

対話システムにおけるユーザの交替潜時に同調した発話制御とその影響分析

Adaptive Control of Utterance Timing based on User Conversational Silence in Spoken Dialogue Systems

宇治川 遥祐^{1*} 高汐 一紀²
Yosuke Ujigawa¹ Kazunori Takashio²

¹ 慶應義塾大学大学院 政策・メディア研究科

² 慶應義塾大学大学院 環境情報学部

Abstract: 円滑で協力的な対人コミュニケーションにおいて、話者間の交替潜時が相互に引き込まれ、同調していく現象は広く知られている。本研究では、非言語的なタイミングの同調が、対話エージェントに対する社会的受容性や対話の質に与える影響を明らかにすることを目的とする。ユーザーの潜時を対話の中で計測し、潜時を動的に制御する対話システムを実装した。「収束」と「同期」の2つの同調手法の有効性、および対話に与える影響について、主観評価と生体指標を用いた客観的指標の両面から分析を行った。

1 はじめに

近年、音声対話システムは、スマートスピーカやスマートフォン上の音声アシスタントとして日常生活に深く浸透している。その役割は、天気予報や家電操作といった単発的なタスク実行から、教育、高齢者の見守り、メンタルヘルスケアといった、人間との長期的な信頼関係や深い相互理解を必要とする社会的存在へと急速に変容しつつある。このような社会的対話において、ユーザの体験価値を決定づけるのは、言語的な内容の正しさだけではない。「いつ、どのような間（ま）で話すか」という非言語的な時間構造が、ユーザの受容性や信頼感に決定的な影響を与える。

人間同士の円滑なコミュニケーションにおいて、対話者間の発話速度、リズム、声の大きさといった韻律的特徴が相互に影響し合い、類似していく現象は「同調」として知られている。心理学や認知科学の知見によれば、この同調現象は単なる表面的な模倣ではない。同調は、相手との社会的距離を縮め、信頼関係を形成するための戦略的行動となる側面や、同調は対話者間の認知リソースを節約し、相互理解を促進する機能的役割を果たす側面を持つ [1, 2]。人間と共生する対話システムにとって、ユーザの振る舞いを感知し、適切に同調する能力は、自然で質の高いインタラクションを実現するために注目されている。

しかし、既存の対話システムにおける同調の実装には、固定のパラメータやユーザの平均的な値で制御を行っているという課題が残されている。従来の研究の多くは、ユーザの対話全体を通じた平均的な特徴量にシステムを近づける「収束」のアプローチを採用してきた。しかし、実際の対話は、思考の深化、迷い、感情の起伏に伴い、テンポが対話の中で変化する動的なプロセスである。ユーザが深く考え込み、テンポが遅くなっている局所的な瞬間に、システムが「平均的な速度」で応答を行えば、ユーザの思考リズムと乖離が生まれ、対話への没入感は損なわれる。すなわち、求められているのは対話全体での値の収束ではなく、ユーザの対話の中での状態変化に追従する「同期」である。

動的な同期を実現するため、本研究では特に「発話潜時」に着目する。対話の時間構造は、発話権が移動する際の「話者交替潜時 (Gap)」だけで決定されるものではない。発話内における沈黙や言い淀みである「発話内潜時 (Pause)」もまた、ユーザの状態が表出された重要な指標である。従来の研究は Gap の最適化（いかにスムーズに交代するか）に注力してきたが、ユーザの思考に寄り添うためには、Gap と Pause を統合的に扱い、対話全体の「間」のテンポを共有する必要がある。

本研究の目的は、この Gap と Pause を統合した発話潜時に対し、ユーザの変化に動的に同期する対話システムを構築し、その有効性を実証することである。本稿では、音声対話コーパス分析による同調現象の解明から、潜時制御を行う対話システム実装、そして被験

*連絡先:

慶應義塾大学大学院 政策・メディア研究科
252-0882 神奈川県藤沢市遠藤 5 3 2 2
E-mail: ujigawa@keio.jp

者実験を通した主観・客観的指標の両側面による対話の評価までを包括的に論じる。

2 関連研究：対話における時間構造と同調

音声対話システムが、機能的な性能に留まらず、社会的な関係性向上を果たすためには、言語的な内容の適切さだけでなく、非言語的な時間構造の制御が不可欠である。本章では、人間同士のコミュニケーションに見られる「同調」の構造を概観し、既存の対話システム研究における実装上の課題を指摘する。その上で、従来の研究で扱われてきた「収束」と「同期」を明確に区別し、本研究で注目する「発話潜時の動的制御」の位置付けを論じる。

2.1 音声対話における同調

人間は対話を行う際、発話速度、音の強弱、ピッチといった韻律的特徴や、使用する語彙、構文構造を、無意識のうちに相手に近づける傾向を持つ。この現象は「同調」と呼ばれ、その発生機序については主に二つの有力な理論によって説明される。第一は、Giles らによるコミュニケーション適応理論 (Communication Accommodation Theory: CAT) である [1]。CAT によれば、同調は相手との社会的距離を縮め、好意や承認を獲得するための無意識的な戦略の適応行動であるとされる。第二は、Garrod と Pickering による相互同調モデル (Interactive Alignment Model: IAM) である [2]。IAM では、同調は高度な意図によるものではなく、知覚と行動のプライミング効果によって、対話者間の認知リソースが共有・節約される自動的なプロセスであると説明される。

これらの理論的背景に基づき、音声対話システムの研究においても、同調の実装が試みられてきた。例えば、Levitan らは、システムがユーザの音声強度や発話速度に同調することで、ユーザからの信頼感や好感度が有意に向上することを報告している [3]。また、Lubold らは教育支援ロボットを用いた実験において、韻律的同調が学習者のモチベーション維持やタスクパフォーマンスの向上に寄与することを示している [4]。しかし、これらの先行研究の多くは、対話全体を通した「平均発話速度」や「平均ピッチ」といった大域的な特徴量を対象としており、次節で述べるような、対話の中での時間的な変化への適応については十分に議論されていない。

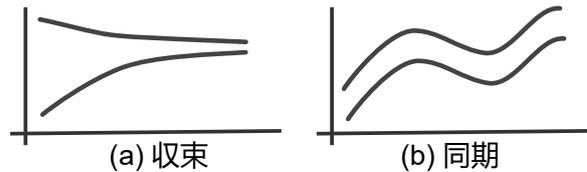


図 1: 同調の分類

2.2 収束と同期の区別

「同調」という用語は広義に用いられるが、その時間的な振る舞いに着目すると、大きく異なる二つの現象が混在している。本研究では、先行研究に基づき、対話システムの制御変数を明確化するため、両者を以下のように区別する [5] (図 1)。

- **収束**: 対話の進行とともに、話者間の特徴量の差異が減少し、特定の値 (例えば二者の平均値) へと漸近していく現象を指す。これは、近接性の動的な変化として捉えることができ、対話の進行に伴って、話者 A と話者 B の特徴空間上の距離が減少していく現象である。例えば、早口のユーザに対してシステムも発話速度を上げ、互いに同程度の速度で会話を行う状態は、収束性が高いと言える。逆に、一方が速く他方が遅いまま平行線をたどる場合は、収束性は低い。Levitan ら [6] や Pardo らの研究 [7] は、韻律的同調の程度を評価するために、対話ペアの特徴量の音声的距離を算出し、その変化を指標とした。
- **同期**: 互いの特徴量の絶対値が一致するかに関わらず、その時系列的な変動傾向 (上がった・下がった) が相関し合う現象を指す。これは、対話者間の行動が時間的に連動しているかを示す指標である。例えば、ユーザが一時的に発話速度を上げた際、システムもそれに呼応して速度を上げ、ユーザが速度を落とせばシステムも落とすといった挙動が見られる場合、たとえ両者の平均速度に差があったとしても、同期性は高いと評価される。Lahiri ら [8] や Rilliard ら [9] は、変化傾向を評価するために、韻律パターン (F0 や強度曲線) の類似度を算出し、対話における韻律的同調の指標として扱った。

Edlund らは、対話者間の韻律的同調を捉えるには、対話全体の平均値などを用いる指標では不十分であり、時間的な変化を考慮した評価が必要であると論じている [10]。実際の対話は、思考の深化や感情の起伏に伴いテンポが刻々と変化する動的なプロセスである。したがって、特定の時点での値を相手に合わせる「収束」

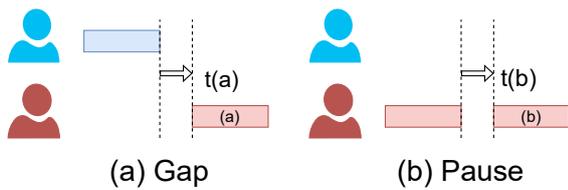


図 2: 発話潜時と発話区間の分類

だけでは、ユーザの瞬時の状態変化に対応することはできない。本研究では、従来扱われている値を近づける「収束」だけでなく、ユーザの対話中の変化に追従する「同期」の観点も含め、多角的に同調を評価する。

2.3 発話潜時の構成要素：Gap と Pause

韻律情報の中でも、発話のタイミングに関わる「時間構造」は、ターン交代のスムーズさや対話の「間」を決定づける基本的な要素である。本研究では、この時間構造を構成する要素として、以下の二つに着目する(図 2)。

- **話者交替潜時 (Gap)**：話者交替の瞬間に生じる沈黙。話者 A の発話終了から、話者 B の発話開始までの「話者間無音区間」である。
- **発話内潜時 (Pause)**：同一話者の発話内 (Turn 内) に存在する沈黙。話者が発話権を保持したまま、次の言葉を検索・計画している間に生じる「話者内無音区間」である。

