修士論文

音声対話システムにおける適応的なマルチモーダル感情分析のための
ストリーム型能動学習

阿慈地 惇人

主指導教員 岡田 将吾

北陸先端科学技術大学院大学
先端科学技術研究科
（情報科学）

令和 8 年 3 月

Abstract

Accurately recognizing a dialogue user ’ s emotional state and adapting system responses accordingly is essential for realizing empathetic dialogue systems [1]. Emotional states are expressed not only through linguistic content but also via nonverbal behaviors such as vocal prosody, facial expressions, and body movements. Therefore, integrating multimodal sentiment analysis (MSA) models that jointly observe linguistic and nonverbal signals is a key challenge in dialogue system design. However, emotional expression patterns and intensities vary substantially across individuals, making it difficult for generic models to capture user-specific characteristics.

One straightforward and reliable approach to addressing this issue is to directly ask users about their emotional states during dialogue. By querying emotions in situ, highly reliable labels can be obtained while mitigating memory decay and eliminating the need for offline annotation. Nevertheless, frequent emotion queries disrupt the natural flow of interaction and impose additional burdens on users. Thus, to balance sentiment estimation accuracy with user experience, it is crucial to strategically determine when to ask emotion labels.

In this study, we formulate personalized multimodal sentiment analysis as a stream-based active learning problem, where data are observed sequentially and the system must decide online whether each sample is worth querying. We propose a framework in which the system determines, at each dialogue turn, whether to request an emotion label based on sequentially observed multimodal features, including speech, language, and visual cues. A reinforcement learning–based label acquisition strategy is designed such that requesting labels for incorrectly predicted samples yields positive rewards, whereas unnecessary queries for correctly predicted samples yield negative rewards. This formulation enables the system to efficiently acquire informative labels under a limited query budget.

The proposed method quantifies uncertainty using confidence estimates from each modality and integrates them through adaptive modality weighting to guide label request decisions. Furthermore, uncertainty thresholds and modality weights are dynamically updated via reinforcement learning, allowing the system to adapt to user-specific emotional expression tendencies over time. Unlike conventional active learning methods with fixed criteria, the proposed approach continuously refines its querying policy as the dialogue progresses.

Simulation experiments using human–agent dialogue corpora demonstrate that the proposed method improves sentiment estimation performance more efficiently than random sampling and conventional uncertainty-based active learning strategies, even under few-shot conditions. These results indicate that the proposed approach effectively achieves accurate personalized multimodal sentiment analysis

while reducing user burden, highlighting its potential as a foundational technology for the practical deployment of empathetic dialogue systems.

目 次

図 目 次

表 目 次

第1章　はじめに

共感的な対話システムには，ユーザの感情状態を適切に把握し，それに応じて応答を調整する能力が求められる [1]．感情は発話内容といった言語的手がかりだけでなく，表情や韻律といった非言語的手がかりによっても伝達される．対話ユーザの感情状態を示すこれらの社会的信号を検出することで，システムはユーザの感情をより正確に認識し，共感的な応答を生成し，豊かな対話体験を提供できる [2]．

しかし，感情の表出パターンは個人によって大きく異なる．例えば，どのモダリティで感情が強く表出されるか，あるいはその強度は人によって異なることが知られている [3, 4, 5]．そのため，全ユーザを一様に扱う一般モデルには推定精度に限界が存在する．先行研究 [6, 7] でも，感情分析において一般モデルよりも個人適応モデルの方が高い精度を示すことが報告されている．

対話システムにおいては，図 1.1 のようにユーザの感情状態を直接尋ねることが，ユーザ固有のマルチモーダル行動と感情との対応関係を把握する上で最も単純かつ有効な方法である．対話中にその場で尋ねることで，記憶減衰の影響を抑えつつ高信頼なラベルを得ることができ，記録やログを後から注釈する必要もなくなるため，ユーザにとって効率的である．

一方で，過度な質問はユーザ体験の低下を招く可能性がある [8]．感情について繰り返し尋ねられると，ユーザは対話を煩わしく，不自然に感じ，システムへのエンゲージメントや利用意欲が低下する恐れがある．そのため，精度向上とユーザ負担の低減とのバランスを取るために，どのタイミングでユーザに感情ラベルを尋ねるべきかを戦略的に判断する仕組みが重要となる．

この課題に対して，有望な解決策として能動学習が挙げられる．能動学習は，情報量の高いサンプルのみに選択的にラベル付けを行うことで，ラベル付けのコストを抑えつつモデル性能を維持できる [9]．特に，サンプルが逐次観測され，その場でラベル取得の要否の決定が必要となる対話シナリオでは，プール型能動学習よりも，ストリーム型能動学習の方が適している．さらに，対話を通じた効率的な知識獲得に関する近年の研究 [10] では，ストリーム型能動学習の枠組みの中で強化学習を用いて「いつ尋ねるか」を最適化する手法が提案され，その有効性が示されている．

しかし，既存研究では，プール型能動学習や単一モダリティを前提とした手法が主であり，逐次観測されるマルチモーダルデータに対して個人適応を目指した能動学習手法はほとんど検討されていない．そこで本研究では，このギャップを埋めるために，個人適応型マルチモーダル感情推定を目的としたストリーム型能動

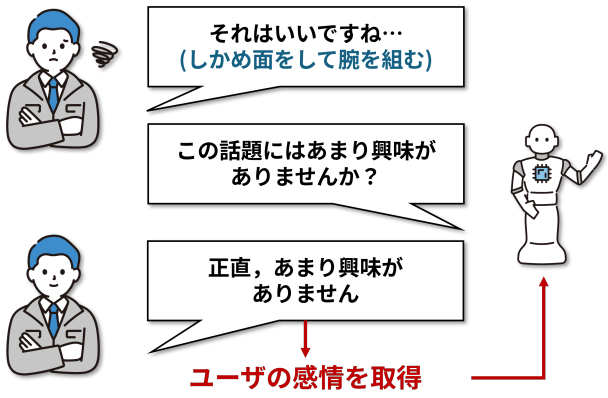


図 1.1: 対話における感情ラベル取得の例.

学習フレームワークを，強化学習に基づいて提案する.

本研究の貢献は以下の 3 点にまとめられる.

