

韻律制御を用いた音声対話システムの共感性向上

齋藤 葵

1. はじめに

近年、音声対話システムのカスタマーサポートへの導入や、メンタルヘルス支援への応用を目的とした研究が進められている。例えば、メンタルヘルス支援においては、患者との対話を通じて支援を行うAIシステムの開発が報告されている[1]。これらの分野に音声対話システムを導入することで、オペレーターなどの支援者が対応する際の心理的・時間的負担の軽減や、業務の効率化が期待されている。そのため、単なる情報提示にとどまらず、ユーザの感情に寄り添った応答を行うシステムが求められている。

機械学習技術、特に深層学習の発展により、音声認識や音声合成の性能は大きく向上した。その結果、従来手法と比較して対話の自然さや円滑さは改善しつつある。一方で、単に正確な内容を伝えるだけでなく、聞き手にとって自然で理解しやすい音声を生成することが必要であり、「自然さ」や「聞きやすさ」の点では人間同士の対話に及ばない。この課題の要因の一つとして、声の高さ(以下、ピッチ)や発話速度などの韻律要素[2]が適切に制御されないことが挙げられ、共感的に受け取られる応答の実現に向けて検討が必要である。

2. 音声対話システムの現状と課題

2.1 音声対話システムの概要

音声対話システムとは、人間の音声入力を受け取り、その内容を解析した上で適切な応答を生成し、音声として出力するシステムである。一般に、音声認識、言語理解、対話管理、応答生成、音声合成といった複数の処理モジュールから構成されている。音声認識では入力音声をテキストに変換し、言語理解ではその意味を解析する。対話管理では対話の状態や文脈に基づいて応答方針を決定し、応答生成によりテキスト形式の応答を生成する。最後に、音声合成によって生成された応答が音声として出力される。

近年の深層学習技術の発展により、音声認識の精度や音声合成の明瞭性は向上している。一方で、人間同士の対話と比較すると、対話の自然さや共感性という観点では、依然として改善の余地が残されている。このような課題に対して、音声対話システムにおいては、話し方の制御や表現方法に関する研究が進められている。

2.2 音声対話システムの韻律制御の必要性

人間の音声には、言語情報に加えて、ピッチや発話速度、イントネーションといった韻律要素が含まれている。これらの情報は、感情表現や相手に対する態度を伝える手段として用いられ、対話の理解や円滑なコミュニケーションにおいて重要な役割を果たしている。西村ら[3]は、利用者にとって円滑な対話を実現する協調的な音声対話システムの実現を目的として、人間同士の対話における印象評価と韻律変化との関係性を分析した。その結果、人間同士の対話において、ピッチや発話速度といった韻律要素が、対話相手の影響を受けて互いに近づく方向へ変化する同調効果¹の相互作用が生じること

を示した。また、Decetyら[4]は、同調的な行動は共感や好意、円滑な交流を促進することを示している。

音声対話システムにおいても、システムが韻律を調整することで、ユーザに自然さや理解されているという印象を与え、共感的に受け取られる可能性を高めると考えられる。特にカウンセリングや対人支援を目的とした応用分野では、韻律の違いが対話に大きな影響を与える可能性がある[5]。

2.3 共感性の認知的側面と情動的側面

Decetyら[4]によれば、共感性は主に認知的側面と情動的側面の二つに大別されるとし、他者の感情状態を認識・理解する側面と、他者の実際または推測される感情状態を情動的に体験する側面から成る概念として定義されている。以下、前者の側面を認知的共感、後者の側面を情動的共感と表す。

本研究では、共感性を認知的側面と情動的側面に分類し、音声対話システムにおける韻律制御がこれら二側面の共感性に及ぼす影響を対象とする。特に、合成音声の韻律制御が、利用者の共感的受容にどのように寄与するかを検討する。