本研究では、対話のテンポ形成に寄与する主要なパラメータとして、Gap および Pause の両者に着目する。従来の研究では、話者交替のスムーズさを示す指標として Gap が主に注目されてきた [11]。しかし、Jens ら [10] が指摘するように、Gap と Pause は互いに独立したのではなく、双方が連動して対話全体のテンポ感を形成している。Gap は「他者の発話終了から、自己が発話権を行使するまでの決定時間」、Pause は「自己の発話継続において、次の言葉を紡ぎ出すまでの決定時間」として解釈ができる。これらは発生の文脈こそ異なるものの、いずれも話者が自身の持つテンポに基づいて沈黙を制御し、対話の時間軸を構築する能動的なプロセスである点において共通している。この解釈に基づき、本研究では測定されたすべての潜時を、その直後の発話者のテンポ特性や認知状態を反映した統合的な指標として分析する。

3 人間同士の対話における潜時同期の分析

本章では、質の高い対話を行うシステム設計の根拠を得るために、人間同士の自然な対話において発話潜時がどのように制御され、その時間的構造が対話の質にどのような影響を与えているかを分析する。具体的には、対話者間の潜時制御において、対話全体を通して平均値へと近づく「収束」が重要なのか、あるいは対話の中で変化する傾向が類似する「同期」が重要なのかを明らかにする。この分析結果は、後述する対話システムの制御モデルにおける基本方針を決定づける重要な指針となる。

3.1 分析の目的と対象データ

3.1.1 分析の目的

本分析の目的は、発話潜時の同調現象における定義の違いが、対話の客観的評価に与える影響を比較検証することにある。先行研究において、対話者間のリズムやタイミングの同調は、ラポール形成や対話の円滑化に寄与すると広く議論されてきた。しかし、その「同調」が具体的にどのような時間的特性を指すのかについては、一意な定義が存在しない。従来研究の多くは、相手の発話速度やポーズ長などの平均的な特徴量に合わせる「収束」を指標としてきたが、実際の対話は思考の深化や感情の起伏に伴いテンポが動的に変化するプロセスである。本分析では、質の高い対話において支配的な同調特性が「収束」と「同期」のいずれであるかを定量的に検証し、対話システムの設計指針にすべき潜時制御の手法を明らかにする。

3.1.2 使用したコーパスの仕様と選定理由

人間同士の自然な対話における発話毎の発話潜時の同調現象と、その対話品質への影響を調査するために、RWCP (Real World Computing Partnership) 知的資源ワーキンググループによって構築された「RWCP 音声対話データベース (RWCP-SP96)」 [12] を用いる。本コーパスの基本仕様を表 1 に示す。

本コーパスを選定した妥当性は以下の 3 点にある。

1. **自然な発話の収録**: 台本のない自発的な対話が含まれており、話者交替潜時 (Gap) や発話内潜時 (Pause) が自然に発生している点
2. **タスク指向性による質の定義**: 「車の購入相談」などのタスク指向対話であり、情報のやり取りや合意形成が目的であるため、対話の活性化や協調性による質の定義が妥当である点。

表 1: RWCP-SP96 の概要

項目	内容
提供機関	技術研究組合 新情報処理開発機構
DOI	10.32130/src.RWCP-SP96
対話形式	対面・タスク指向型対話 (1 対 1)
使用言語	日本語
タスク	自動車の購入, 海外旅行の計画
話者構成	説明員 (専門家) vs 質問者 (顧客)
データ量	計 48 対話 (各タスク 24 対話)

3. 役割の非対称性: 収録タスクが, 専門的な知識を持つ「説明員」と, 情報を集める「質問者」という役割分担の元に行われている点. この情報の非対称性は, 知識ベースを持つシステムと情報を検索するユーザという, 一般的な音声対話システムの構図と極めて親和性が高く, 対話システム設計への直接的な応用可能性が高いと考えられる.

本研究では, このコーパスから抽出した対話データに対し, Gap および Pause の 2 種類の潜時データを抽出した. なお, 一般的に扱われる発話の種別である Overlap (発話の重なり) や, Backchannel (実質的な発話権の移動を伴わない短い相槌等) などについては, 沈黙時間を中心としたテンポ分析のノイズとなるため, 分析対象から除外した. 分析に使用された総発話数は 7,108 件であり, その内訳として, 話者交替を伴う Gap が 3,104 件, 同一話者による継続発話間での沈黙である Pause が 2,793 件含まれている.

3.2 コーパスの分析手法

対話の質と潜時同調の関係を明らかにするため, 本研究では「収束」と「同期」という異なる特性に対応する 2 つの評価指標を導入した. また, 分析の軸として, 対話の質を定量化するために「対話活性度 (全発話時間の総和)」と「対話協調性 (話者間の発話時間の差の少なさ)」を定義し, これらが共に高い対話を「良質な対話 (同調群)」, 低い対話を「非同調群」として比較を行った.

3.3 対話の質の客観的評価指標

同調した対話がどのような質的特徴を持つかを定量的に評価するため, 以下の 2 つの客観的指標を導入する.

まず, 対話の盛り上がりや測る指標として「対話の活性度 (Activity)」を定義する. これは対話全体で行われた話者の発話の総数である. 本指標の値が大きい

ほど, 積極的に発話が多く行われ, 対話が活性化した状態であると解釈する.

$$Activity = N_A + N_B \quad (1)$$

次に, 対話のバランスを測る指標として「対話の協調性 (Cooperativity)」を定義する. これは総発話数に対する話者間の発話数差の割合である. この値が小さい (0 に近い) ほど, 発話権が一方に偏らず, 双方が対等に参加しており, 協調性が高い対話であると解釈する.

$$Cooperativity = \frac{|N_A - N_B|}{N_A + N_B} \quad (2)$$

3.3.1 収束の評価手法: 相関分析

対話の進行に伴って二者間の差異が漸的に減少していく「発話潜時の収束」の評価手法について述べる. 収束とは, 時系列データの形状そのものではなく, 「時間経過」と「発話特徴の差異」という 2 変数間の関係性として捉えられる時間発展的なプロセスである. 本分析では, この評価指標として Pearson 相関係数を採用した. 収束現象が成立する場合, 時間の経過に伴って差異は減少傾向を示すため, 両者の間には負の線形関係が観測されるはずである [6]. Pearson 相関係数は, この変量間の線形な依存関係の強さを検定する統計手法として標準的であり, 対話全体を通じた変化の傾向を数学的に検証する上で妥当であると考えられる.

対話全体を通じた発話潜時の値の線形的な一致度をもって, 収束の指標とする. この値が正に大きい (1.0 に近い) ほど, 同じタイミングで同じような潜時をとる (収束している) と定義する.

3.3.2 同期の評価手法: DTW による時系列解析

特定の時点において二者の変化やパターンが整合している「発話潜時の同期」の評価手法について述べる. 人間の対話行動には, 知覚から運動生成に至るまでの反応潜時という生理的な制約が不可避である [13]. また, 発話速度の個人差や感情によって, 同調行動であってもその継続長は動的に変動する. そのため, 実質的な同期は物理的な同時性ではなく, 一定の位相差や時間的伸縮を含んだ追隨行動として現れる. 一般的なユークリッド距離や単純な相関分析は時間軸の厳密な一致や等しい長さを前提とするため, 線形補完等で正規化を行ったとしても, 局所的な韻律構造を歪め, 人間特有の「ゆらぎ」を過剰な不整合として判定してしまう問題がある [9]. そこで本分析では, この評価指標として動的な時間伸縮法 (Dynamic Time Warping: DTW) を採用した. DTW は時間軸の非線形な伸縮によって

位相差および発話長の差異を吸収することが可能であり、遅れや速度差を伴う同調を「最小累積距離」という距離コストとして評価関数に組み込むことができる。これにより、生理的な制約を考慮した上で、時系列パターン of 構造的な類似度を適切に測定することが可能となる [8].

DTW は、時間軸の非線形な伸縮を許容しながら系列間の距離を最小化する手法である。算出された累積距離が 0 に近いほど、互いの変化パターンの距離が近い（同期している）と定義する。

3.4 分析結果：変化への同期の優位性

定義した 2 つの同調指標に基づき、全 48 対話をそれぞれ「高同調群」と「低同調群」に中央値を用いて分割し、客観的評価指標（活性度・協調性）に有意な差異が見られるかを分析した。

各指標の正規性を Shapiro-Wilk 検定により確認した結果、多くの変数において正規性が棄却されたため ($p < 0.05$)、群間の比較にはノンパラメトリック検定である Mann-Whitney の U 検定を用いた。さらに、サンプルサイズに依存する p 値のみの報告による誤解を避けるため、実質的な差の大きさを表す効果量として r 値 ($r = |Z|/\sqrt{N}$) を算出し、Cohen の基準 (0.1: Small, 0.3: Medium, 0.5: Large) に基づき評価を行った。

分析の結果、対話の質に対して「収束」と「同期」が与える影響には明確な差異が確認された。

3.4.1 Gap と Pause の統合的理解の重要性

Gap 発話および Pause 発話の各潜時データに対し、Pearson 相関係数と、DTW 距離を用いた場合の結果比較を表 2 に示す。表中の数値は、活性度の指標となる発話総数、および協調性の指標となる発話数差の中央値と四分位範囲を表している。

分析の結果、Pearson 相関を用いた場合、Gap と Pause のいずれの潜時においても、対話の活性度 (Activity) および協調性 (Cooperativity) に対する有意な関連性は確認されなかった ($p > 0.05$)。これに対し、DTW を用いた分析では、Pause 単体のデータにおいて、対話の活性度に対し有意な差が検出された ($p < 0.05$, $r = 0.38$)。具体的には、継続発話の間である Pause 発話が同期している高同調群は、低同調群と比較して有意に多い発話総数を記録した。

一方で、協調性の指標である発話数差については、いずれの手法および潜時においても有意な差は見られなかった ($p > 0.05$)。この結果は、Pause 発話の同期が発話量の増加には寄与するものの、特定の話者に偏らない発話量のバランスまでは保証しないことを示唆し