- 人と AI の相互作用において観察される個人ごとの感情表出の多様性に対応するため，個人適応的マルチモーダル感情分析をストリーム型能動学習問題として定式化した.
- 逐次観測されるマルチモーダル行動に対してラベル要求の可否を決定する強化学習フレームワークを提案した.
- 2 つの人—エージェント対話コーパスを用いた実験により，提案手法が few-shot 条件下で感情推定性能を向上させることを示した.

第2章　関連研究

2.1　対話を通じた知識獲得

対話システムにおける重要な機能の一つに，ユーザとの相互作用を通じて必要な知識を獲得する能力が挙げられる．固定的な知識に依存する既存モデル，例えば大規模言語モデルは，新たに出現する語彙や地域特有の表現，さらにはユーザ固有の嗜好や感情的ニュアンスを十分にカバーできない場合がある [11, 12]．この限界を解消するため，ユーザとの対話を通じて知識を補完・更新する枠組みを探究する研究が増えている．

先行研究では，ユーザ発話から新たな語彙や事実知識を獲得する手法 [13, 14]，ユーザ満足度や嗜好を推定する手法 [15]，さらにはロボットが対話を通して新しい物体カテゴリや空間概念を学習する手法 [16, 17, 18] など，多様な観点から知識獲得が検討されてきた．これらの研究は総じて，固定のデータセットのみに依存するのではなく，実際の相互作用を通じて環境やユーザに適応していく重要性を示している．

2.2　対話システムにおける効率的な知識獲得のための能動学習アプローチ

章 2.1 では，対話を通じてユーザの嗜好や語彙など，新たな知識を獲得する重要性を示したが，過度な質問はユーザ体験を損なう可能性がある [8]．そのため，システムは有益な情報を効率的に取得しつつ，質問回数を最小限に抑える設計が求められる．この課題に対処するため，近年の研究では，ユーザの知識や嗜好の獲得を能動学習の枠組みで捉え，システムが「いつ尋ねるべきか」を学習する手法が提案されている [10]．これらのアプローチは，対話システムが長期的な報酬を考慮しつつ，柔軟な質問戦略を獲得できる可能性を示している．

2.3　感情認識における能動学習

能動学習（Active Learning，AL）は，ラベルなしデータの中から情報量の高いサンプルに対してのみ選択的にラベルを要求することで，学習効率を向上させる

枠組みである [9]．AL には主に 2 つの代表的な設定が存在する．1 つはプール型設定であり，あらかじめ用意されたラベルなしデータ集合を保持し，その中からラベル付けすべきサンプルを選択する方法である．もう 1 つはストリーム型設定であり，データが逐次的に観測される状況を想定し，各サンプルを観測したタイミングで直ちにラベル要求の必要性を判断する方法である．

AL は，感情認識においてもラベル付けのコスト削減を目的として幅広く適用されてきた．Li らは，情報量とモダリティ間整合性を統合的に利用するプール型 AL 手法 GRACE を提案し，限られたラベル付きデータであっても高い性能を維持できることを示した [19]．Abdelwahab と Busso は音声感情認識に AL を適用し，ごく少量のラベル付きサンプルであっても性能向上が可能であることを報告している [20]．さらに近年では，Moreno-Acevedo らが情報量と多様性の双方を同時に考慮するストリーム型 AL 手法を提案し，より少ないラベルで高精度を達成できることを実証した [21]．加えて，Karnjanapatchara らはマルチタスク学習とアノテータ間の一致率モデリングを統合することで，逐次的なラベル獲得プロセスにおいても信頼性の高い学習を実現している [22]．

しかし，これらの既存研究の多くはプール型設定を前提としているか，あるいは単一モダリティを対象としていることが多く，逐次観測されるマルチモーダルデータを即時に扱う手法については十分に検討されていないのが現状である．また，個人ごとに異なる感情表出の特徴に適応する個人適応の重要性が先行研究において繰り返し指摘されている一方で [7]，能動学習の枠組み自体に個人適応を明示的に組み込んだ研究は極めて限定的である．

これらのギャップに着目し，本研究では，逐次観測される各マルチモーダルデータが，感情認識における即時的な個人適応に有益かどうかを判断するストリーム型能動学習フレームワークを提案する．本手法は，オンライン環境下での適応性を重視しつつ，限られたラベル取得の機会を最大限に活かすことを目的として設計されている．

第3章　定式化

本研究の目的は，対象ユーザとの対話の初期段階においてできる限り有益な特徴—ラベル対を取得し，その情報を対話後半における感情推定性能の向上に活用することである．この枠組みにおいて最も重要となる課題は，限られた問い合わせ予算のもとで，どのタイミングでラベルを取得すべきかを適切に判断するためのラベル要求方策を学習する点にある．特に，過剰なラベル要求はユーザ体験を損なう可能性があるため，できる限り少ない問い合わせ回数で最大限の性能向上を実現する手法設計が求められる．

本研究では，対話を離散的な時間ステップ $t = 1, 2, \ldots, T$ の列としてモデル化する．各ステップにおいて，システムはマルチモーダル特徴ベクトル $x_t \in \mathbb{R}^d$ を観測し，その時点に対応する真の感情ラベルを $y_t \in \mathcal{Y}$ と定義する．一方，システムが推定したラベルを $\hat{y}_t$ と表す．ここで重要なのは，真のラベル y_t は，システムがユーザに対して明示的にラベル要求を行った場合にのみ取得可能であり，それ以外の場合にはアクセスできないという点である．

ラベル要求を行うか否かの判断は，方策 π_ϕ によって決定され，この方策はパラメータ ϕ によって特徴づけられる．各ステップにおける報酬 r_t は，以下のように定義される．すなわち，ラベル要求を行うことによって誤推定を未然に防ぐことに成功した場合には正の報酬 ρ^+ を与え，逆に，ラベル要求前の時点で正しい推定をできていた場合には負の報酬 ρ^- を与える．また，ラベル要求を行わなかった場合には報酬を 0 とする．このような報酬設計に基づくと，方策学習は期待報酬の最大化問題として定式化され，限られた問い合わせ回数の制約下で最も有益なラベルのみを選択的に取得するための最適な方策を習得することが本問題の中心となる．

$$\max_{\pi_\theta} \ \mathbb{E}_{\pi_\theta}\left[\sum_{t=1}^{T} r_t\right]. \tag{3.1}$$

