連続した音声区間を 1 つとしている．この対話データでは，話し手が 4,940 件，聞き手が 2,865 件の合計 7,805 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 9 種の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　対話コーパスにおける相槌の機能別の表情分析

表 5.1: Action Units の詳細

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，聞き手の相槌の機能と顔部の振舞いの関係を明らかにするための分析について述べる．本研究では，聞き手の相槌の機能別の表情の分析を行うために，4.1 節の対話データから話し手の視覚的，韻律的，言語的な特徴を発話ごとに抽出した．抽出する特徴量は先行研究 [4–7] と同様であり，詳細については 5.1 節で述べる．

5.1 特徴量について

5.1.1 視覚的特徴量

顔画像処理ツールである OpenFace [24] を用いて話し手の映像データから頭部，視線，Action Units [25] に関する特徴量を抽出した．頭部に関する特徴量として，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，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，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

話し手の音声データから音声情報処理ツールである openSMILE [26] 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いて，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 [27] 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: 音声特徴量の詳細

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: 算出した統計量

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

話し手の発話内容から，自然言語処理モデルである BERT [28] 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，話し手の発話内容を 768 次元のベクトルに変換したものを言語的特徴量とした．

5.2　相槌の機能に基づいた表情分析

4.1 節の 2 者対話データの中で分析対象となるデータについて，信頼度を向上させるために本稿ではアノテータ間で相槌ラベルの一致度が高いデータを対象とした．具体的には 3 名いるアノテータのうち，2 名以上のアノテータが同一の相槌ラベルを付与しているものを対象とした．さらに，相槌ラベルが付与された発話の中で，全体の 5%以下であった相槌ラベル NP，C，R，S に関しては相槌ラベル O と統合し，N，P，E，A，O の 5 種類の相槌ラベルを扱った．

4.1 節で記録した聞き手の映像データから OpenFace を用いて，発話中の Action Units の強度を相槌ラベル別に抽出した．外れ値の影響をなくすため，平均値から標準偏差の 3 倍以上離れた値は今回の分析の対象外とした．上記の集計した値から相槌ラベル別に各

Action Units の統計量（50 パーセンタイル値，75 パーセンタイル値，最大値）を算出した．付録 A.1 は各 AU の相槌の機能別の箱ひげ図を示す．

本稿では，顔部の部位ごとの動きを分析するために，相槌ラベル別の Action Units の強度の統計量の分布を確認した．ラベル E については，AU06，AU07，AU10，AU12 において 50 パーセンタイル値，75 パーセンタイル値，最大値の全てが他の相槌ラベルと比較して高かった．さらに AU09 と AU25 においても 75 パーセンタイル値と最大値が高かった．このことから感情の動きを表すような相槌を打つ際には，目の周りと唇の動きが活発になると考えられる．次に，ラベル A については AU17 の 50 パーセンタイル値，75 パーセンタイル値，最大値の全てが他の相槌ラベルより高くなっているため，話し手の話題を先取りするような相槌を打つ際には顎に動きがあると考えられる．ラベル N については AU15，AU23 の値が最大値において他のラベルと比較して高く，唇の動きが活発であると考えられる．