2.4 韻律制御を用いた音声対話システムの課題

本節では、音声対話システムにおける韻律制御に関する先行研究のうち、本研究と関連の深いものを取り上げる。特に共感的応答の実現という観点から、既存手法の問題点を整理し、それらに起因する未解決の課題を明らかにする。

加藤ら[6]は、人間同士の音声データを体系的に収集・整理したデータセット(以下、音声コーパス)からLSTMベースのピッチ推定モデルを学習し、画面上にアバターを表示する対話条件で、ユーザ発話に応じてシステム発話のピッチを動的に制御する手法を提案した。その結果、ピッチ制御により共感性に関する主観評価が向上することを報告している。しかし、本研究が対象とする応用場面を考慮すると、以下の点でさらなる検討が必要である。第一に、制御対象とする韻律要素がピッチのみに限定されている。実際の人間同士の対話における共感的応答は、単一の韻律要素のみの制御では十分でない可能性がある[2][3][4]。したがって、より自然で共感的な応答を実現するためには、複数の韻律要素を統合的に制御することが重要であると考えられる。第二に、アバター表示を伴うシステムでの評価であったため、観察された共感性の向上が韻律制御のみによるものかは明確ではない。実際の応用場面、特にメンタルヘルス支援などでは、電話相談のように音声のみでの対話が想定される場合も多い。そのため、視覚情報に依存せず、音声の韻律制御のみによって共感性を向上させる手法の検討が必要である。第三に、共感性を認知的側面と情動的側面に分けて評価しておらず、包括的な印象評価にとどまっている。加藤らの研究では「人間らしさ」や「豊かさ」といった一般的な印象を測定しているが、共感性を認知的側面と情動的側面に分けて詳細に評価するには至っていない。メンタルヘルス支援などの応用を考えると、共感性をより多面的に捉える評価手法について検討する余地がある。

一方、関ら[7]は、対話相手の発話速度に応じて話者の発話速度が同調することを示している。また、発話速度の同調は、

¹ 対話において対話相手の影響を受けて韻律要素が類似した値へ

の収束するように変化するこ

共感がなくても生じる一方で、共感の程度が高い場合には、より大きく安定して現れることを報告している。しかし、発話速度の類似性が対話の印象に関係するという知見があるにもかかわらず[8]、発話速度制御を音声対話システムに実装し、その効果を検証した研究は十分に行われていない。

そこで本研究では、合成音声のピッチに加えて発話速度も制御可能な音声対話システムを提案する。提案システムと従来システムを被験者に使用させた上で、韻律制御が認知的共感および情動的共感に及ぼす影響について比較検証する。

3. 音声対話システムの構成

3.1 音声対話システムの全体構成と処理手順

本研究で使用した主要構成要素と役割について表 1に、構築した音声対話システムの構成を図 1に示す。まずユーザの発話音声に対して音声区間検出および音声認識を行う。音声区間検出により発話の開始・終了時刻を推定し、発話時間の算出と発話区間の抽出をする。その発話区間に対してピッチ抽出を行い、16 kHz・フレーム長10 msの条件で、各フレームの基本周波数(F0)をHz単位で算出する。その後、無声区間を除外したフレーム単位のF0の平均を、各発話の代表ピッチ値とする。また、音声認識結果をテキスト化し、モーラ数²[10][11]の算出に用いる。なお、モーラ数は小書きの仮名、句読点、空白を除いて、促音と長音を含めて計測した。算出されたモーラ数を発話秒数で除することで、発話速度を計算する。音声認識で得たユーザ発話テキストをChatGPT APIに投入し、応答文を生成する。取得したピッチと発話速度は各推定モデルに投入し、次発話の変化率を推定する。推定結果に基づき応答音声のピッチと発話速度を設定し、設定値と応答文を音声合成器に投入して音声応答を出力する。