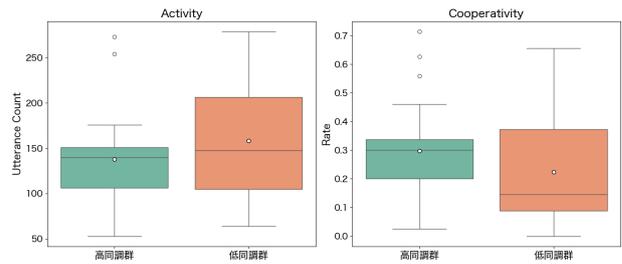


図 3: Pearson 相関に基づいた収束の定義による対話への影響

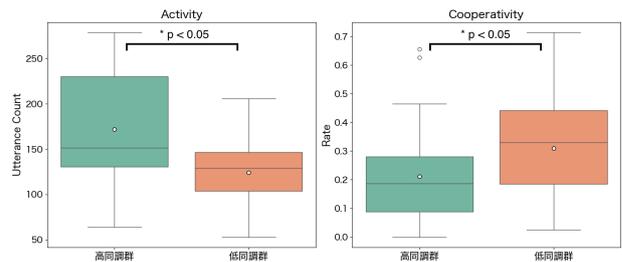


図 4: DTW に基づいた同期の定義による対話への影響

ている。したがって、活性度と協調性の双方を向上させる要因を特定するため、次節ではこれらを統合したデータによる分析を行う。

3.4.2 収束の限界と同期の有意性

前項の結果を踏まえ、Gap と Pause を統合した時系列データにおける分析結果を表 2 に示す。分析結果の基本統計量および検定結果の一覧から適用した同調指標によって統計的な有意差の有無が異なる結果が読み取れる。以下に、各同調指標に基づいた群分け結果の詳細を述べる。

Pearson 相関係数（発話潜時の絶対値の一致度）に基づき群分けを行った結果（図 3）、高同調群と低同調群の間で、対話の活性度および協調性のいずれにおいても有意な差は確認されなかった ($p > 0.05$)。対話全体を通して、対話者同士が互いに「同じ長さの間」を取り合おうとすることは、対話の質の向上には直結しないことを示唆している。

DTW 距離（時系列的な変化パターンの類似度）に基づき群分けを行った結果（図 4）、高同調群は低同調群と比較して、対話の質において統計的に有意な優位性を示した。

具体的には、対話の活性度 (Activity) について、高同調群の中央値は 151.50 であったのに対し、低同調群は 129.00 にとどまり、有意な差が確認された ($p < 0.05$, $r = 0.34$)。同様に、対話の協調性 (Cooperativity) についても、高同調群の発話量の差の中央値は

表 2: 分析対象データ別の同調手法に基づく対話指標比較

対象データ	評価手法	指標	高同調群 Median (IQR)	低同調群 Median (IQR)	Sig.
Gap	収束	Activity	136.00 (47.75)	148.50 (94.25)	n.s.
		Cooperativity	0.24 (0.19)	0.20 (0.26)	n.s.
	同期	Activity	143.00 (80.75)	137.00 (59.25)	n.s.
		Cooperativity	0.18 (0.24)	0.29 (0.25)	n.s.
Pause	収束	Activity	153.00 (71.50)	140.50 (34.75)	n.s.
		Cooperativity	0.20 (0.22)	0.19 (0.22)	n.s.
	同期	Activity	165.50 (107.25)	140.00 (34.50)	*
		Cooperativity	0.19 (0.19)	0.22 (0.23)	n.s.
Gap・Pause 統合	収束	Activity	140.00 (45.00)	147.50 (101.50)	n.s.
		Cooperativity	0.30 (0.14)	0.15 (0.28)	n.s.
	同期	Activity	151.50 (99.75)	129.00 (43.00)	*
		Cooperativity	0.19 (0.19)	0.33 (0.26)	*

(* : $p < 0.05$, n.s.: not significant)

0.19, 低同調群は 0.33 であり, 高同期群の方が有意に値が小さい, すなわち特定の話者に偏らず対等な発話交換が行われていることが示された ($p < 0.05, r = 0.31$).

これらの全体統計の結果は, 収束と比較した際の同期の有意性を統計的に支持するものである. 効果量 r はいずれも中程度の水準を示しており, 発話潜時の「値の差」ではなく「変化のパターン」が共有されている対話においては, 発話が活発に行われ, かつ双方が対等に関与する傾向が強いことが明らかになった.

3.5 システム実装への適応

本分析の結論として, 質の高い対話を実現するために, 対話システムは単に「人間らしい平均的な間」を固定的に設定したり, 直前の Gap のみに反応するだけでは不十分と言える. 「変化パターンへの同期」および「Gap と Pause の統合的な効果」という分析結果に基づき, システムに求められる要件は以下の 2 点に集約される.

1. **対話中の動的な変化パターンへの追従**: ユーザの潜時の絶対値そのものではなく, その時間的な変化の傾向を対話中に検知し, 対話システム側もその変化に合わせて応答タイミングを決定すること.
2. **対話中の沈黙の包括的な制御**: 対話システムは「発話交替時間 (Gap)」のタイミングのみを制御するのではなく, 「発話内沈黙 (Pause)」もテン

ポの一部として認識し, ユーザの発話潜時を含めた包括的な時間制御を行うこと. つまり, ユーザの発話終了からの応答時間だけでなく, システムが継続発話を行う際の潜時も制御することで, 対話全体の「間」を同期させる制御が重要となる.

これらの知見に基づき, ユーザの変化するテンポに対して, Gap と Pause の区別なく包括的に同期する制御モデルを構築することで, ユーザとの親和性を高め, 対話の活性化が期待できる. 次章では, この指針を取り入れた音声対話システムを実装する.

4 実験用音声対話システムの実装

前章の分析により, 質の高い対話においては, 静的な平均値への収束ではなく, 対話の文脈に応じた動的な潜時変化への同期が重要であることが示された. しかし, これを実際の音声対話システムで実現するには, 大規模言語モデルの返答生成遅延やネットワーク遅延といった物理的な制約を克服し, 数十ミリ秒単位で発話タイミングを制御する必要がある. 本章では, これらの課題を解決し, ユーザの潜時に対話の中で同調可能な対話システムの設計と実装について述べる.

4.1 音声対話コーパス分析に基づく設計要件

コーパス分析を通して得られた知見に基づき, 本実験システムには以下の 3 つの設計要件が求められる.

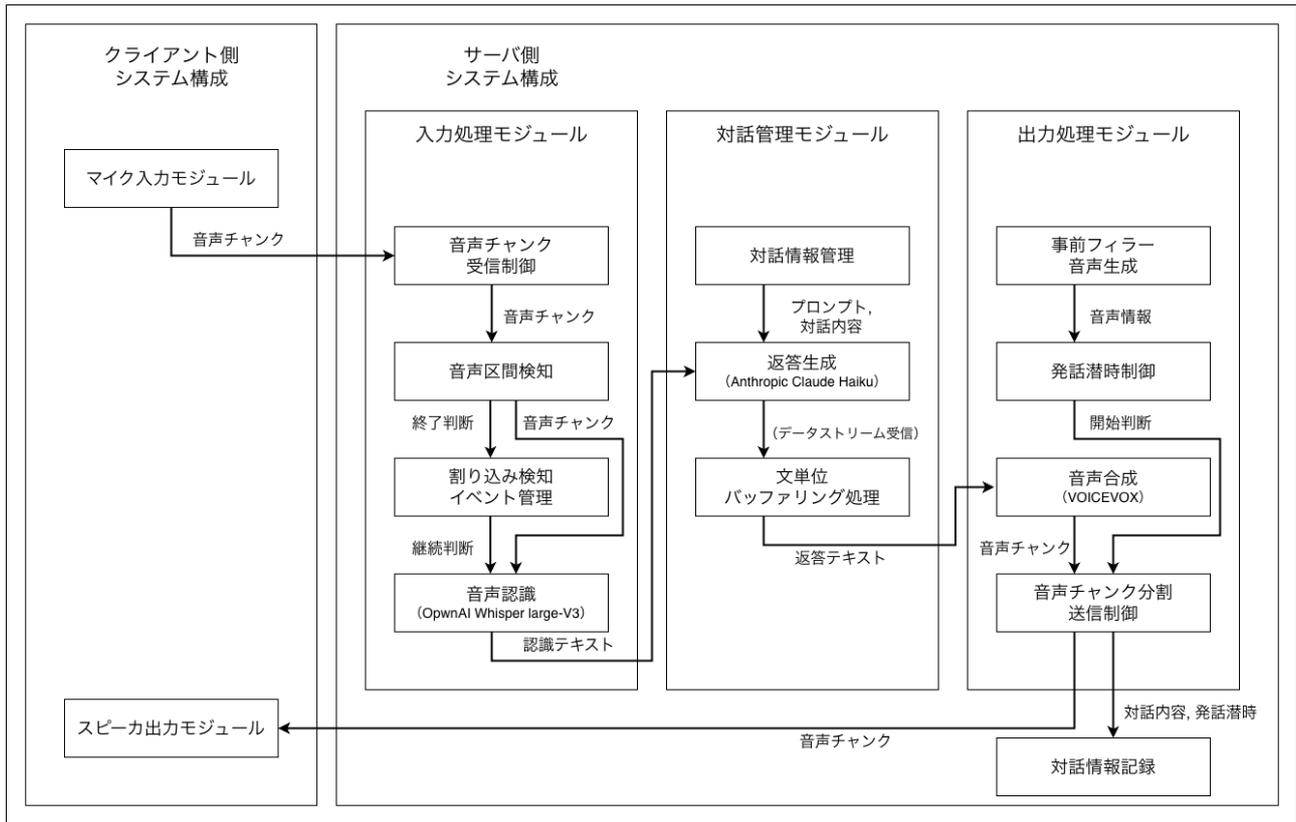


図 5: 実験用音声対話システムの構成図

【要件 1】 ユーザ潜時の対話中の計測: 同調の基準となるユーザのテンポ状態を把握するために、ユーザの発話開始・終了タイミングおよび発話区間情報を対話中に検知し、潜時を定量的に計測する機能が必要である。

【要件 2】 システム発話における Gap と Pause の統合制御: 「統合的なテンポ」を再現するためには、システムは単に応答を開始するまでの時間 (Gap) を調整するだけでなく、発話内容に含まれる間 (Pause) の長さも意図的に制御できなければならない。