第4章　提案手法

本研究では，対話中に観測されるデータを逐次処理しながらユーザ固有の感情表出特性に適応するために，マルチモーダル感情分析（MSA）を対象としたストリーム型能動学習手法を新たに提案する．提案手法では，Reinforced Active Learning（RAL）[23] に基づく方策学習フレームワークを採用し，この枠組みをマルチモーダル入力に拡張するために，不確実性推定をマルチモーダル特徴に基づいて行う仕組みを導入している．これにより，我々はMSA に特化した強化学習ベースの能動学習手法である RAL-MSA を構築した．

RAL-MSA の目的は，対象ユーザとの対話の初期段階において，モデルの性能改善に特に寄与する情報量の高い特徴–ラベル対を効率的に取得し，その取得済みの個人固有データを活用して対話後半における感情推定性能を向上させることである．このような逐次的な推論とラベル取得の統合は，個人の感情表現が多様で時間的にも刻々と変化しうる実際の対話状況において，ユーザ適応をリアルタイムで実現する上で極めて重要である．また，対話を中断してユーザにラベルを尋ねる行為は，頻度が高い場合にユーザの没入感や体験品質を損なう可能性があるため，限られた問い合わせ回数の中で最大限の性能向上を達成するための戦略的なラベル要求方策の学習が不可欠となる．

提案手法の全体的な構成と情報の流れを示した概要図を 図 4.1 ，アルゴリズムをアルゴリズム 1 に示す．本図におけるオラクルとはラベル付けを要求したサンプル x_t のラベル y_t を提供する関数である．本図に示されるように，RAL-MSA は逐次的に観測されるマルチモーダル特徴に基づき，不確実性および予測状況を評価しながらラベル要求の有無を判断する方策を学習する構造を備えており，この枠組みによってリアルタイムかつ個人適応的な感情推定が可能となる．

4.1　学習手順

RAL-MSA の全体的な学習プロセスは，以下に示す 4 つの段階から構成される．これらの各段階は相互に密接に関連しており，逐次的に進行することで，ターゲットユーザの個別的な特徴に適応した感情推定モデルの構築を可能にする．

1. **事前学習：**まず初めに，学習に用いるマルチモーダル分類器およびラベル要求ポリシーを，事前学習用ユーザデータを基盤として初期化する．この段階では，複数のモダリティから得られる特徴分布を大まかに捉え，基本的な分類

Algorithm 1 RAL-MSA algorithm

Input:
x_t (sample), C (per-modality learners),
w_t (modality weights), θ_t (certainty threshold),
ϵ (ϵ-greedy threshold), η (learning rate)
Init:
$M \leftarrow \{\text{audio}, \text{text}, \text{visual}\}$,
$decisions \leftarrow \{\}$

for t=1...T **do**
    for $m \in M$ **do**
        $decisions[m] \leftarrow C[m].askCertainty(x_t^m) < \theta_t$
    end for
    $committeeDecision \leftarrow \mathbb{I}[\sum_{m \in M} w_t[m] decisions[m] \geq 0.5]$
    $p \leftarrow U(0, 1)$
    if $p < \epsilon$ **or** $committeeDecision = 1$ **then**
        $y_t \leftarrow acquireLabel(x_t)$
    end if
    if $committeeDecision = 1$ **then**
        $r_t \leftarrow getReward(x_t, y_t)$
        $w_{t+1} \leftarrow updateWeights(r_t, w_t, decisions, committeeDecision, \eta)$
        $\theta_{t+1} \leftarrow updateThreshold(r_t, \theta_t, \eta)$
    end if
end for

function UPDATEWEIGHTS($r_t, w_t, decisions, committeeDecision, \eta$)
    for $m \in M$ **do**
        if $decisions[m] = committeeDecision$ **then**
            $w_{t+1}[m] \leftarrow w_t[m] \exp(\eta \times r_t)$
        end if
    end for
    $w_{t+1} \leftarrow w_{t+1} / \sum_{m \in M} w_{t+1}[m]$
    return w_{t+1}
end function

function GETREWARD(x_t, y_t)
    if $\hat{y}_t = y_t$ **then**
        return ρ^-
    else
        return ρ^+
    end if
end function

function UPDATETHRESHOLD(r_t, θ_t, η)
    $\theta_{t+1} \leftarrow \min\{\theta_t(1 + \eta \times (1 - 2^{\frac{r_t}{\rho^-}})), 1\}$
    return θ_{t+1}
end function

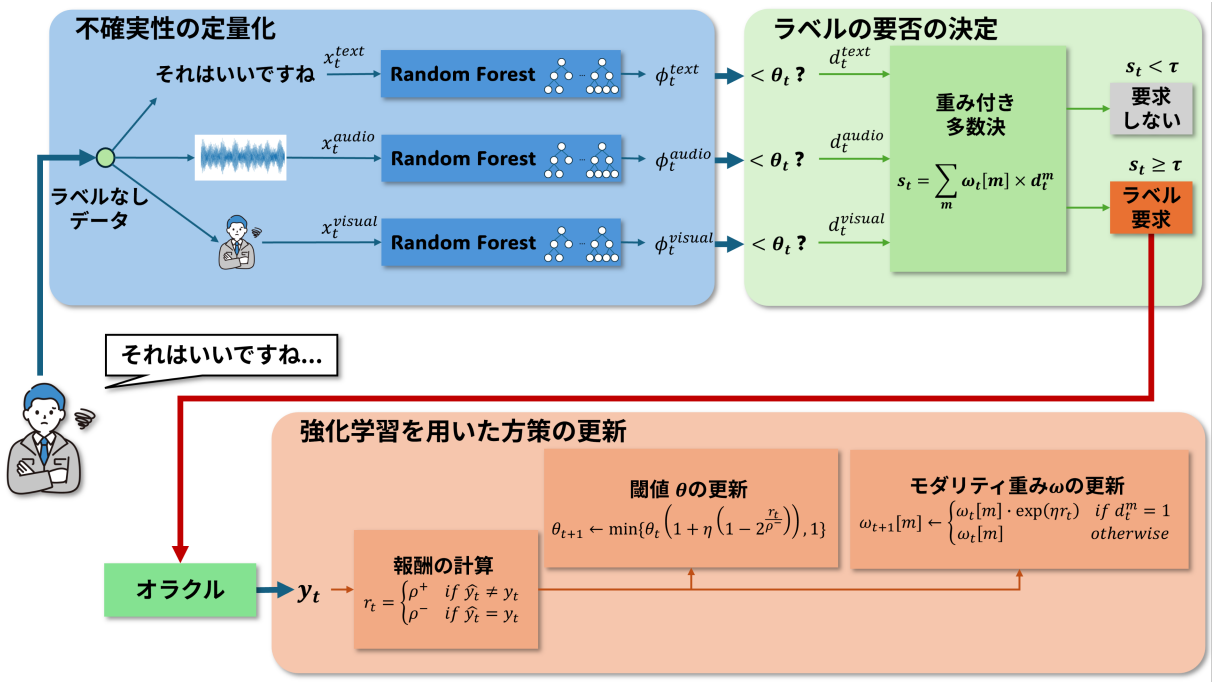


図 4.1: RAL-MSA の概要.