第6章　相槌提示システム

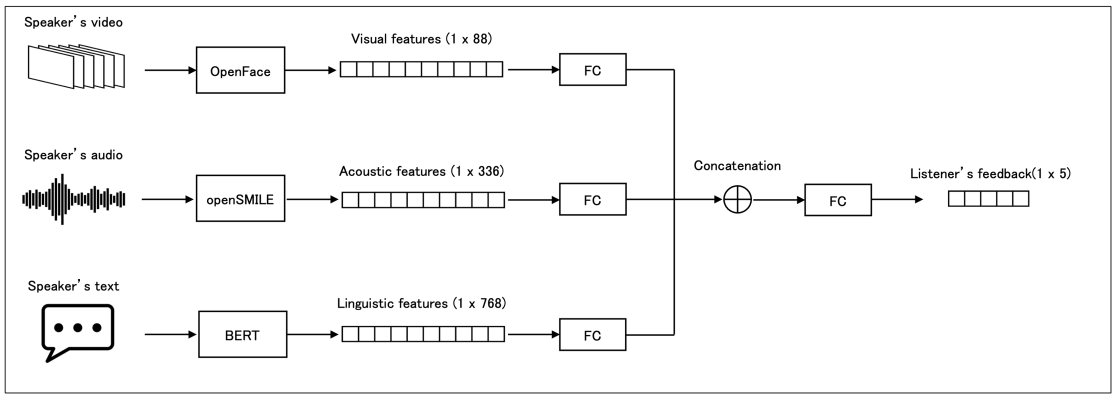

図 6.1: 本システムで扱う機械学習モデルのネットワーク図

本章では，5 章で述べた分析結果を基に，3.2 節で述べた研究課題を達成するためのシステムの概要について述べる．

6.1　モデル構築

先行研究 [4–7] で構築した相槌ラベルを推定する機械学習モデルを参考に，4 章で説明した対話コーパスから 5.1 節で説明した特徴量を抽出し，機械学習モデルを構築した．視覚的，韻律的，言語的モダリティを全て用いた機械学習モデルの推定性能は，F 値が 0.462 であった．

6.2　システム

提案システムのフローチャートを図 6.2 に示す．提案システムは入力部，特徴量抽出部，推定部，出力部からなる．

6.2.1　入力部

入力部では，話し手が発話を行っている映像データを受け取る．受け取った映像データから音声を抽出する．映像データは，話し手がカメラの正面を向いて発話を行っているシーンを撮影したものである．音声データは，話し手の音声のみが収録されたものである．受け取った映像・音声データを特徴量抽出部へと渡す．

6.2.2　特徴量抽出部

特徴量抽出部では，6.2.1 節で入力された映像と音声データから 5.1 節で述べた特徴量の抽出を行う．抽出した特徴量を推定部へと渡す．

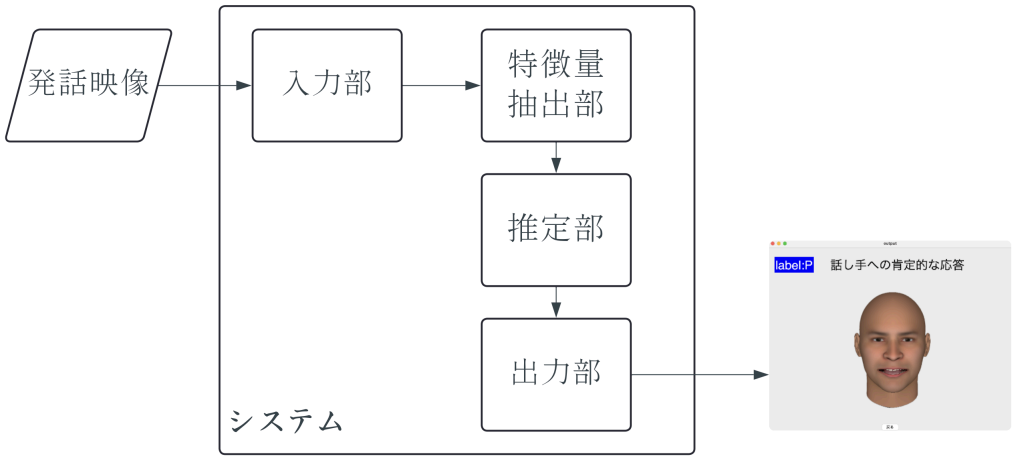

図 6.2: 提案システムのフローチャート

6.2.3　推定部

推定部では，6.1 節で構築した機械学習モデルを用いて，特徴量抽出部で抽出した特徴量から 4.2 節で述べた相槌ラベルの推定を行う．推定された相槌ラベルを出力部へと渡す．

6.2.4　出力部

出力部では推定部で予測された相槌ラベルに基づいて，自然な相槌の表情として顔画像をユーザに提示をする．システムの出力画面で提示する顔画像の生成手法について述べる．相槌の表情を生成するために，Action Units の値に基づいた人間の顔の 3D グラフィックモデルを作成することが可能な FaceGen [29] を用いた．5.2 節の分析結果を用いて，相槌ラベル別に各 Action Units の統計量（50 パーセンタイル値，75 パーセンタイル値，最大値）を算出した．相槌ラベル別に生成した顔画像を図 6.3，図 6.4，図 6.5 に示す．図 6.3 は 50 パーセンタイル値，図 6.4 は 75 パーセンタイル値，図 6.5 は最大値を各部位に反映して生成した顔画像である．

システムの最終的な出力画面について，推定されたラベルが P であり，Action Units の最大値を用いた出力例を図 6.6 に示す．画面上部にはラベルとその機能，画面下部にはラベル別の表情画像を表示した．ユーザはフィードバックとして提示された表情画像を受け取ることで，適切な相槌を打つ際の表情を確認することができる．