【要件 3】 制御目標に対する低遅延性: 同調条件において、算出された目標潜時通りに発話を開始するためには、システムの処理遅延による意図しない遅れを排除しなければならない。特に、速いテンポへの同調が求められる場面において、生成処理が間に合わず発話が遅れる事態を防ぐため、極めて低いレイテンシでの動作が不可欠である。

4.2 システム構成と実装詳細

上述の要件を満たすため、本研究では図 5 に示す実験用音声対話システムを開発した。システムはクライ

アント・サーバモデルを採用し、Python の `asyncio` ライブラリを用いた非同期処理によって実装することで、音声ストリームの送受信と並列して推論処理を行うイベント駆動型アーキテクチャを実現した。各コンポーネントの実装詳細は以下の通りである。

4.2.1 音声区間検出による潜時計測

要件 1 (ユーザ潜時の計測) に対応するため、音声認識の前段に DNN ベースの音声区間検出エンジンである Silero VAD を配置した。本モジュールは、入力音声のパワーと確率分布に基づき、ユーザの発話終了を約 0.3 秒の閾値で高速に判定する。システムはこの判定時刻を基準としてユーザの直前の発話潜時を確定させ、次項の制御ロジックへ即座に伝達する役割を担う。

4.2.2 ストリーミング生成による低遅延化

要件 3 (正確な実時間動作) における課題である「生成遅延」を解消するため、対話生成エンジンには Anthropic 社の Claude 3 Haiku API を採用し、ストリーミング生成モードを実装した。従来のターンアラウンド方式 (全文生成後に合成) では数秒単位の遅延が不

可避であり、速いテンポでの「間に合わせる」制御が不可能となる。対して本システムでは、文単位（句読点区切り）で生成が完了したトークンから順次、並列的に音声合成プロセスへ引き渡すパイプライン処理を実装した。これにより、First Token の出力までの時間を最小化し、システムが目標とする Gap での発話開始を物理的に可能とする時間的余白を確保した。

4.2.3 音声合成と出力タイミングの制御

要件 2 (Gap と Pause の制御) および要件 3 の実現として、音声合成エンジンには VOICEVOX を採用し、ローカル環境にて稼働させた。本システムでは、計測されたユーザ潜時に基づいて決定された「目標 Gap」と「目標 Pause」に従い、以下の 2 点の制御を行う。

- **Gap 制御**：ストリーミング生成された音声データの再生開始時刻をスケジュールすることで、意図したタイミングでの応答開始を実現する。
- **Pause 制御**：発話テキスト内の発話単位に対応する無音区間の長さをパラメータとして操作し、発話内部のテンポを調整する。

目標潜時が短く、応答生成が間に合わない場合には、発話終了検知後から目標時間が経過した時点で、事前生成されたフィラー（「はい」「えーっと」などの返答音声）を再生する制御を行った。これにより、システム性能のばらつきに左右されず、制御を十ミリ秒単位の精度で安定して実行することを可能とし、設計通りの「間」を厳密に提示することが可能となった。

5 同調を行う潜時制御の評価実験

本章では、前章までに構築した「Gap と Pause を統合的に制御可能な音声対話システム」を用い、被験者実験を通して発話潜時の動的制御が対話に与える影響を検証する。コーパス分析では、人同士のコミュニケーションにおいて、発話潜時の変化が同調した対話では、そうでない対話と比較して、対話の活性や効率に影響を与えることが示唆された。しかし、同調が対話の質を向上させる要因なのか、質の高い対話者同士において同調する現象が副次的に観察されたのかは明らかではない。そこで、「固定的な応答」と「平均値への収束」「変化への同期」を比較することで、ユーザの認知プロセスや対話体験にどのような因果的效果をもたらすかを明らかにする。

5.1 検証の目的と検証項目

本実験の主たる目的は、システムによる発話潜時の制御手法の違いが、ユーザの記憶効率、生理的集中度、および社会的印象に与える影響を定量的に評価することである。先行研究および本研究でのコーパス分析の知見に基づき、以下の 3 つの検証項目 (Evaluation Questions: EQ) を設定した。

EQ-1: 記憶効率への効果：潜時による同調は、ワーキングメモリへの干渉を抑制し、情報の記憶効率を向上させるか

ユーザの情報処理速度や発話生成に伴う認知プロセスの変動に即したタイミングで情報が提示された場合、情報の符号化に必要な認知リソースが確保されやすいと考えられる [2]。平均的なペースを合わせる「収束」よりも、その瞬間の状態に合わせる「同期」の方が、提示情報の記憶定着率を向上させるかを確認する。

EQ-2: 生理的集中の維持：潜時変化への同期は、ユーザの認知負荷を低減し、対話タスクへの生理的な集中状態を促進するか

自身の内部テンポと乖離した外部刺激は認知的な干渉を引き起こすとされる [14]。システムがユーザの動的な潜時変化に同調することで、高い時間的整合性が保たれ、結果としてユーザが対話内容へ没入し、生理的な集中状態が高水準で維持されるかを検証する。

EQ-3: 社会的価値の創出：変化への同調は、システムに対する「知性」や「丁寧さ」といった社会的評価を向上させるか

相手の思考の間や迷いといった微細な機微を汲み取り、それに合わせて間を調整する振る舞いは、コミュニケーション適応理論において相手への配慮として機能する [1]。対話の中の状態に合わせる「同期」において、システムが最も肯定的かつ知的な社会的印象を獲得するかを調査する。

5.2 実験タスクと条件

5.2.1 情報収集対話タスク

実験タスクには、ユーザの事前知識によるバイアスを排除し、対話を通じた純粋な情報獲得量を評価するため、「未知の郷土料理に関する情報収集タスク」を採用した。ユーザは、「マンティ」や「インジェラ」といった海外の馴染みのない郷土料理の名前が記載されたリストを提示され、システムに対して能動的に質問を行うことで、その料理の特徴、材料、調理方法などの情



図 6: 実験参加者の環境の様子

報を収集する。本タスクでは、筆記行為による注意の分散を防ぎ、対話そのものへの没入度を評価するため、対話中のメモ等の外部記録行為を一切禁止した。ユーザは5分間の対話セッション終了後、対話のみから得られた情報を記憶を頼りに記述することが求められる。

実際の対話の様子および対話後の記述結果の一例を、それぞれ図6と、表3,4に示す。

5.2.2 比較条件の設定

発話潜時制御の効果を検証するため、以下の3条件を設定した。なお、全ての条件において、システムはGapだけでなくPauseも統合して制御を行う。

固定潜時条件 (Fixed Condition) 対照群として、ユーザの潜時に関わらずシステムの発話潜時を常に一定値 300ms に固定する条件である。人間同士の会話における平均的な話者交替潜時は約 200ms であることが知られているが [15]、実際の音声対話システムにおいては、発話終了判定 (Voice Activity Detection: VAD) のために一定の無音区間を監視する必要がある。本システムでは、この判定に必要な最小限の-marginとして 300ms を設定している。したがって、この値は、意図的な遅延操作を行わず、VAD の処理時間を考慮した上で技術的に実現可能な最速の応答を採用した時間設定である。

収束条件 (Convergence Condition) 対話者間の「平均的なテンポ」が揃っていく現象を再現する条件である。直近3発話のユーザ発話潜時の加重移動平均を算出し、以下の式(3)に基づきシステムの発話潜時 $L_{sys}^{(n)}$ を決定する。これにより、値(Value)の近接性は高まるが、局所的な変化への

追従は行われない。

$$L_{sys}^{(n)} = \frac{3L_{user}^{(n-1)} + 2L_{user}^{(n-2)} + 1L_{user}^{(n-3)}}{6} \quad (3)$$

同期条件 (Synchrony Condition) ユーザの「変化パターン」への即応を重視した条件である。以下の式(4)を用い、ユーザが直前のターンで潜時を長くした(考え込んだ)場合はシステムも遅らせ、短くした(即答した)場合は早く返すという、変化量に基づく追従制御を行う。

$$L_{sys}^{(n)} = L_{sys}^{(n-1)} + (L_{user}^{(n-1)} - L_{user}^{(n-2)}) \quad (4)$$

5.3 評価指標

上述のEQを検証するため、以下の3つの側面から多角的な評価を行った。

5.3.1 客観評価：記憶定着率

EQ-1(記憶効率)を検証するため、対話終了直後に実施した「内容記述課題」のスコアを用いた。本評価では、対話によって提示された情報の定着度を定量化するため、対話文および被験者の記述文に含まれる単語の中から、文意の構成に不可欠な実質的意味を持つ語を「意味語」と定義し、その出現数を計測した。意味語の抽出処理には日本語形態素解析ライブラリ Janome を用い、解析された形態素のうち名詞、動詞、形容詞、および副詞を抽出対象とした。その際、頻出するが文脈固有の情報を持たない形式名詞(例:「こと」、「もの」)や、補助動詞(例:「～している」の「いる」)などの非自立語は、情報要素としての重要性が低いため除外処理を適用した。算出された意味語の総数を用い、システムの発話に含まれていた情報の何割が、被験者の記述によって再現されたかを以下の式により算出した(式5)。

$$Score_{recall} = \frac{N_{recalled}}{N_{presented}} \quad (5)$$

ここで $N_{presented}$ は対話タスク中にシステムが発話した文章に含まれる意味語の総数であり、 $N_{recalled}$ は被験者の記述文章に含まれる意味語の数である。なお、実験条件(応答潜時の制御)によって対話のテンポや総ターン数が変動し、結果としてシステムからの提示情報量($N_{presented}$)自体に差異が生じる可能性がある。情報の絶対量の多寡が記憶成績に与える影響を排除するため、本指標では提示総数に対する割合($Score_{recall}$)を採用することで、条件間の公平な比較を可能とした。