能力や初期の方策の挙動を確立することが目的となる．これにより，オンライン適応に入る前の初期性能を保証しつつ，後続の段階でユーザ固有のデータに効率的に適応できる基盤を整える．

2. **オンライン適応とラベル要求：**次に，ターゲットユーザから取得されるデータを 1 サンプルずつ逐次的に処理する．各時点において，ポリシーに基づきラベル要求の要否を判断し，要求する場合には実際にユーザから正解ラベルを取得する．その後，得られたラベルを用いて報酬を計算し，その報酬に従ってポリシーパラメータを適応的に更新する．このプロセスによって，システムはユーザの感情表出の傾向や誤推定が生じやすい状況を逐次的に学習し，リアルタイムで個人適応を進めることが可能となる．

3. **再学習：**新たに獲得されたラベル付きサンプル数が所定の閾値に達した時点で，それらのサンプルを学習用データプールに追加し，分類器の再学習を実施する．これにより，ターゲットユーザ特有の特徴分布が分類器に反映され，推定精度の向上が期待される．再学習後，再び「オンライン適応とラベル要求」ステージに戻り，ポリシー更新とラベル要求を継続する．この反復サイクルは，事前に設定されたラベル要求予算が尽きるまで継続される．

以上の一連の手順により，RAL-MSA は対話の進行に合わせてモデルと方策を段階的に改善し，限られたラベル要求回数のもとでも効率的かつ高性能な個人適応的感情推定を実現する．

4.2 提案手法の概要

RAL-MSA フレームワークは，基本的には Wassermann らによって提案された RAL [23] の枠組みに基づいて構築されている．しかし，本研究では，この既存枠組みを単に踏襲するだけではなく，マルチモーダル感情分析（MSA）の文脈に適用するための新たな機能拡張を行っている．具体的には，音声・言語・映像といった異種のモダリティにまたがる特徴量に基づいて不確実性を定量化する新たなコンポーネントを導入することで，従来の単一モダリティに依存した不確実性推定では捉えきれない複合的な不確実性を精緻に捉えられるように設計している．この拡張により，ユーザの感情表出が複数のモダリティに渡って多様に現れるという MSA 特有の困難さに対して，より柔軟かつ頑健に適応可能なラベル要求方策の学習が可能となる．

以上の設計方針に基づき，本研究で提案するラベル要求方策は，以下の三つの主要な構成要素を通じて学習される．すなわち，(a) 不確実性の定量化，(b) ラベル要否の決定，および (c) 強化学習を用いた方策の更新である．それぞれの構成要素は密接に連携し，相補的に機能することで，ユーザへの不要な質問を抑制しながらも高い推定性能を維持すること，さらには逐次的に観測されるデータからリアルタイムに個人の感情推定精度向上を達成することを目的として設計されている．このように，多段階の学習プロセスを統合した RAL-MSA は，実際の対話環境に対応するための枠組みを提供する．

不確実性の定量化．本研究における第一の目的は，モデルの予測に対する不確実性が特に高いサンプルを同定することであり，そのようなサンプルに対して真のラベルを付与することが精度向上に最も寄与すると期待される．本手法では，各モダリティに対して個別に不確実性を推定することで，音声，言語，映像といった情報源それぞれの観点から予測の信頼性を多面的に評価できるように設計しており，これにより，単一モダリティに依存した場合では捉えきれない不確実性を総合的に把握することが可能となる．

ラベル要求の判断．次に，各モダリティから得られた不確実性の指標を重み付き和として統合し，この統合値に基づいてラベル要求の要否を判断する．このアプローチにより，モダリティそれぞれの重要度を考慮しつつ全体のラベル要否の決定をし，その結果として，単一モダリティのみでは判断が困難なケースであっても，より適切なラベル要求判断を行えるようにすることを目指している．

強化学習による方策更新．さらに，ラベル要求方策を強化学習の枠組みに基づいて継続的に更新することにより，システムが自身の行動結果から学習し，報酬最大化を指向した振る舞いを獲得できるようにしている．具体的には，取得した

報酬に基づいて方策のパラメータを逐次的に更新することで，ユーザごとに異なる感情表出の傾向や不確実性パターンを徐々に学習し，個人差をより適切に反映したラベル要求判断が可能となるように設計されている．これにより，対話システムは，単なる静的な不確実性評価に基づく単発的な判断にとどまらず，経験に応じて柔軟に意思決定方策を改善していく適応的な学習プロセスを実現する．

4.3 マルチモーダル特徴に基づく不確実性推定

各時刻 t において，システムが処理する入力は以下のように，音声特徴，言語特徴，および映像特徴から構成されるマルチモーダル特徴ベクトルとして表現される．

$$x_t = \{x_t^{\text{audio}}, x_t^{\text{text}}, x_t^{\text{visual}}\}. \tag{4.1}$$

ここで，各モダリティに対応する特徴量 x_t^m は，まずそのモダリティに特化して学習された Random Forest モデルへ入力され，モデルは感情ラベル集合 $\mathcal{Y} = 1, \ldots, C$ に対する予測分布を出力する．これにより，各モダリティごとに独立した視点から感情推定が行われるとともに，それぞれが持つ情報量の違いを反映した予測が得られる．