表 1 主要構成要素と役割

Python(Version:3.10.11)	プログラミング言語
Google Cloud Speech-to-Text ³	音声認識
Silero VADで公開されている学習済みニューラルネットワークモデル ⁴	ユーザ発話の音声区間検出・抽出
TorchCrepes ⁵ [9]	ピッチ抽出
ChatGPT API (GPT-4.1-miniモデル ⁶)	応答文生成
Google Cloud Text-to-Speech ⁷	生成した応答文を音声合成 ⁸

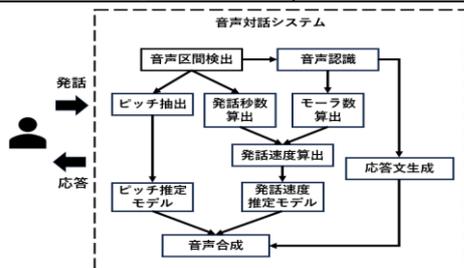


図 1 音声対話システムの構成図

また本研究では、音声の制御方法の違いが対話に与える影響を比較するため、3つの応答手法を比較条件として用意した。手法1は韻律制御を行わないベースライン手法であり、ピッチおよび発話速度には音声合成器の標準値を用いて応答を生成する。手法2はピッチのみを制御する従来手法であり、発話速度には音声合成器の標準値を用いて応答を生成する。

手法3はピッチおよび発話速度の両方を制御する手法であり、本研究の提案手法とする。

3.2 LSTM による韻律推定モデルの構築

本研究では、ピッチ推定モデルと発話速度推定モデルをLSTM (Long Short-Term Memory) に基づくニューラルネットワークとして構築した。LSTMは時系列データにおける長期的な依存関係を考慮した学習が可能であり、発話に伴う韻律要素の時間的変化を扱うのに適している。また先行研究[6]においても同様の手法が用いられているため、本研究でもLSTMを採用した。ピッチ推定モデルと発話速度推定モデルは、それぞれ独立したニューラルネットワークとして構築した。両モデルとも2層のLSTMと3層の全結合層で構成され、ピッチ推定モデルはLSTM隠れ層サイズ64、発話速度推定モデルは128とした。学習にはGPU⁹を使用し、最適化アルゴリズムにはAdam (学習率0.001) [12]、損失関数は平均二乗誤差を採用した。バッチサイズは32、最大エポック数は400に設定した。

学習データには、日本語共感的音声対話コーパス (STUDIES)¹⁰[13]に収録されている対話データのうち、1対話当たり10から20ターン¹¹の対話150件を使用した。この音声コーパスは、女性講師と生徒(男子または女子)との1対1の模擬音声対話から構成される。本研究では、支援的・助言的対話を対象とするため、講師-生徒という非対等な関係性の下で感情の読み取りや支援的応答が含まれる本コーパスを、学習データとして用いた。モデル学習にあたり、先行研究[6]に倣い、連続する発話間のピッチ変化率および発話速度変化率を算出し、これらを時系列データとして扱った。変化率は次式により算出する。

変化率 = V_n / V_{n-1} (V_n : n 番目の発話に対応する韻律量)
ここで V_n は推定対象に応じて異なり、ピッチ推定では各発話の代表ピッチ値、発話速度推定では各発話の発話速度を表す。以下、生徒の変化率を S_n 、講師の変化率を T_n と表す。

各モデルでは、3ターン分の対話(生徒3発話、講師3発話)から学習サンプルを生成する。各話者の連続する3発話からは、隣接発話間の変化率を2点ずつ算出できる。本研究では、生徒の2つの変化率(S_n, S_{n+1})と講師の1つ目の変化率(T_n)を入力し、講師の2つ目の変化率(T_{n+1})を出力として学習させた。すなわち、生徒側の連続する韻律変化に対して、講師側が次に示す韻律変化を予測することで、生徒の韻律変化に対する講師の応答パターンをモデル化する(図 2)。学習・評価にあたっては、生成した学習サンプルからなるデータセットを訓練用80%、テスト用20%に分割し、訓練データで学習したモデルをテストデータで評価した。発話速度の学習に変化率を使用した理由は、話者ごとの絶対的な発話速度の差を抑え、対話中における相対的な変化に着目するためである。また、ピッチについても変化率を用いて学習を行っており、両韻律要素を同一の表現形式で扱うことで、モデル設計および解釈の一貫性を保つことを目的とした。