表 3: 実験における対話例

話者	発話内容
ユーザ	マンティはどここの国の料理ですか？
システム	マンティは中央アジア諸国，特にウズベキスタンやカザフスタンを中心とした地域で食べられる伝統的な料理です。
システム	蒸し餃子に似た形をしており，羊肉や野菜などを詰めた生地を蒸して作ります。
ユーザ	へえ，マンティの調理方法を教えてください。
システム	まず，小麦粉と水，塩で生地を作り，その中に挽き肉や玉ねぎ，ニンニクなどを詰めます。
システム	生地の形を整えて蒸し上げるのが基本的な作り方です。
ユーザ	次に，インジェラに合うおすすめ飲み物食べ物をお願いします
システム	まず，飲み物としてはトゥッラというホットの発酵茶が良く合います。
システム	ほのかな酸味と甘みが特徴で，インジェラの味わいと相性が良いです。
⋮	⋮

表 4: 対話後の記述例

記述内容
マンティは羊の肉を使った料理。カザフスタンやウズベキスタンの料理である。小麦粉、水、塩胡椒などを用いる。インジェラに合うのは温かい飲み物である。

5.3.2 生体評価：心拍変動 (HRV)

EQ-2 (生理的集中) を検証するため，ウェアラブル心拍センサ (WHS-3[16]) を用いて対話中の心拍変動 (Heart Rate Variability: HRV) を計測した。HRV は自律神経系の活動を非侵襲的に推定する指標として広く用いられており，特に認知的なりソースの投下量やメンタルワークロードの変動を鋭敏に反映することが知られている。本実験では，対話システムの応答潜時の違いが，ユーザの生理的な覚醒水準や集中維持に与える影響を検証することを目的とした。評価指標の算出にあたっては，短時間の自律神経活動の動的な変化を捉えるため，60 秒間の移動窓を設定し，以下の 3 つの指標を算出した。

- **HR (Heart Rate):** 1 分間あたりの心拍数であり，生理的な覚醒水準を反映する指標である。認知的な負荷や情動的な興奮が高まると交感神経活動が優位となり，心拍数は上昇する傾向にある。
- **RMSSD (Root Mean Square of the Successive Differences):** 隣接する RRI の差分の二乗平均平方根であり，副交感神経活動の活性度を反映する指標である。精神的負荷が高い状態や集中状態においては副交感神経活動が抑制され，RMSSD の値は低下することが知られている。本研究では，対話への没入や集中を示す主要な指標として用いる。

- **CVRR (Coefficient of Variation of R-R intervals):** RRI の標準偏差を平均値で除した変動係数であり，自律神経活動の総量を表す指標である。一般に，精神的な緊張状態では低下し，リラックスした状態では上昇する傾向がある。

分析においては，これらの指標の変化が単なる「不快感に伴うストレス」によるものか，あるいはタスクへの能動的な関与に伴う「ポジティブな集中」によるものかを判別する必要がある。そのために，内容記述課題のスコアや主観評価との相関を考慮して結果の解釈を行う。この多角的なアプローチにより，対話システムの発話潜時制御がユーザの認知状態および精神的負荷に与える影響を包括的に評価する。

5.3.3 主観評価：印象アンケート

EQ-3 (社会的価値) を検証するため，各セッション終了後に主観評価アンケートを実施した。評価項目は，システムの社会的受容性や満足度を問う「(a) 全体的な印象」，音声対話のターン交代の適切さを問う「(b) 発話の印象」，および対話の流暢性や負荷感を問う「(c) 会話の印象」の 3 つのカテゴリに分類した。特に本研究の主眼である「間 (発話潜時)」の適切さを直接的に問う項目 (「システムは話しやすい『間』であったか」等) に加え，潜時制御が対話の「自然さ」や「スムーズさ」といった高次な印象に波及するかを検証す

表 5: 各条件における対話統計量および効率性指標の平均値 (Mean ± SD)

項目	固定潜時 (F)	収束 (C)	同期 (S)
提示単語数 ($N_{dialogue}$)	440.8 ± 53.6	358.2 ± 44.2***	395.3 ± 47.9***
記述単語数 (N_{desc})	44.1 ± 26.7	43.1 ± 18.7	48.8 ± 22.5
記憶定着率 ($N_{desc}/N_{dialogue}$)	0.108 ± 0.063	0.121 ± 0.051	0.126 ± 0.054*

(* $p < .05$, *** $p < .001$ vs cond. F)

る項目を設けている。回答形式として、定量的評価が可能な多くの項目については、原則として「1: 全くそう思わない」から「7: とてもそう思う」までの7段階のLikert尺度を採用した。これにより、潜時制御の違いがユーザの主観的な対話体験に与える心理的影響を明らかにする。

6 結果

6.1 対話データの概要と処理

本実験では、計35名（男性15名、女性20名、平均年齢 $M = 22.0$ 歳、 $SD = 2.2$ ）の参加者からデータを収集した。参加者の要件として、全員が日本語を母語とし、視覚・聴覚の障害および心疾患の既往歴がないことを事前に確認している。サンプルサイズは、予備実験の結果から効果量を中程度 (Cohen's $f = 0.30$) と仮定し、3条件の対応あり分散分析を想定してG*Powerを用いた事前分析 ($\alpha = 0.05, 1 - \beta = 0.80$) に基づいて決定された。分析に先立ち、データの整合性および実験手続きの妥当性を検証した結果、以下の基準に基づき一部のデータを除外した。まず、内容記述課題の分析については、実験教示に対する理解の齟齬により、適切な手順で記述が行われなかった2名分のデータを除外した ($N = 33$)。また、心拍変動解析については、実験中のセンサ装着不備により、解析に十分な精度の生体信号が取得できなかった1名分のデータを除外した ($N = 34$)。

データの統計的解析にあたっては、有意水準を5%に設定した。データの正規性が保証されない群が含まれていたため、3条件間の比較にはFriedman検定を用い、事後検定としてShaffer法による多重性補正を行ったWilcoxonの符号付き順位検定を用いた。さらに、サンプルサイズに依存するp値のみの報告による誤解を避けるため、実質的な差の大きさを表す効果量として r 値 ($r = |Z|/\sqrt{N}$) を算出し、Cohenの基準 (0.1: Small, 0.3: Medium, 0.5: Large) に基づき評価を行った。

6.2 対話統計量と記憶効率の分析

内容記述課題の統計量および効率性指標の結果 (表5) からは、発話潜時の同期制御が情報の提示量を抑制しつつも、情報の定着効率を有意に向上させるという傾向が明らかになった。

まず、提示単語数について比較すると、同期条件は固定潜時条件よりも有意に少ない値を示した ($p < 0.001, r = 0.53$)。これは、同期条件ではユーザの沈黙に合わせてシステムも待機するため、5分間という限られた時間内での総ターン数が減少したこと起因すると考えられる。一方で、対話後に記述された「内容記述単語数」の絶対量については、条件間に有意な差は見られなかった ($p > 0.05$)。すなわち、同期条件では提示された情報量が少なかったにも関わらず、固定潜時条件と同等の絶対量の情報を獲得できていたことを意味する。

この結果を「学習効率」の観点から検証するため、対話量に対する記述の比率 (記述単語数 / 提示単語数) について分析を行った。結果として、同期条件の優位性が統計的に裏付けられた。オムニバス検定により条件間の有意差が確認され ($p < 0.05$)、多重比較の結果、同期条件は固定潜時条件と比較して有意に高い比率を示した ($p < 0.05, r = 0.32$)。一部には分母 (提示量) の減少が比率上昇の主因であるとの見方も成立しうるが、重要な点は分子 (記述量) が低下していない事実である。この結果は、同期制御によって生じた「間」が情報の提示密度を適正化し、情報の歩留まり (定着効率) が向上したとする検証項目EQ-1を強く支持するものである。

6.3 対話中の集中度の比較

対話タスク中の参加者の生理的覚醒度および集中状態を評価するため、心拍変動 (HRV) 指標を用いた分析を行った。対話全体 (0-300秒) の平均値および、ベースラインからの変化量を用いて比較を行った。各指標の基本統計量および有意差検定の結果を表6に、時系列推移を図7に示す。

表 6: 各条件における生理指標の基本統計量 (Mean \pm SD) と有意差検定の結果

指標・フェーズ	固定潜時 (F)	収束 (C)	同期 (S)
Heart Rate (bpm)			
Total (0-300s)	77.56 \pm 2.08	77.14 \pm 1.97	78.15 \pm 2.15**
P1 (0-50s)	78.16 \pm 1.46	77.78 \pm 1.07	78.69 \pm 1.43*
P2 (50-100s)	76.73 \pm 0.82	77.27 \pm 0.98*	78.08 \pm 0.91**
P3 (100-150s)	77.18 \pm 0.89	76.93 \pm 1.02	78.18 \pm 0.99**
P4 (150-200s)	77.90 \pm 0.88	77.01 \pm 1.06	77.77 \pm 1.03
P5 (200-250s)	77.87 \pm 0.83	77.13 \pm 0.93	77.94 \pm 0.86
P6 (250-300s)	77.52 \pm 0.91	76.74 \pm 0.81	78.24 \pm 1.09#
RMSSD (ms)			
Total (0-300s)	32.87 \pm 5.07	32.27 \pm 5.95*	30.73 \pm 4.37
P1 (0-50s)	33.17 \pm 2.47	33.07 \pm 3.06	30.71 \pm 2.71
P2 (50-100s)	35.44 \pm 2.68	33.37 \pm 2.99	31.50 \pm 1.92
P3 (100-150s)	35.66 \pm 2.51	34.28 \pm 2.78	30.79 \pm 2.23
P4 (150-200s)	30.33 \pm 2.45	31.41 \pm 2.89	30.22 \pm 2.38
P5 (200-250s)	31.11 \pm 2.20	31.09 \pm 1.87	30.28 \pm 2.35
P6 (250-300s)	31.49 \pm 1.89	30.35 \pm 2.67	30.89 \pm 2.31
CVRR (%)			
Total (0-300s)	6.76 \pm 1.65	6.37 \pm 1.53	6.37 \pm 1.40
P1 (0-50s)	7.09 \pm 1.15	6.42 \pm 0.94	6.67 \pm 1.41##
P2 (50-100s)	6.80 \pm 0.71	6.41 \pm 0.90	6.05 \pm 0.58#
P3 (100-150s)	7.37 \pm 0.78	7.08 \pm 0.83	6.22 \pm 0.67
P4 (150-200s)	6.19 \pm 0.83	6.37 \pm 0.81	6.33 \pm 0.72
P5 (200-250s)	6.49 \pm 0.67	6.09 \pm 0.67	6.47 \pm 0.64
P6 (250-300s)	6.60 \pm 0.63	5.82 \pm 0.78**	6.46 \pm 0.78