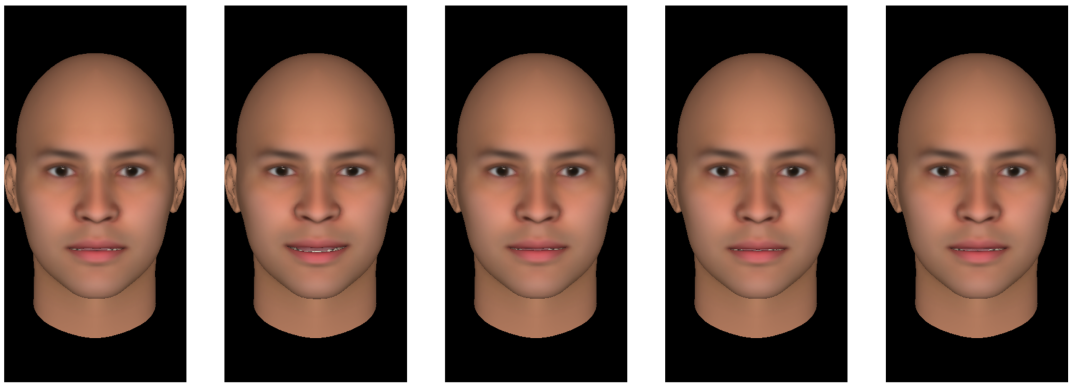

図 6.3: 50 パーセンタイル値を用いて作成した顔画像

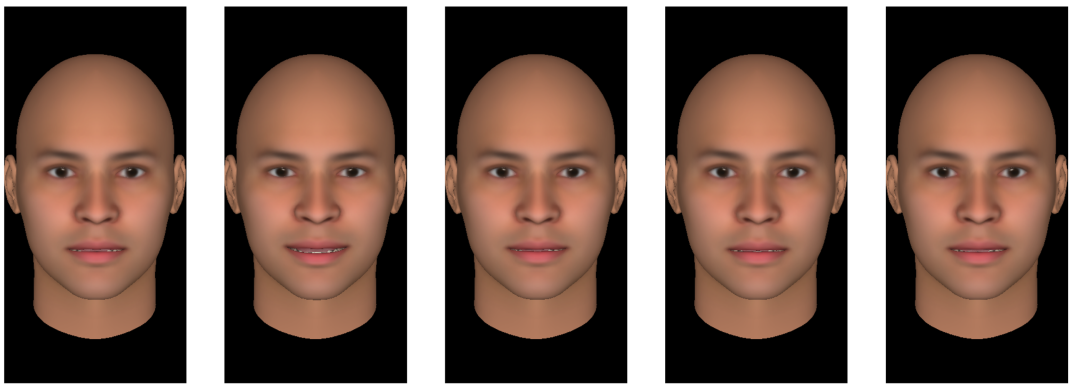

図 6.4: 75 パーセンタイル値を用いて作成した顔画像

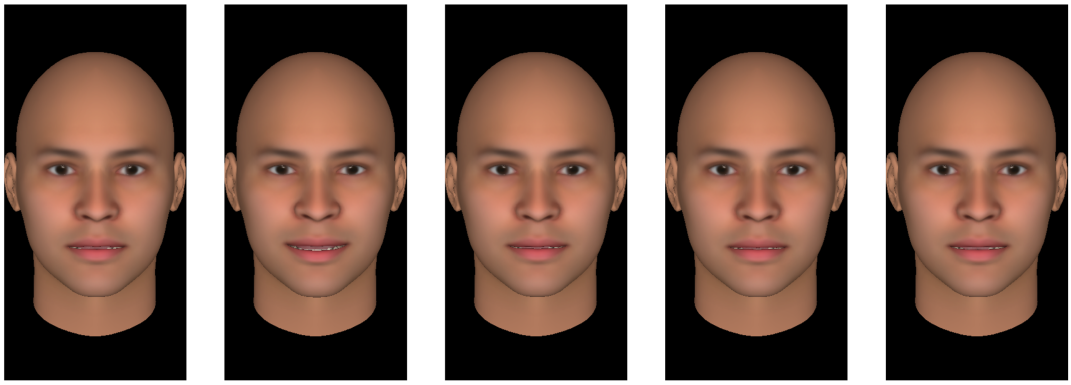

図 6.5: 最大値を用いて作成した顔画像

図 6.6: システムの出力例

第7章　結論

本稿では，適切な相槌を提示するシステムの実現を目指し，対話の流れに合う適切な相槌を打つために言語・非言語行動をどのように用いると良いかを明らかにするための分析を行った．具体的には，Action Units の値を用いて相槌の機能別に統計量の分布を確認することで，聞き手の相槌の機能に基づいた顔部の振舞いを分析した．その結果，感情の動き表すような相槌を打つ際には目の周りと口の動きが活発になることや，感情を含まない相槌を打つ際には他の相槌機能と比較して唇に動きがあることが確認できた．