本研究では、女性講師の発話特性を推定対象としてモデルを学習した。ピッチはTorchCrepesにより発話区間のフレームごとに基本周波数を推定し、平均値を各発話の代表値とした。学習用の韻律特徴量は、3.1節で述べた手順に従って算出し

² 発音を区切る最小のリズム単位の数

³ <https://cloud.google.com/speech-to-text?hl=ja>

⁴ <https://github.com/snakers4/silero-vad>

⁵ <https://github.com/maxmorrison/torchcrepe>

⁶ <https://platform.openai.com/docs/models/gpt-4.1-mini>

⁷ <https://cloud.google.com/text-to-speech?hl=ja>

⁸ 韻律を制御しながらテキストを音声に変換すること

⁹ NVIDIA GeForce GTX 1650

¹⁰ <https://research.nii.ac.jp/src/STUDIES.html>

¹¹ 生徒の発話とそれに対する講師の発話からなる1往復の対話を1ターンと定義

たが、モーラ数の算出には音声認識結果ではなくアノテーション済みデータの発話内容を用いた。これらにより、算出した発話単位のピッチ値および発話速度から変化率を計算し、モデル学習に使用した。

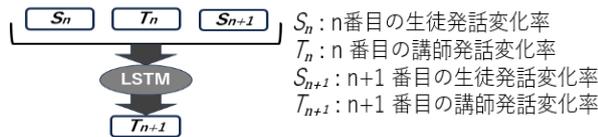


図 2 LSTMを用いた講師発話変化率予測モデルの構成

4. 音声対話システムの評価

4.1 被験者実験の手順

本研究では、音声対話システムの共感性向上の効果を検証するため、被験者実験を実施した。被験者は20代25名(男性10名, 女性15名)であり, 主に会津大学短期大学の学生を中心に募集した。音声入力にはUSBマイクであるUber Mic¹²を使用し, 音声出力にはASUS ROGflowX13のノートPC内蔵スピーカーを使用した。環境雑音の影響を抑え実験条件を統一するため, 静かな屋内環境で実施した。被験者は3手法についてそれぞれ約5分間の音声対話を体験し, 各対話後に共感性を7段階尺度で評価するアンケートに回答した。なお, 各手法の提示順による影響を抑えるため, 提示順は全6通りとし, 各順序を被験者に偏りなく割り付けた。

被験者には生成された応答内容ではなく, 音声応答の話し方(音声表現)に着目して評価を行うよう指示した。ただし, 評価バイアスを避けるために各手法の差異がピッチおよび発話速度にあることは事前に説明していない。被験者の感情表出を伴う対話を促し, 共感的応答が成立しやすい状況を確保するため, 本実験では「悩み事」「現在関心をもってること」「好きなこと」の3つを発話テーマとして設定したが, 各手法において発話内容の詳細までは統一していない。本システム側の応答文も生成モデルによって都度生成されるため, 手法間で対話内容が完全に一致する条件にはなっていない。そのため, 共感性の評定には, 韻律制御の違いに加えて, 話題や応答内容の差異が影響している可能性がある。

4.2 評価指標

本研究では, 音声対話システムに対する共感性を評価するため, 主観評価アンケートを用いた。加藤ら[6]と笠井ら[14]が提案する質問項目を参考に, 認知的共感および情動的共感に関する8項目と対話品質評価項目の3項目からなる構成とした(表 2)。