本研究では，これらの出力確率分布を用いて不確実性を定量化するため，各モダリティ m に関して，予測分布の中で最も高い確率値を確信度スコアとして採用する．この最大確率は，モデルがそのモダリティから得られる情報に基づいて最も「確信している」ラベルに対する確信度を直接的に表す指標となるものであり，確信度が低い場合には当該サンプルが曖昧である，すなわちラベルを取得することで分類境界の改善に寄与する可能性が高いサンプルであることを意味する．したがって，この確信度スコアを通じて，各モダリティが内包する不確実性を細やかに評価でき，マルチモーダル環境における情報のばらつきや表現の曖昧さを適切に扱うための基盤が構築される．

$$\phi_t^m = \max_{c \in \mathcal{Y}} P(y = c \mid x_t^m). \tag{4.2}$$

各モダリティを独立に評価することにより，システムは異なる情報源に基づく多角的な観点から各サンプルに対する予測の確信度を詳細に把握できる．このような評価は，音声・言語・映像といったモダリティごとに存在する特徴の偏りや表出傾向の違いを適切に反映し，単一モダリティでは捉えきれない不確実性をより精緻に抽出することを可能にする．

4.4 ラベル要求決定

各モダリティ m に対して，確信度スコア ϕ_t^m を閾値 θ_t と比較し，その結果に基づいて指標変数 $d_t^m = \mathbb{I}[, \phi_t^m < \theta_t,]$ を定義する．このとき，θ_t を下回る低確信度

の予測結果のみを意思決定の対象として扱う．これは，確信度が低いサンプルは，特徴空間のこれまで十分に探索されていない領域や，解釈が曖昧で判断が難しい領域に位置している可能性が高く，そのラベルを取得することがモデル性能の改善に寄与しやすいためである．

次に，これらの指標変数を，モダリティごとに学習される重み $\omega_t[m]$ を用いて集約し，ラベル要求スコアを以下のように算出する．

$$s_t = \sum_m \omega_t[m] d_t^m. \tag{4.3}$$

この値 s_t が閾値 τ（本研究では $\tau = 0.5$ に設定）以上である場合に，ラベル要求が実行される．ここで τ は，複数のモダリティがどの程度一致して「不確実である」と判断したときにラベル要求を行うかを制御するパラメータとして機能する．

さらに，ラベル要求が特定の条件に偏りすぎて早期に停止することを防ぐため，本研究では ϵ-greedy 機構も導入している．具体的には，一様乱数 $u \sim \mathcal{U}[0,1]$ が ϵ を下回った場合には，$s_t < \tau$ であってもラベルを要求する．この追加の探索機構により，高確信度でありながら誤分類されているケースや，これまでに観測されていない新規の表現パターンを検出する機会を確保でき，結果としてポリシーが過度に保守的になることを防ぎつつ，より柔軟で頑健なラベル要求方略を学習できる．

4.5 強化学習によるポリシー更新

各ラベル要求がどの程度有用であったかは，その時点で得られる報酬 r_t によって定量的に評価される．具体的には，

$$r_t = \begin{cases} \rho^+, & \text{if a query is made and } \hat{y}_t \neq y_t, \\ \rho^-, & \text{if a query is made and } \hat{y}_t = y_t, \\ 0, & \text{if no query is made.} \end{cases} \tag{4.4}$$

すなわち，モデルが誤った推定を行っている状況においてシステムがラベル要求を発行した場合には，正の報酬 ρ^+ が与えられ，要求が結果として誤りを回避したことが明示的に評価される．一方で，モデルの推定が既に正しいにもかかわらずシステムが不要な問い合わせを行った場合には，負の報酬 ρ^- が割り当てられ，無駄なクエリを抑制する方向へと学習が誘導される．このような報酬設計により，システムは予測を間違えるようなサンプルに対してのみラベル要求を行うよう最適化される．なお，ϵ-greedy によって純粋に探索目的で発行された問い合わせについては，ポリシー更新には利用しない点にも注意が必要である．ポリシー更新は，モダリティ全体の決定（$s_t \geq \tau$）によってラベル要求が妥当と判断された場合にのみ適用される．

信頼度閾値 θ_t の更新は，RAL [23] に従い，次式で与えられる．

$$\theta_{t+1} \leftarrow \min\left\{\theta_t\left(1+\eta\left(1-2^{\frac{r_t}{\rho^-}}\right)\right),\, 1\right\}, \tag{4.5}$$

この式から分かるように，もし $r_t = \rho^-$ すなわち「不要な問い合わせ」が発生した場合には，$2^{\frac{r_t}{\rho^-}}$ が急激に増加するため θ_{t+1} は急速に減少し，システムはより保守的に問い合わせ判断を行うよう調整される．一方で，$r_t = \rho^+$ の場合には，問い合わせが有益であったことを反映して閾値が緩やかに増加し，将来的に同様の状況で問い合わせをやや積極的に行うよう促される．ただし，急激な変動を防ぐために更新は上限 1 に制約されている．

さらに，各モダリティの寄与度を表す重み $\omega_t[m]$ についても，問い合わせがどのモダリティの判断に基づいて行われたかに応じて更新が行われる．特に，問い合わせに寄与したモダリティ（$d_t^m = 1$）に対しては，次式で示されるように，その重みを指数関数的に増減させる更新則が用いられる．

$$\omega_{t+1}[m] \leftarrow \begin{cases} \omega_t[m] \cdot \exp(\eta, r_t), & \text{if } d_t^m = 1, \\ \omega_t[m], & \text{otherwise.} \end{cases} \tag{4.6}$$

この設計により，あるモダリティの判断が全体モデルの決定と一致し，かつその問い合わせが実際にメリット（正の報酬）をもたらした場合には，そのモダリティの影響力が強化される．逆に，その判断が不適切（負の報酬）であった場合には重みが減少し，将来的にはそのモダリティの影響が相対的に小さくなる．この後，重みベクトル全体は以下のように正規化され，確率単体上に射影される．

$$\omega_{t+1}[m] \leftarrow \frac{\omega_{t+1}[m]}{\sum_{m' \in M} \omega_{t+1}[m']} \quad \forall m. \tag{4.7}$$

以上の一連の更新において，学習率 η はこれらの動的な更新の変化度合いを制御する役割を担っており，単一の問い合わせ結果によってポリシー全体が過度に変動してしまうことを防止する．このように，RAL-MSA におけるポリシー更新機構は，報酬構造・閾値調整・モダリティ重み更新といった複数の要素が相互に連携することで，対象ユーザの感情表出特性に徐々に適応しつつ，効率的な問い合わせ戦略の獲得を可能としている．