さらに，ユーザに対して適切な相槌の表情をフィードバックとして提示するようなシステムの検討を行った．具体的には，話し手の映像データと音声データを入力すると，聞き手の相槌の機能を予測した上で適切な相槌の表情画像を出力するシステムの検討を行った．

本研究にはいくつかの制約がある．1 つ目に，本研究では話者の言語・非言語行動の中から表情に着目している．そのため，どのような言語，韻律，ジェスチャー等の言語・非言語行動が適切な相槌の振舞いであるか明らかになっていない．今後，本稿で扱った表情と合わせて，適切な相槌の振舞いを明らかにする取り組みを行いたい．2 つ目は，生成した 3DCG モデルは全て FaceGen に搭載されている標準の男性の顔を利用している．今後，ユーザに対してフィードバックを提示する際に，性別や国籍によってどのような影響があるかを調査する取り組みを行いたい．

今後の展望として，6 章で述べたようなシステムの構築と，顔部の分析を拡充するために相槌の機能別に寄与する特徴量を特定した上で表情画像の生成を行った上で実際にフィードバックシステムの効果を検証する予定である．さらに，現状では表情のみに着目したフィードバックにとどまっているが，韻律や言語に関するフィードバックについても追加を予定している．

参考文献

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

[2] Yasuharu Den, Nao Yoshida, Katsuya Takanashi, and Hanae Koiso. Annotation of japanese response tokens and preliminary analysis on their distribution in three-party conversations. In *2011 International Conference on Speech Database and Assessments (Oriental COCOSDA)*, pp. 168–173, 2011.

[3] 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.

[4] 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.

[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), 第 2023 巻, pp. 352–358, 2023.

[6] Toshiki Onishi, Naoki Azuma, Shunichi Kinoshita, Ryo Ishii, Atsushi Fukayama, Takao Nakao, and Akihiro Miyata. Prediction of various backchannel utterances based on multimodal information. In *Proc. the 23rd ACM International Conference on Intelligent Virtual Agents(IVA '23)*, Vol. 47, pp. 1–4, 2023.

[7] 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),, 第 2024-CN-121 巻, pp. 1–7, 2024.

[8] 鹿摩大智, 岡哲平, 大西俊輝, 東直輝, 石井亮, 深山篤, 宮田章裕. 聞き手の相槌種類に応じた表情生成システムの基礎検討. 情報処理学会インタラクション 2024 論文集, pp. 658–661, 3 2024.

[9] 森大河, 伝康晴. 相槌の特徴に一致した頷き生成モデル. 人工知能学会論文誌, Vol. 37, No. 3, pp. IDS–H_1, 052022.

[10] 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.

[11] 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.

[12] Philippe Blache, Massina Abderrahmane, Stéphane Rauzy, and Roxane Bertrand. An integrated model for predicting backchannel feedbacks. IVA '20, New York, NY, USA, 2020. Association for Computing Machinery.

[13] Haoyang Su, Wenzhe Du, Xiaoliang Wang, and Cam-Tu Nguyen. Sample efficiency matters: Training multimodal conversational recommendation systems in a small data setting. In *Proc. of the 32nd ACM International Conference on Multimedia(MM'24)*, MM '24, p. 2223–2232, 2024.

[14] Vidit Jain, Maitree Leekha, Rajiv Ratn Shah, and Jainendra Shukla. Exploring semi-supervised learning for predicting listener backchannels. In *Proc. 2021 CHI Conference on Human Factors in Computing Systems*, Vol. 395, pp. 1–12, 2021.

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

[16] Soumia Dermouche and Catherine Pelachaud. Generative model of agent's behaviors in human-agent interaction. In *2019 International Conference on Multimodal Interaction (ICMI '19)*, pp. 375–384, 2019.

[17] Patrik Jonell, Taras Kucherenko, Gustav Eje Henter, and Jonas Beskow. Let's face it: Probabilistic multi-modal interlocutor-aware generation of facial gestures in dyadic settings. Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA '20), pp. 1–8, 2020.

[18] Benedetta Bucci, Alessandra Rossi, and Silvia Rossi. Action unit generation through dimensional emotion recognition from text. In *31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 1071–1076, 2022.