表 2 共感性主観評価項目

No.	質問項目	共感の分類
1	私が何を考えているか理解しているように感じた	認知的
2	私の状況を把握しているように感じた	認知的
3	私の意図をくみ取っていた	認知的
4	私が何を求めているのか分かっていた	認知的
5	私の感情に共感してくれた	情動的
6	私の気持ちに寄り添ってくれた	情動的
7	私の感情を受け止めてくれた	情動的
8	私を気にかけてくれた	情動的
9	親しみやすかった	対話品質評価
10	人間らしさを感じた	対話品質評価
11	対話しやすかった	対話品質評価

※7 段階尺度 (1:全くそう思わない~7:非常にそう思う)

4.3 実験結果

評価アンケートの結果について3つの手法間の比較を行っ

た。本研究では同一被験者が複数条件を評価しているため, 対応のある比較が必要であり, 評価尺度が順序尺度であることから, 分析方法としてノンパラメトリック検定のフリードマン検定を用いた。フリードマン検定は手法間に全体として差があるか判断できるが, どの手法間に差があるかまでは明らかにできないため有意差が認められた場合には, 多重比較を行った。

表 3 フリードマン検定の結果(N=25)

	P値	*:P<0.05 **P<0.01
共感性	0.00048	**
認知的共感	0.00078	**
情動的共感	0.0198	*
対話品質評価	0.0069	**

分析では, 各被験者について手法ごとに尺度得点を算出し, それらを用いて3手法の差を検定した。尺度得点は, 共感性(項目1~8の平均), 認知的共感(項目1~4の平均), 情動的共感(項目5~8の平均), 対話品質評価(項目9~11の平均)として求めた。

フリードマン検定の結果, いずれの尺度でもP<0.05で手法間に有意差が認められた(表 3)。この結果を受けて, 有意差の所在を明らかにするため, フリードマン検定で有意差が認められた尺度について, 多重比較を行った。多重比較にはScheffe法を用い, 有意水準はScheffe法により補正した上で, 各手法間の平均順位差を検定した。

表 4 共感性の平均順位 の多重比較(N=25)

比較する手法	平均順位	P値	*P<0.05 **P<0.01
手法1 vs 手法2	1.620 vs 1.760	0.8802	
手法1 vs 手法3	1.620 vs 2.620	0.0015	**
手法2 vs 手法3	1.760 vs 2.620	0.0081	**

表 5 認知的共感の平均順位 の多重比較(N=25)

比較する手法	平均順位	P値	*P<0.05 **P<0.01
手法1 vs 手法2	1.660 vs 1.740	0.9588	
手法1 vs 手法3	1.660 vs 2.600	0.0030	**
手法2 vs 手法3	1.740 vs 2.600	0.0077	**

表 6 情動的共感の平均順位 の多重比較(N=25)

比較する手法	平均順位	P値	*P<0.05 **P<0.01
手法1 vs 手法2	1.660 vs 1.940	0.5766	
手法1 vs 手法3	1.660 vs 2.400	0.0214	*
手法2 vs 手法3	1.940 vs 2.400	0.2263	

表 7 対話品質評価の平均順位 の多重比較(N=25)

比較する手法	平均順位	P値	*P<0.05 **P<0.01
手法1 vs 手法2	1.640 vs 1.880	0.6818	
手法1 vs 手法3	1.640 vs 2.480	0.0092	**
手法2 vs 手法3	1.880 vs 2.480	0.0913	

多重比較の結果, 共感性平均および認知的共感平均において, 手法2(ピッチのみ)と手法3(ピッチおよび発話速度)の間にP<0.01で有意差が認められた。また, 解析対象とした全指標において手法1(ベースライン)と手法3の間に有意差が認められた一方, 手法1と手法2の間には有意差が認められなかった。さらに, 平均順位では, 有意差が得られなかった指標においても手法3が最も高く, 高評価となる傾向が示された。