4.6 分類モデル

本研究における分類モデルは，ラベル要求ポリシーとは独立して学習される構成となっている．この分離により，ラベル要求の戦略学習が分類性能に直接干渉することなく，両者がそれぞれの目的に応じて最適化される点に大きな利点がある．本研究では，複数のモデル候補を比較検討した結果，表 6.1 で示すように最

も優れた総合性能を示した Random Forest に基づくアンサンブルモデルを採用した．Random Forest は多数の決定木を統合することで高い汎化性能とロバスト性を確保できるほか，モダリティごとに独立した特徴量空間を扱いやすいという点でも本研究のマルチモーダル設定と極めて相性が良い．

具体的には，各時刻 t において，モダリティごとに構築された分類器はそれぞれ入力特徴 x_t^m に基づき，感情ラベル集合 $\mathcal{Y}$ に対する事後確率分布 $P(y \mid x_t^m)$ を出力する．このようにして得られた複数の確率分布は，音声・言語・映像といった情報源から得られる独立した判断を表しており，これらを統合することでより信頼性の高い推定が実現される．最終的な推定ラベル $\hat{y}_t$ は，すべてのモダリティにおける確率値を比較し，最も高い確率を与えるクラスを選択するという単純な方式によって決定される．この手法は，マルチモーダル情報を扱う際に生じがちなモダリティ間の情報量の偏りや不均衡の影響を緩和するとともに，各モダリティの強みを生かした形で推定結果を統合できるという点で有効である．

以上の理由から，本研究では Random Forest を基盤とするアンサンブル型の分類モデルを採用し，マルチモーダル感情推定における信頼性と適応性を両立させる設計としている．

$$\hat{y}_t = \arg\max_{c \in \mathcal{Y}} \frac{1}{|M|} \sum_{m \in M} P(y = c \mid x_t^m). \tag{4.8}$$

第5章　実験設定

本研究では，提案手法の有効性を検証するために，既存のベースライン手法と同一条件下で比較実験を行った．以下では，実験に用いたデータセットの詳細，マルチモーダル特徴量の抽出手順，評価指標および評価手順，全モデルに共通する設定や学習に用いたハイパーパラメータなど，実験設定全体を説明する．

5.1　データセット

本研究では，人間参加者とエージェントとの対話から構成される公開コーパスである Hazumi1902，Hazumi1911 [24, 25] を用いる．これらのコーパスは，人と対話システムとの相互作用を対象として収録されたデータセットであり，各交換ごとに実験参加者が主観的感情状態（Subjective Sentiment: SS）を自己申告するという特徴を有している．ここでいう「交換（exchange）」とは，システムによる発話に続いてユーザが発話する一連の対話単位を指す．参加者構成や平均交換数などの統計情報を表 5.1 に示す．交換参加者は，その対話単位に対して「会話を全く楽しんでいない」を表す 1 から「会話を非常に楽しんでいる」を表す 7 までの 7 段階の尺度を用いて感情状態を評価している．

本研究では，先行研究 [22] に従い，得られた SS スコアを 3 クラスに離散化した．すなわち，スコア 5〜7 をポジティブ，4 をニュートラル，1〜3 をネガティブとして扱う．詳細なデータ分布に関しては付録 A に記載する．

Hazumi シリーズの対話は，Wizard-of-Oz 法によって収録されている．この方法では，表面的には対話エージェントが応答しているように見えるものの，実際には人間のオペレータがエージェントを操作することで，より自然で柔軟な対話が可能となる．対話はすべてビデオ録画されており，これにより音声・言語・映像の各モダリティを含むマルチモーダル特徴の抽出が可能となっている．

Hazumi1902 には 28 名（うち女性 19 名）の参加者が含まれ，一方の Hazumi1911 には 26 名（うち女性 14 名）が含まれている．これらのコーパスは，個人差を含む多様な感情表出を捉えており，本研究における個人適応型マルチモーダル感情推定の評価に適したデータセットといえる．また，両コーパスは収録時期や参加者構成に差異があるため，複数データセット間での再現性や一般化性能を検証する上でも有用である．

表 5.1: Hazumi1902 と Hazumi1911 の統計情報.

	Hazumi1911	Hazumi1902
参加者数	26	28
平均年齢	44.6 ±16.7	44.6 ±15.2
平均会話時間	20.5 min	17.7 min
平均交換数	95	83
総会話時間	534.0 min	495.3 min
総交換数	2468	2337

5.2 特徴量抽出

本研究では，先行研究 [26] において Hazumi データセットを用いたマルチモーダル感情分析で一般的に採用されている設定に厳密に従い，音声・言語・映像の 3 つのモダリティから特徴量を抽出する．これにより，各モダリティが持つ情報を統合しつつ，先行研究との比較が可能な形で実験を構築している．以下では，各モダリティに対する特徴抽出手順について詳細に述べる．

言語特徴：音声認識によって得られた発話の書き起こしについて，日本語形態素解析器である MeCab [27] を用いて形態素解析を行う．そのうえで，品詞ごとのトークン頻度や Bag-of-Words（BoW）特徴を抽出する．語彙数の差異により特徴次元数は各データセットで異なり，Hazumi1902 では 984 次元，Hazumi1911 では 2613 次元となる．これらの特徴は，対話中の言語的手がかりを豊富に捉えるために有効である．

音声特徴：音声モダリティについては，openSMILE ツールキット[1]を使用し，INTERSPEECH 2009 Emotion Challenge で定義された IS09 特徴セット [28] を各発話単位で抽出する．このセットはピッチやエネルギーといった特徴を含む全 384 次元から構成されており，話者の発話スタイルや感情表出における韻律的変化を捉えることができる．

映像特徴：映像モダリティでは，OpenFace [29] を用いて 10 か所の顔特徴点をトラッキングし，30 fps のフレーム系列から眼周囲および口周囲の 12 点に対してフレームごとの速度・加速度を算出する．その後，各ターン交換ごとにこれらの信号の最大値・平均値・標準偏差を特徴として抽出し，さらに AU（Action Unit）活性度の平均を併せて用いる．加えて，Microsoft Kinect V2 によって取得される頭部および肩関節の位置情報から，身体動作の速度・加速度指標を計算する．最終的な映像特徴ベクトルは，顔表情情報（66 次元）および身体動作情報（20 次元）を組み合わせた全 86 次元となり，視線・表情・体動といった多様な非言語的手がかりを包括的に表現する．