(* $p < 0.05$, ** $p < 0.01$ vs cond. F; # $p < 0.05$, ## $p < 0.01$ vs cond. C)

Time Series Trends (Baseline: -180s ~ -60s)

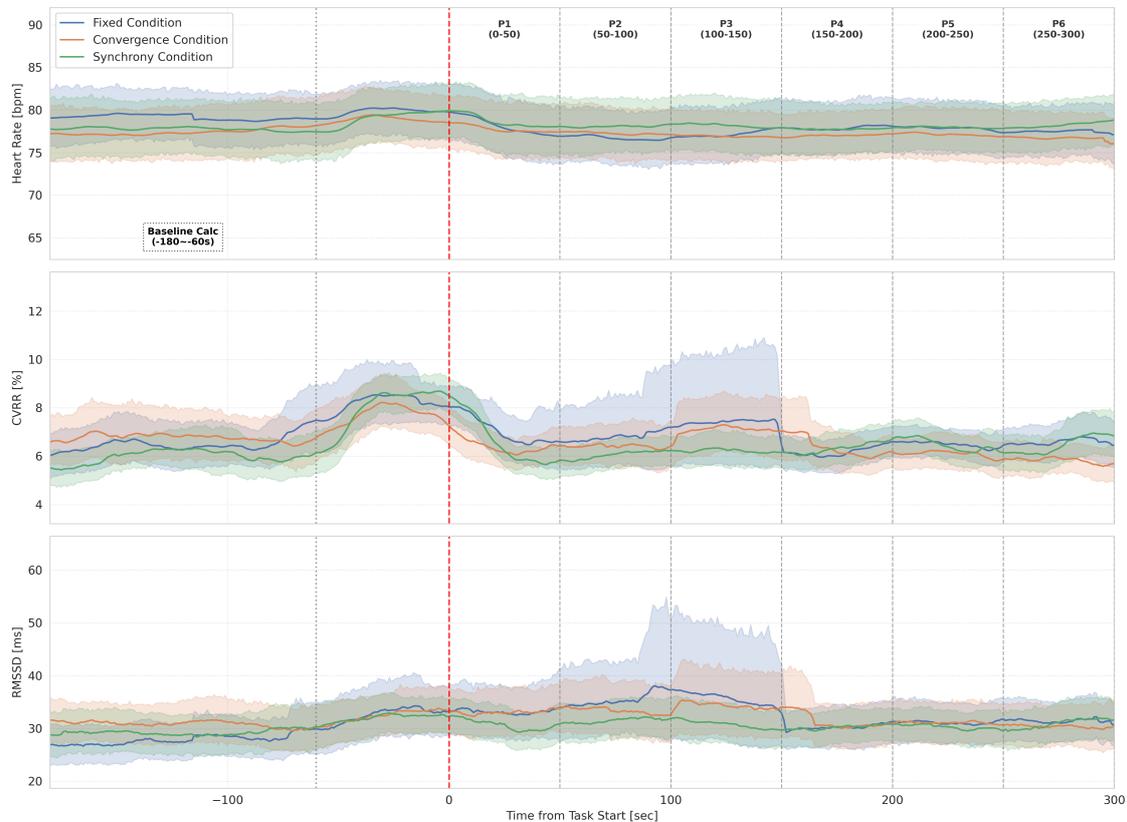


図 7: 対話中の心拍データの時系列推移

6.3.1 心拍データの前処理

取得された R-R 間隔 (RRI) には、体動等のノイズやパケットロスによる欠損、不整脈等の生理的な外れ値が含まれる。データの信頼性を担保するため、以下の前処理を適用した。

まず、線形補間法により 1Hz の等間隔時系列へリサンプリングし、隣接変化率が 20% を超える区間を異常値として除去・再補間した。次いで、個体差排除のためベースライン補正を行った。直前の準備動作や会話による生理状態への干渉を考慮し、タスク開始直前の 60 秒間を除外した上で、開始 180 秒前から 60 秒前までの 120 秒間をベースライン区間とした。各指標の算出においては、このベースライン区間の平均値を基準値とし、タスク中の計測値から減じた変化量を用いることで、タスク実施に伴う生理反応の純粋な変動分を抽出した。

6.3.2 対話全体における生理指標の比較

まず、副交感神経活動を反映しリラックス状態の指標となる RMSSD に着目する。固定潜時条件が 32.87ms と最も高く、次いで収束条件が 32.27ms、同期条件が 30.73ms と最も低い値を示した。一般に RMSSD の低下は、メンタルワークロードの上昇や集中状態（交感神経の優位）を示唆する。変化量を用いた検定の結果、固定潜時条件と収束条件の間に有意な差が確認され ($p < 0.05, r = 0.39$)、同期条件においても他の条件よりベースラインからの低下量が大きい傾向が見られた。

次に、平均心拍数 (HR) に着目すると、同期条件が 78.15bpm と最も高い値を示した。変化量を用いた検定の結果、固定潜時条件と同期条件の間に有意な差が確認された ($p < 0.01, r = 0.49$)。これは、固定潜時条件では心拍数の上昇が緩やかであったのに対し、同期条件では心拍数が有意に上昇し、ユーザが高い覚醒水準で対話タスクに取り組んでいたことを示唆している。

情報収集という能動的なタスクの性質を考慮すると、固定潜時条件では低い覚醒状態に近かったのに対し、同期条件では適度な緊張感を持ち、積極的にタスクへ没入・集中していたと解釈できる。これにより、EQ-2 は支持された。

6.3.3 対話フェーズごとの時系列詳細分析

対話の進行に伴う生理状態の変化を詳細に捉えるため、対話時間全体を 6 つのフェーズ (P1~P6) に分割し分析を行った。特に注目すべきは以下の 3 点である。

1. **対話序盤 (P1: 0-50 秒)** : CVRR において、同期条件は収束条件と比較して有意に値が高かった

($p < 0.05, r = 0.43$)。これは、対話開始直後から自律神経活動が活発に変化し、同期への適応が始まっていたことを示す。

2. **対話中盤 (P2: 50-100 秒)** : このフェーズにおいて、心拍数 (HR) の変化量 (Delta) が、固定潜時条件において有意に大きかった (vs 収束: $p < 0.05, r = 0.45$; vs 同期: $p < 0.01, r = 0.53$)。これは、固定潜時条件にて、より生理反応が不安定であったことを示している。また、同期条件では、この区間における心拍数の維持率が高く、対話が軌道に乗る段階において、ユーザの対話への集中が高まっていた可能性が示唆される。
3. **対話終盤 (P6: 250-300 秒)** : 対話セッションの終了間際において、生理指標に条件間差が認められた。CVRR の変動係数 (CV) については、固定潜時条件と収束条件の間に有意差 ($p < 0.01, r = 0.59$) が確認され、条件間で自律神経活動の安定性に差異が生じたことが示された。また、心拍数 (HR) の平均値においては、収束条件 ($M = 76.74$) と同期条件 ($M = 78.24$) の間に有意差 ($p < 0.05, r = 0.36$) が見られた。この結果は、同期条件において収束条件よりも有意に高い心拍水準が維持されたことを示しており、タスク終了まで覚醒度や集中力が途切れなかった可能性が示唆される。

6.4 印象評価アンケートの結果

アンケート結果 (図 8) の分析から、評価構造は「機能的な流暢性」と「社会的印象」の 2 側面に大別される傾向が確認された。

まず、対話の機能面に関する評価では、固定潜時条件が最も優れたパフォーマンスを示した。「スムーズさ」(vs 同期: $p < 0.05, r = 0.38$) や「間の適切さ」(vs 収束: $p < 0.05, r = 0.39$) などの項目で他条件を有意に上回り、最短応答 (300ms) による流暢性の高さが評価された。

一方、対話の社会面・対人面に関する評価では異なる傾向が見られた。「丁寧さ」において、同期条件は収束条件と比較して有意に高い評価を得た ($p < 0.05, r = 0.39$)。特筆すべきは、「総合印象」や「丁寧さ」といった社会的指標において、同期条件が固定潜時条件と同等、あるいはそれ以上の評価 (総合印象平均値: 5.38 で最大) を獲得している点である。

この結果は、同期条件が「スムーズさ」などの機能面で固定潜時条件に劣るにも関わらず、総合的な好印象では上回る傾向にあることを示している。一般に、対話システムにおいて流暢性の欠如は印象低下に直結す