共感性(8項目平均)では, 手法3は他の手法と比較して中央値が高く, 高評価となる傾向が示された。一方, 手法1および手法2では分布が近く, 両者の間には大きな差は見られなかった(図 3)。また, 認知的共感では, 手法3は手法1および手法2より中央値が高く, 分布も上方に位置している(図 4)。有意差が認められない比較においても, 平均順位および中央値は手法3が最も高くなる傾向が見られた(図 5, 図 6)。

¹² <http://m-audio.jp/uber-mic/>

以上の結果から、ピッチおよび発話速度の両方を制御した手法3は、制御しない手法1とピッチのみ制御した手法2と比較して、評価アンケートにおいて総じて高評価であった。特に、共感性に関する8項目および認知的共感において、手法2と手法3の間に有意差が認められ、発話速度制御の付加が評価向上に寄与する可能性が示唆された。一方で、情動的共感および対話品質評価では、手法1と手法3の間に有意差が認められたが、手法2と手法3の間では有意差が認められなかった。しかし、中央値および平均順位はいずれも手法3が最も高く、手法3が高評価となる傾向が示された。

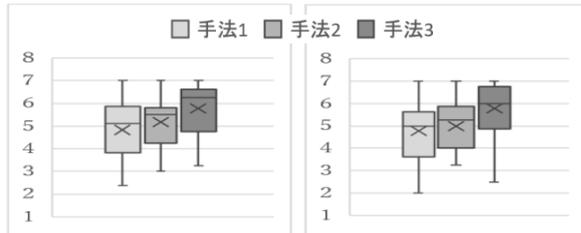


図 3 共感性評価(8項目平均)の箱ひげ図(左)¹³

図 4 認知的共感評価(4項目平均)の箱ひげ図(右)

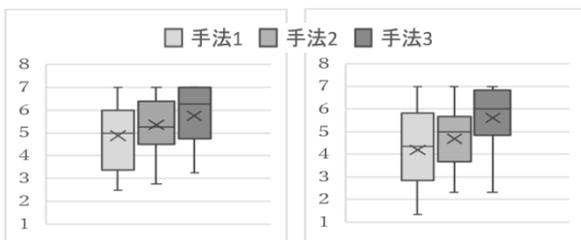


図 5 情動的共感評価(4項目平均)の箱ひげ図(左)

図 6 対話品質評価(3項目平均)の箱ひげ図(右)

4.4 考察

実験の結果、提案手法である手法3(ピッチおよび発話速度を制御)は、ベースライン手法である手法1(韻律制御なし)および従来手法である手法2(ピッチのみ制御)と比較して、共感性(8項目平均)と認知的共感において一貫して高い評価を示し、多重比較でも手法2と手法3の間に有意差が認められた。このことから、ピッチに加えて発話速度も制御することで、発話の韻律要素がより多面的に変化し、対話相手に「理解されている」という印象が形成されやすくなった可能性がある。韻律要素は感情表現や相手に対する態度の手がかりとなるため、複数の韻律要素を同時に制御することが、共感的印象の向上に寄与したと考えられる。

手法2と手法3を比較した結果、認知的共感において有意差が認められた一方、情動的共感では有意差が認められなかった。このことから、提案手法は感情の共有そのものよりも、話し手の状態を「理解してくれている」と感じさせる側面により強く影響した可能性がある。加えて、手法1と手法2の間に有意差が認められなかった理由として、ピッチのみの制御では韻律変化が限定的であり、被験者にとって明確な違いとして知覚されにくかったと推測される。

一部の被験者では、提案手法である手法3と他手法の評価差が小さかった。ユーザ自身の発話が抑揚に乏しく、ピッチや発話速度の変動が小さい場合には、ユーザの韻律特性に基づいて推定される制御量も小さくなり、応答音声の韻律変化が限定的となる。その結果、手法による差が知覚されにくかった可能性がある。また、手法3の評価が相対的に低い被験者もあり、韻律制御による話し方が被験者の期待と一致しない場合には、かえって違和感を生じさせたと考えられる。