[1]https://www.audeering.com/research/opensmile/

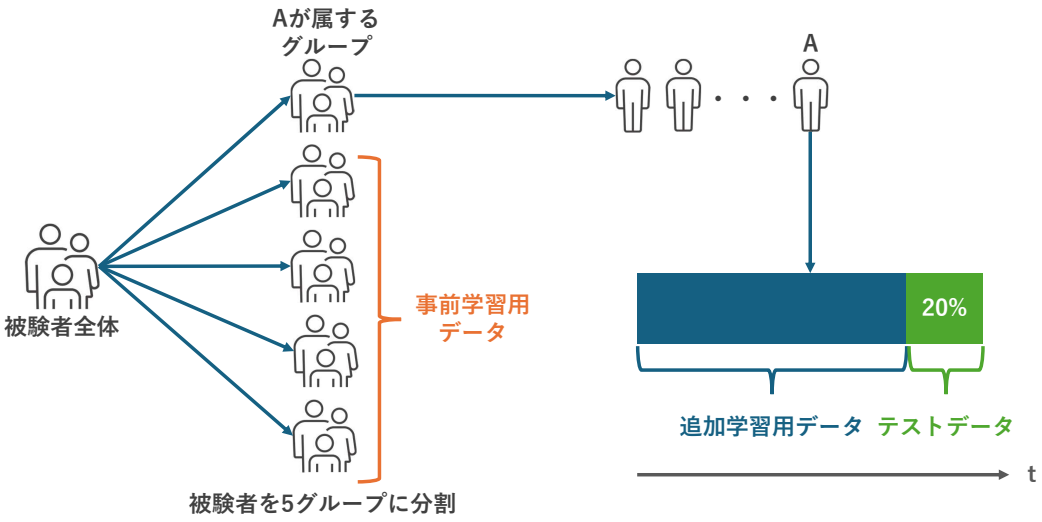


図 5.1: 交差検証方法.

5.3 評価手順

本研究では，評価の信頼性および再現性を確保するために，図 5.1 のような被験者単位でグループ化したクロスバリデーション方式を採用する．具体的には，全参加者を 5 つのグループ（Groups 1–5）に分割し，各実験において，そのうち 1 つのグループをターゲットグループとして扱い，残りの 4 グループを分類器およびラベル要求方策の事前学習に用いる．例えば，グループ 1 に所属する被験者 A をターゲットとする場合には，グループ 2–5 のデータを用いて事前学習を行う．このような設定により，ターゲットユーザに対する適応性能を評価でき，提案手法の参加者間での汎用性についても検証することが可能となる．

事前学習が完了した後，各ターゲットグループ内の参加者ごとにストリーム型能動学習を実行する．各参加者については，時間順に並んだ対話データのうち，最初の 80%を追加学習用データとして使用し，この部分でラベル要求と逐次的なモデル更新を行う．残りの 20%はテストセットとして保持し，追加学習後のモデル性能を評価するために用いる．ラベル要求は 1 サンプルずつ逐次的に実行され，新たに取得されたラベル付きサンプルが 5 件に達するたびに分類器の再学習を行う．このように逐次適応と部分的な再学習を繰り返すことで，システムはユーザ固有

の特徴を徐々に反映させながら更新される．得られた感情推定性能は，各参加者について個別に算出した後，ターゲットグループに属する全参加者で平均化することで集計される．

さらに，現実的な対話システム運用を想定すると，ユーザに頻繁にラベル提供を求めることは対話の自然性を損ない，ユーザ体験の低下を引き起こす可能性がある．そのため，本研究では，実際の使用状況を踏まえ，ユーザに対して 5 回から 10 回程度という，実運用上許容可能な範囲に絞った少数ショット条件（5-shot, 10-shot）で各手法を評価する．これにより，限られたラベル付与回数のもとでどれだけ効果的に性能向上が達成できるかを詳細に検証する．

性能指標としては，多クラス分類におけるクラス不均衡の影響を抑制できる Balanced Accuracy（BA）を採用する．BA は，各クラスにおける再現率（True Positive Rate）を平均することで定義され，クラス分布に偏りがあるデータセットに対しても公平な評価を実現する．クラス数を C とすると，BA は次式で表される．

$$\mathrm{BA} = \frac{1}{C} \sum_{j=1}^{C} \frac{TP_j}{TP_j + FN_j}, \tag{5.1}$$

ここで，TP_j および FN_j はそれぞれクラス j の真陽性数と偽陰性数を表す．この指標により，どのクラスに対しても均等な重要度を持つ性能評価が可能となり，本研究のように各クラスのデータ量に偏りが存在するタスクにおいて特に有用である．

5.4 ハイパーパラメータ

本研究では，分類器およびラベル要求ポリシーの両方において，各モダリティごとに独立した Random Forest モデルを用いて学習を行った．具体的には，いずれのモダリティに対しても，100 本の決定木から構成されるアンサンブル（n_estimators=100）を構築し，決定木の深さに関しては制限を設けず（max_depth=None），分岐基準としては Gini 不純度（criterion='gini'）を採用し，さらにブートストラップサンプリングを有効化する設定とした．このような構成により，モダリティ間で特性の異なる特徴空間を扱う場合においても，各モデルが十分な表現能力を維持しつつ汎化性能を確保できるよう配慮している．

また，ラベル要求ポリシーにおける探索戦略としては，ϵ-greedy 方式を採用し，探索率 ϵ を 0.05 に設定した．この設定により，方策の収束を妨げない範囲でランダムな探索を導入し，高信頼度の予測が連続する状況でも未知のサンプルを能動的に取得しうる柔軟性を確保している．報酬設計については，誤推定を防ぐ有益なラベル要求に対して正の報酬 $\rho^{+} = 1$ を与え，逆に不必要なラベル要求に対しては負の報酬 $\rho^{-} = -1$ を与えることで，ポリシーが有益な問い合わせのみを選択するようにした．

さらに，閾値およびモダリティ重みを更新する際の学習率 η は，事前学習において 5×10^{-3} とし，対象ユーザへの個人適応時には 1×10^{-2} に設定した．これは，事前学習段階では比較的安定した更新を行いつつ，個人適応段階ではユーザ固有の表出特性に迅速に適応できるようにするための設計である．加えて，能動学習におけるラベル要求の総予算は 0.5 に設定し，過度な問い合わせがユーザの対話体験を損なうことを防ぐと同時に，限られたラベル要求回数で最大限の性能向上を得られるよう配慮した．