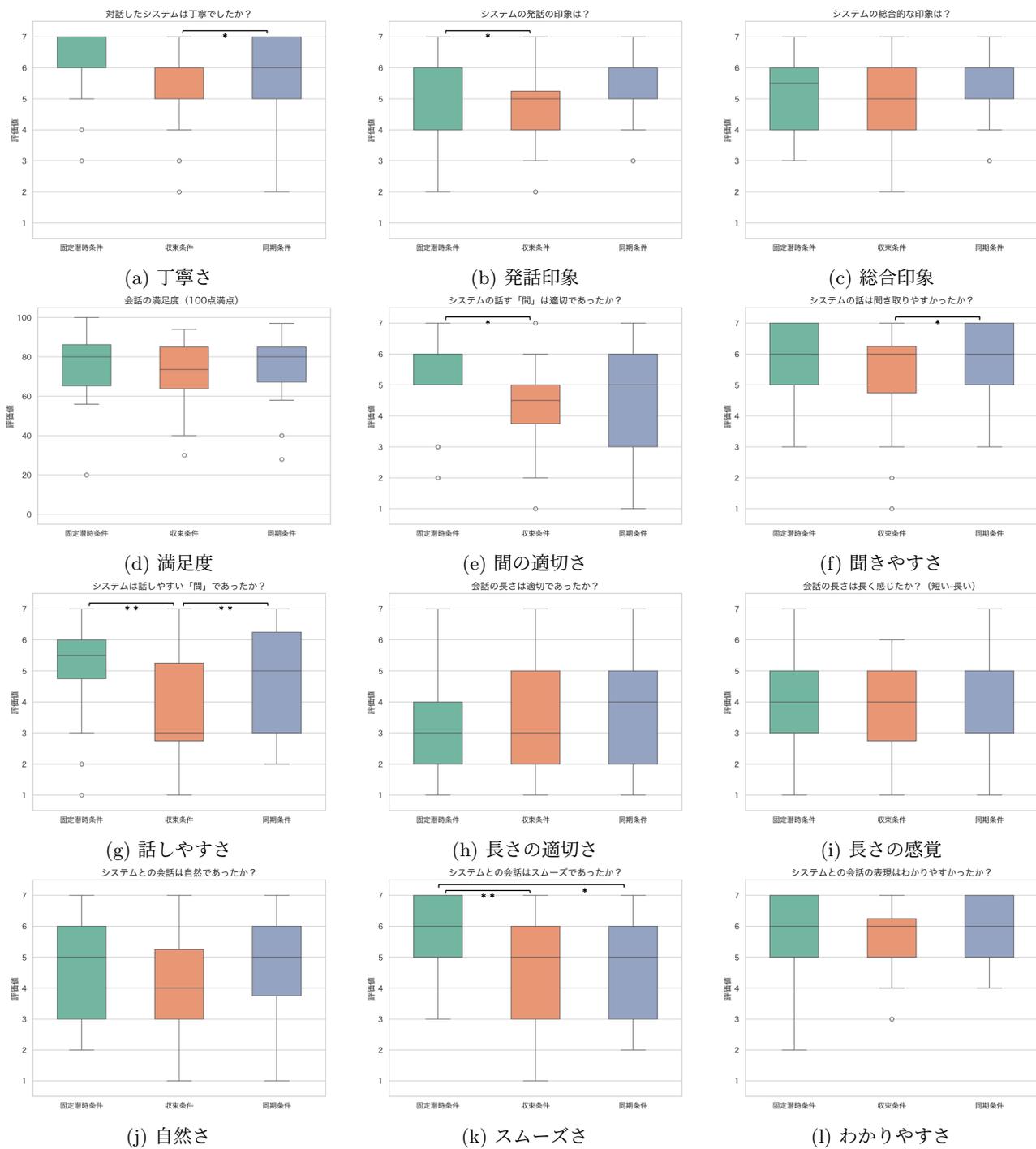


図 8: 印象評価アンケートの結果一覧 (* $p < 0.05$, ** $p < 0.01$)

ると考えられるが、本実験の結果はこの通念とは異なる。すなわち、ユーザは単なる応答の流暢さ以上に、自身のテンポへの同期から感じられる「丁寧さ」や「配慮」といった質的価値を重視しており、EQ-3は支持されたと言える。

7 考察

本章では、評価実験の結果に基づき、発話潜時の動的制御がユーザの認知プロセス、生理的状态、および対話体験に与える影響について多角的に考察する。特に、同期制御がもたらした「認知的猶予」の効果と、収束制御の限界要因、そして社会的インタラクションとしての「間」の機能について詳細に論じる。

7.1 認知的猶予の提供による学習効率の最大化

内容記述課題の結果は、対話システムにおける情報提示のタイミング制御が、ユーザの知識獲得プロセスに対して決定的な役割を果たすことを示唆している。本実験における最大の知見は、発話潜時の同期制御が記憶される情報の「絶対量（記述単語数）」そのものを増やすのではなく、提示された情報に対する「定着率（記述/提示比率）」を有意に向上させた点にある ($p < 0.05$)。

固定潜時条件では、システムが常に300msという極めて短い潜時に即答するため、単位時間あたりの情報提示密度が高まる。Swellerらの認知負荷理論[17]によれば、人間のワーキングメモリの容量と持続時間には厳格な限界が存在する。外部からの連続的な刺激を長期記憶へと定着させるためには、情報をチャンク化し構造化する内的な処理過程が不可欠であり、それには一定の時間的猶予が必要となる。しかし、固定潜時条件のような「間」のない対話は、ユーザ自身のペースでの処理を許容しないシステム主導の構造となる。ユーザが直前の情報を脳内で統合している最中に次の情報が強制的に提示されるため、情報の分断や文脈復帰のための外在的認知負荷が増大する。この結果、情報処理要求がユーザの認知容量を超過する「認知的オーバーフロー」の状態を招き、対話内容の深い理解が阻害されていた可能性が高い。

対照的に、同期条件では、ユーザが思考のために間を置いた際、システムも同様に沈黙することで、ユーザに対して情報の咀嚼に必要な「認知的猶予」を自然な形で提供している。この意図的な空白が、情報の提示ペースをユーザの処理能力に動的に適合させ、結果として記憶効率を最大化したと考えられる。すなわち、「いかに多くの情報を与えるか」よりも「いかにユーザ

の情報処理を待つか」という設計視点が、実質的な知識伝達において重要であることが実証された。

以上のことから、検証項目EQ-1（記憶効率への効果）に対し、肯定的な結論が得られた。

7.2 生理的覚醒と生産的な集中状態の共起

生理指標の解析結果において、同期条件では他条件と比較して心拍数（HR）の有意な上昇が確認された。一般に、対話中の心拍上昇は、システムの不具合や会話の不成立に対する「不快感に伴うストレス」反応と、課題に対する「集中・没入」反応の双方で生じうるため、生理指標単独での解釈は困難である。

しかし、本実験におけるパフォーマンス指標（内容記述課題の結果）を併せて分析することで、この心拍上昇の質的な解釈が可能となる。もし心拍上昇がシステムへの不満や焦りによるネガティブなストレスに起因するものであれば、認知資源は感情調整などのタスク外処理に浪費され、記憶パフォーマンスは低下するはずである。ところが実際の結果は、同期条件において「記憶定着率」が有意に向上しており、高い覚醒度と高いパフォーマンスが共存している。

この事実は、同期条件においてユーザが感じた「間」や「緊張感」が、不快なものではなく、情報を処理するための生産的な集中状態を誘発していたことを強く示唆する。システムがユーザの思考に合わせて沈黙することで生じた「場」の緊張感が、ユーザの意識を対話内容に強く向けさせ、結果として深い情報処理を促したと考えられる。したがって、発話潜時の同期制御がユーザのタスクに対する集中を高めるという検証項目EQ-2（生理的集中の維持）は支持された結論付けられる。

7.3 ユースケースに合わせた制御手法の選択

実験結果は、対話システムの利用目的に応じて「固定潜時制御」と「同期制御」を使い分けるべきであるという、設計上の指針を示している。アンケート分析の結果は、対話の「流暢さ」と「社会的印象」が必ずしも正の相関関係にないという興味深い知見をもたらした。

まず「スムーズさ」の項目において、同期条件は固定潜時条件よりも有意に低い評価 ($p < 0.05$) を受けている。従来の対話システム評価指標に照らせば、スムーズさが損なわれることは、ユーザ満足度を低下させる要因となる。しかしながら、本実験の「丁寧さ」や「総合印象」のスコアにおいて、同期条件は固定潜時条件と同等の高評価を維持しており、統計的な有意差こそないものの、記述統計的には全条件中で最も高い平

均値 (Mean= 5.38) での総合評価を示した。この「スムーズではないが、印象は良い」という結果は、同期制御によって生じた「間」や「応答の遅れ」が、単なるシステムのエラーやラグとしてではなく、相手の状態に配慮した「丁寧な振る舞い」という社会的印象としてポジティブに受容されたことを意味する。

同期条件は、ユーザが早く話せばシステムも即答し、ユーザが考えながら遅く話せばシステムも間を置くという、動的な相互作用を生み出す。これにより、即座に想起できた内容にはテンポよく返答がなされ、思考を要する複雑な内容には十分な間を持って返答されるという、会話文脈に即した自然な構造が形成される。アンケートの自由記述において、システムの発話内潜時の長発話に対して「システムの発話自体もゆっくりに感じた」という報告があった点は注目に値する。実際にはシステムの発話速度自体は全条件で一定であったにも関わらず、潜時の調整が対話全体のテンポとして知覚され、システムの発話全体がユーザの状態に同調しているという印象を与えたと推察される。

コミュニケーション適応理論 (CAT) [1] が示唆するように、相手のペース (潜時) に合わせるという同調行動は、「私はあなたの状態に関心を持っています」という相手への関心や好意的な態度を表明する非言語的な表現として機能する。本実験においてユーザは、機能的な話しやすさが若干損なわれたとしても、システムが自分に合わせてくれているという感覚に対し、より高い社会的価値を見出したと解釈できる。教育、カウンセリング、案内といった、効率性以上に信頼関係が重視されるドメインにおいては、この「社会的価値」の創出こそが重要となる。

以上のことから、潜時制御による同調行動が社会的価値を高めるという検証項目 EQ-3 (社会的価値の創出) は、統計的傾向を含めて支持された。

7.4 収束条件における同調精度の限界と要因

本実験において、収束条件は、同期条件ほど顕著な効果を示さなかった。この要因として、人間同士の自然対話と本実験のタスク特性の乖離、および制御アルゴリズムの性質的な限界が挙げられる。人間同士の自然対話では、潜時の平均値は徐々に収束する傾向にあるが、本実験のような「情報検索型タスク」では、ユーザが回答を思い出そうとして長考する場合がある。Heldnerらの研究においても、認知負荷のかかるタスク対話では、自然な雑談よりも長い Gap 時間が計測される現象が確認されている [18]。

収束条件のアルゴリズムでは、この「思考のための長考」もテンポの一部として平均値計算に取り込んでしまう。そのため、ユーザが思考を終えて話し出した次

のターンで、システムが文脈上不必要に長い潜時 (数秒間) をとってしまう現象が観察された。事後の自由記述アンケートにおいて、比較的長い発話潜時に対して「ゆっくり話しているのか、システムのエラーで止まっているのか判断できず不安だった」との記述が見られたことから、この不自然さが確認できる。「思考状態」と「発話テンポ」を区別せず、単純な移動平均で追従しようとした結果、不必要な沈黙が生成され、対話の円滑さを阻害してしまったことが収束条件の限界であったと言える。