これらの結果から、ピッチと発話速度を合わせて制御する韻律制御手法は、共感的印象の向上に有効である可能性を持つ一方、適応の度合いが小さい場合には制御の効果が知覚されにくく、個人差の影響を受けやすいことが示唆される。今後は、ユーザごとの韻律特性や嗜好を考慮した、より精緻な適応制御の設計が必要である。

5. むすびに

本研究では、音声対話システムにおける共感性向上を目的として、ピッチと発話速度の両方を制御する手法を提案し、被験者実験によってその有効性を示した。実験の結果、提案手法は、制御を行わない手法および単一要素のみを制御する手法と比較して、共感性評価において高い評価を得た。特に認知的共感では明確な効果が見られ、複数の韻律要素を統合的に制御することの有効性が示された。

本研究の意義は、共感性を認知的・情動的側面から測定する枠組みを提示し、複数の韻律要素の統合制御の効果を実証的に示した点にある。一方で被験者の限定性、個人差への対応、長期使用における効果検証などの課題が残されている。今後は、被験者数を増やした追実験、個人差を考慮した適応制御手法の改善、他の韻律要素の統合、および実環境での評価が必要である。

参考文献

- [1] 岡山大学, AIを活用したメンタルケアサポートシステムを開発〜患者さんとの対話で心に寄り添う AI〜, https://www.okayama-u.ac.jp/tp/release/release_id1248.html (参照:2026-1-2).
- [2] 前川喜久雄, “韻律とコミュニケーション”, 日本音響学会誌, 55 巻 2 号, pp.119-125, 1999.
- [3] 西村良太ほか, “音声対話における韻律変化をもたらす要因分析”, 音声研究, Vol.13, No.3, pp.66-84, 2009.
- [4] J.Decety, P.L.Jackson, “The Functional Architecture of Human Empathy”, Behavioral and Cognitive Neuroscience Reviews, Vol. 3, No.2, pp. 71-100, 2004.
- [5] D.Tao et al., “A Study on Prosodic Entrainment in Relation to Therapist Empathy in Counseling Conversation”, in Proc. INTERSPEECH 2023, pp.3662-3666, 2023.
- [6] 加藤弘泰, 李晃伸, “音声対話システムにおける同調効果のためのコーパスに基づくピッチ制御”, 情報処理学会第 84 回全国大会講演論文集, 2022 巻 1 号, pp.301-302, 2022.
- [7] 関洋平, 竹内勇剛, “対話場面における相手の発話に対する応答の発話速度の同調”, 日本認知科学学会大会発表論文集, Vol.26, pp.2-13, 2009.
- [8] D.B.Buller, R.K.Aune, “The Effects of Speech Rate Similarity on Compliance: Application of Communication Accommodation Theory”, Western Journal of Communication, Vol.56, No.1, pp.37-53, 1992.
- [9] J.W.Kim et al., “CREPE: A Convolutional Representation For Pitch Estimation”, in Proc. IEEE ICASSP, pp.161-165, 2018.
- [10] 石黒翔, 齊藤智, “日本語のモーラ数を数数するシェルスクリプトプログラムの提案”, 日本認知心理学会第 16 回大会発表論文集, P1-036, 2018.
- [11] 李在鎬ほか, 認知音韻・形態論, くろしお出版, 2013.
- [12] D.P.Kingma, J.L.Ba, “Adam: A Method for Stochastic Optimization”, Proc. ICLR, 2015.
- [13] 齋藤祐樹ほか, “STUDIES: 表現豊かな音声合成に向けた日本語共感的対話音声コーパス”, 日本音響学会研究発表会講演論文集, 2022 巻, 2-3P-15, 2022.
- [14] 笠井有華ほか, “対話型 AI の共感性向上へのアプローチ”, 人工知能学会全国大会論文集, 38 巻, 1N3-GS-9-03, 2024.

¹³ 箱は第 1 四分位数から第 3 四分位数を、横線は中央値を、×印は

平均値を、ひげは最小値と最大値を示す。