[19] Kevin El Haddad, Hüseyin Çakmak, Emer Gilmartin, Stéphane Dupont, and Thierry Dutoit. Towards a listening agent: a system generating audiovisual laughs and smiles to show interest. In *Proc. 18th ACM International Conference on Multimodal Interaction (ICMI'16)*, p. 248–255, 2016.

[20] 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.

[21] Atsushi Ito, Yukiko I. Nakano, Fumio Nihei, Tatsuya Sakato, Ryo Ishii, Atsushi Fukayama, and Takao Nakamura. Estimating and visualizing persuasiveness of participants in group discussions. *Journal of Information Processing*, Vol. 31, pp. 34–44, 2023.

[22] 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.

[23] 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.

[24] 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.

[25] Paul Ekman and Wallace V. Friesen. Manual for the facial action coding system. *Palo Alto: Consulting Psychologists Press*, 1977.

[26] 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.

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

[28] 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.

[29] Singular Inversions. Facegen modeller. `https://facegen.com/modeller.htm`. (accessed 2023/12/6).

付録

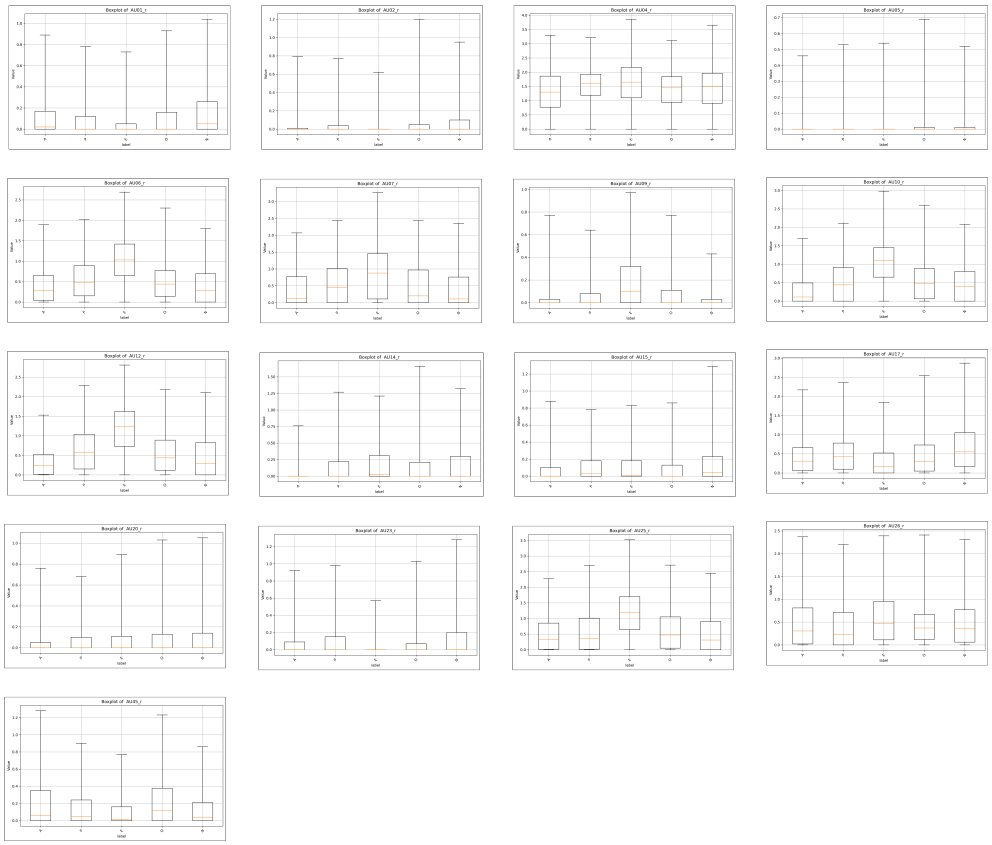

図 A.1: 相槌ラベル別の各 ActionUnits の箱ひげ図

研究業績

研究会・シンポジウム

(1) 鹿摩大智, 岡哲平, 大西俊輝, 東直輝, 石井亮, 深山篤, 宮田章裕: 聞き手の相槌種類に応じた表情生成システムの基礎検討, 情報処理学会インタラクション 2024 論文集, Vol.2024, pp.658–661 (2024 年 3 月).

(2) 鹿摩大智, 岡哲平, 大西俊輝, 東直輝, 石井亮, 宮田章裕: マルチモーダル情報に基づいて適切な相槌の表情を提示するシステムの検討, 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 (DICOMO2024), Vol.2024, pp.468–474 (2024 年 6 月).