またこの結果は、インタラクションにおける「予測可能性」と「随伴性」の欠如によっても解釈できる。固定潜時条件は、機械的ではあるが常に一定であるため予測可能であり、ユーザはそれに合わせて発話を計画できる。一方、同期条件は、予測は難しいものの、ユーザの直前の振る舞い (沈黙) に対して即座に反応するため、自分の行動に対する高い随伴性を持つ。これらに対し、収束条件は過去の平均値に基づくため、現在の文脈とは乖離したタイミングで応答され、予測もできず自分の直前の行動とも整合しない。この二重の欠落が、ユーザに対して「タイミングが合わない」という強い違和感を抱かせ、対話の流暢性を阻害したと結論付けられる。

7.5 音声対話における同調の構成要素

本研究の実験結果は、人間と対話システムが円滑かつ効果的に相互作用するために、同調がいかなる機能を果たしているかを浮き彫りにした。音声対話における同調現象の本質的要素として、以下の3点が導出される。

1. 認知的リソースとしての「間」の機能

従来の対話システム研究において、応答潜時は単なる「処理遅延」として扱われ、その最小化が至上命題とされてきた。しかし、本研究の結果は、適切な「間」がユーザの情報処理を促進し、記憶定着率を有意に向上させる「認知的猶予」として機能することを示している。ユーザが思考を深化させる局面において、過度な即時性は認知容量の飽和を招く要因となり得る。したがって、対話における潜時は、単なる待ち時間ではなく、対話の理解・学習を最大化するための動的なリソースとして再定義されるべきである。

2. 収束に対する同期の優位性

人間の発話テンポは文脈に応じて刻々と変化する動的なものである。本実験において、過去の平均値に回帰させる「収束」アプローチよりも、直近の変化に追従する「同期」アプローチが有効

であったという事実は、対話における随伴性の重要性を裏付けている。ユーザが求めているのは、統計的な安定性ではなく、自身の直前の振る舞い（迷いや加速など）に対する即応的なフィードバックである。このことから、人間らしい対話テンポの構築には、大域的な平均化処理よりも、局所的な変動への鋭敏な適応能力が不可欠であると言える。

3. 社会的統合としての時間制御

発話潜時の調整は、話者交替という構造的な役割だけでなく、対話相手への態度や社会的属性を伝達する非言語情報として機能する。相手のペースに自身の振る舞いを同調させる過程が、傾聴や配慮といった社会的意味を内包しており、ラポール形成の基盤となる。教育、医療、カウンセリングなど、信頼関係が重視されるドメインにおいて、この時間的振る舞いは、生成される発話内容の流暢さと同等以上に、インタラクションの質的評価を決定づける本質的な要素である。

7.6 限界と今後の課題

本研究の限界と今後の展望として、以下の2点が挙げられる。

第一に、複合的な情報による状態推定の精緻化である。本実験では発話潜時に着目した制御を行ったが、対話における全体的なテンポ感は、発話速度、基本周波数の変動、音量変化などの複合的な要素によって形成される。今後は、潜時制御に加え、これらの韻律特徴を多角的に捉え、相互の関係性を考慮しながら同調するモデルへの拡張が有効であると考えられる。異なるパラ言語要素が補完し合いながら対話のテンポを形成する制御が実現できれば、対話システムの自然性をさらに高めることに寄与すると期待される。

第二に、文化的背景の限定性である。本実験の被験者は主に日本の大学生であり、得られた知見には日本の文化的・言語的コンテキストが反映されている可能性がある。特に、沈黙に対する許容度や、会話の重複に対する選好が異なる文化圏において、「同期」が同様の効果をもたらすか、あるいは文化に応じた要素の調整が必要かを検証が必要である。

しかしながら、本研究が実証した「動的な潜時制御が人間の認知機能（記憶・集中）に直接介入できる」という事実は、従来の「平均への収束」に留まっていた対話システム研究のパラダイムを拡張するものであり、今後のHCI研究における重要な基盤となると確信する。

8 おわりに

本研究では、人間と音声対話システムとの共生を見据え、対話における非言語的な時間構造、特に「発話潜時」の動的な制御がユーザの認知プロセスおよび対話体験に与える影響を明らかにした。

従来の対話システム研究において、同調は主にユーザの平均的な特徴量への「収束」として実装されてきた。しかし、音声対話コーパスの分析を通じ、質の高い対話においては、値の近接ではなく、対話の文脈や思考プロセスに応じて刻々と変化する潜時の揺らぎへの「同期」の有意性を明らかにした。この知見に基づき、話者交替潜時（Gap）と発話内潜時（Pause）を統合的に扱い、ユーザの潜時変化に対して対話の中で同調する対話システムを構築した。

被験者実験による比較検証の結果、「同期」条件は、対話内容の記憶定着率を有意に向上させ、かつユーザの高い集中状態（心拍変動指標による評価）を維持することが実証された。これは、システムがユーザの思考テンポに合わせて適切な認知的猶予を提供することで、情報の符号化に必要な処理時間が確保され、ワーキングメモリへの過度な負荷が抑制されたことに起因すると結論付けられる。また、主観評価においては、同期制御による応答の遅延が、単なる処理ラグではなく、相手への「配慮」や「丁寧さ」といった社会的印象として肯定的に受容されることが確認された。

本研究の学術的貢献は、対話システムにおける「間」のデザインが、単なる流暢さや好感度の向上に留まらず、ユーザの認知リソースを適切に管理し、学習や理解を支援する機能的役割を果たせることを実証した点にある。この成果は、教育支援、認知トレーニング、意思決定支援など、ユーザの深い思考と関与が求められる対話インタフェースの設計において、ユーザに適応した韻律制御の指針を提供するものである。

今後は、本研究で扱った時間情報に加え、発話速度や基本周波数の変動などを複合的に評価することで、より高精度にユーザの状態を推定し、相互の関係性を考慮しながら同調するモデルへと拡張していく。また、ハイコンテキスト文化圏で得られた本知見の普遍性を検証するため、異なる文化的背景を持つ話者間での比較研究を展開し、より適応的で包括的な対話制御モデルの確立を目指す。

謝辞

本研究は、森泰吉郎記念研究振興基金の助成を受けて実施されたものである。本研究では、国立情報学研究所 音声資源コンソーシアムから提供を受けた「RWCP

音声対話データベース - 96 年版 (RWCP-SP96)」を利用した。

参考文献

- [1] Howard Giles, Justine Coupland, and Nikolas Coupland, editors. *Contexts of Accommodation: Developments in Applied Sociolinguistics*. Editions de la Maison des Sciences de l'Homme and Cambridge University Press, 1991.
- [2] Martin Pickering and Simon Garrod. Toward a mechanistic psychology of dialogue. *The Behavioral and brain sciences*, Vol. 27, pp. 169–90; discussion 190, 05 2004.
- [3] Rivka Levitan, Stefan Benus, Ramiro Gálvez, Agustín Gravano, Florencia Savoretti, Marián Trnka, Andreas Weise, and Julia Hirschberg. Implementing acoustic-prosodic entrainment in a conversational avatar. pp. 1166–1170, 09 2016.
- [4] Nichola Lubold, Erin Walker, and Heather Pon-Barry. Effects of voice-adaptation and social dialogue on perceptions of a robotic learning companion. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 255–262, 2016.
- [5] Camille J. Wynn and Stephanie A. Borrie. Classifying conversational entrainment of speech behavior: An expanded framework and review. *Journal of Phonetics*, Vol. 94, p. 101173, 2022.
- [6] Rivka Levitan and Julia Hirschberg. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Proceedings of Interspeech 2011*, pp. 3081–3084, 2011.
- [7] Jennifer S. Pardo, Kelly Jordan, Rolliene Mallari, Caitlin Scanlon, and Eva Lewandowski. Phonetic convergence in shadowed speech: The relation between acoustic and perceptual measures. *Journal of Memory and Language*, Vol. 69, No. 3, pp. 183–195, 2013.
- [8] Rimita Lahiri, Md Nasir, Manoj Kumar, So Hyun Kim, Somer Bishop, Catherine Lord, and Shrikanth Narayanan. Interpersonal synchrony across vocal and lexical modalities in interactions involving children with autism spectrum disorder. *JASA Express Letters*, Vol. 2, No. 9, p. 095202, 09 2022.
- [9] Albert Rilliard, Alexandre Allauzen, and Philippe Boula de Mareüil. Using dynamic time warping to compute prosodic similarity measures. In *Interspeech 2011*, pp. 2021–2024, 2011.
- [10] Jens Edlund, Mattias Heldner, and Julia Hirschberg. Pause and gap length in face-to-face interaction. pp. 2779–2782, 09 2009.
- [11] Emanuel A. Schegloff. Overlapping talk and the organization of turn-taking for conversation. *Language in Society*, Vol. 29, No. 1, pp. 1–63, 2000.
- [12] Real World Computing Partnership. Rwcsp spoken dialogue corpus – 1996 edition (rwcp-sp96), 2006. Dataset.
- [13] Stephen Levinson and Francisco Torreira. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, Vol. 6, , 06 2015.
- [14] 一川誠. 大人の時間はなぜ短いのか. 2008.
- [15] Paul T. Brady. A technique for investigating on-off patterns of speech. *The Bell System Technical Journal*, Vol. 44, No. 1, pp. 1–22, 1965.
- [16] 心拍センサ WHS-3 — ユニオンツール株式会社. <https://www.uniontool.co.jp/product/sensor/whs3.html>.
- [17] John Sweller, Jeroen J. G. Van Merriënboer, and Fred Paas. Cognitive architecture and instructional design. *Educational Psychology Review*, Vol. 10, pp. 251–296, 09 1998.
- [18] Mattias Heldner and Jens Edlund. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, Vol. 38, No. 4, pp. 555–568, 2010.