5.5 比較モデル

提案する RAL-MSA の有効性を検証するため，本研究では以下に示す複数の比較モデルを設定し，それらと同一条件下で性能評価を行った．これらの比較により，RAL-MSA がどの要素によって性能向上を実現しているのかを明らかにすることができる．

Random sampling．最も単純なベースラインとして，対話ストリーム中に出現するサンプルの中から，ラベル要求を一様ランダムに選択する手法を用いる．この手法は，能動学習を行わずにランダムなラベル選択を行う場合の性能下限を与えるものであり，提案手法による効率的なラベル選択および性能改善が偶然によるものではないことを明確に示すための基準点となる．

w/o threshold θ．本手法の構成要素である信頼度しきい値 θ の個人適応更新を行わず，事前学習データから得られた固定値をそのまま使用する手法である．事前学習データは複数の非対象ユーザの対話から構成されているため，この固定値は「平均的なユーザ」に最適化されたしきい値とみなすことができる．この比較により，しきい値 θ をユーザごとに逐次適応させることの重要性を定量的に検証できる．

w/o weight ω．モダリティ重み ω のオンライン更新を行わず，事前学習で得られた固定重みを用いる手法である．この固定重みも，しきい値と同様に多数の実験対象ではないユーザのデータに基づいて推定された「平均的ユーザ」に最適化されたパラメータである．この比較により，各モダリティの貢献度を対話中に逐次調整する個人適応メカニズムが性能向上にどれほど寄与しているかを評価できる．

Confidence(Pool)．推定器の出力における最大確信度に基づくプール型能動学習手法を比較対象として用いる．当該手法はプール型能動学習に分類されるため，追加学習用データを逐次的に取得し，各サンプルごとにラベル要求の要否を判断するストリーム型能動学習とは異なり，追加学習用データ全体を事前に俯瞰した上で，推定器の確信度が最も低い上位 5 サンプルを選択し，それらに対してラベル付与および学習を行う．この手法との比較を通じて，データ全体を俯瞰可能なプール型能動学習と，逐次的な判断を必要とするストリーム型能動学習を採用する本手法との間に生じる推定精度の差異を定量的に評価することが可能となる．

Entropy(Pool). さらに，比較手法として，推定器の出力分布のエントロピーに基づくプール型能動学習手法を用いる．エントロピーベースの手法は，予測クラスの最大確信度のみならず，クラス分布全体の不確実性を考慮できる点に特徴がある．本研究では，最大確信度ベースの手法と同様に，追加学習用データ全体を俯瞰した上で，各学習ステップにおいてエントロピーが最も高い上位 5 サンプルを選択し，ラベル付与およびモデル更新を行う．

これらのモデルを含めた比較実験により，RAL-MSA が獲得している性能向上が，個人適応的なラベル要求戦略やモダリティ重みの更新といった最適化の仕組みに起因するものなのか，あるいは単に複数ユーザを平均化したパラメータに依存した結果にすぎないのかを体系的に検証することが可能となる．このように，多面的な比較を通じて，提案手法の有効性およびその構成要素の寄与度を示すことができる．

第6章　結果と考察

6.1　感情推定モデルの選定

表 6.1 に，Hazumi1902 データセットにおける各種感情推定モデルの Balanced Accuracy を示す．ここで示す Balanced Accuracy は，参加者ごとに算出した値を平均したものであり，あわせて 95% 信頼区間を示している．太字は，各モデルにおいて最も高い性能を示した結果を表す．両データセットにおいて，ランダムフォレストモデルは k 近傍法，決定木，多層パーセプトロンと比較して最も高い性能を示した．これらの結果に基づき，本研究では感情推定モデルとして ランダムフォレストモデルを採用した．ここで，太字はそれぞれのモデルの中で最も良い性能を表している．

表 6.1: 各モデルの 0-shot Balanced Accuracy (±95% 信頼区間).

Models	**Hazumi1902**	**Hazumi1911**
k-nn	0.424 ± 0.084	0.457 ± 0.075
Decision Tree	0.403 ± 0.063	0.433 ± 0.072
Random Forest	**0.467** ± 0.083	**0.475** ± 0.078
Neural Network	0.425 ± 0.066	0.418 ± 0.075

6.2　能動学習戦略の比較

Table 6.2 には，両データセットにおける 0-shot，5-shot，10-shot の各条件で得られた Balanced Accuracy(±95% 信頼区間) を示す．また，Confidence(Pool), Entropy(Pool) を除いて，最も良い性能を太字にしている．なお，Hazumi1911 における被験者 M6002 は，5-shot 条件において一度もラベル要求が発生しなかったため，分析の一貫性を保つ観点から除外した．

まず，全体的な傾向として，RAL-MSA は 5-shot および 10-shot の両条件において，Random Sampling と比較して最大 1.5 %の精度向上を示し，限られたラベル問い合わせ回数であっても有効な性能向上が可能であることが確認された．一方で，Hazumi1911 の 5-shot 条件においては，閾値適応を行わないモデル（w/o

表 6.2: Hazumi1902 ($n = 28$) と Hazumi1911 ($n = 25$) における能動学習戦略の性能比較.

Method	Hazumi1902			Hazumi1911		
	0-shot	5-shot	10-shot	0-shot	5-shot	10-shot
Random Sampling	0.467 ± 0.083	0.468 ± 0.083	0.470 ± 0.080	0.485 ± 0.079	0.477 ± 0.081	0.470 ± 0.086
w/o Threshold (θ fixed)		0.475 ± 0.078	0.471 ± 0.084		**0.491** ± 0.081	0.484 ± 0.078
w/o Weight (ω fixed)		**0.476** ± 0.078	**0.472** ± 0.085		0.490 ± 0.082	**0.485** ± 0.082
Ours (RAL-MSA)		**0.476** ± 0.078	**0.472** ± 0.085		0.490 ± 0.082	**0.485** ± 0.082
Confidence(Pool)		0.494 ± 0.094	0.500 ± 0.090		0.483 ± 0.093	0.468 ± 0.090
Entropy(Pool)		0.474 ± 0.088	0.494 ± 0.089		0.480 ± 0.092	0.481 ± 0.090

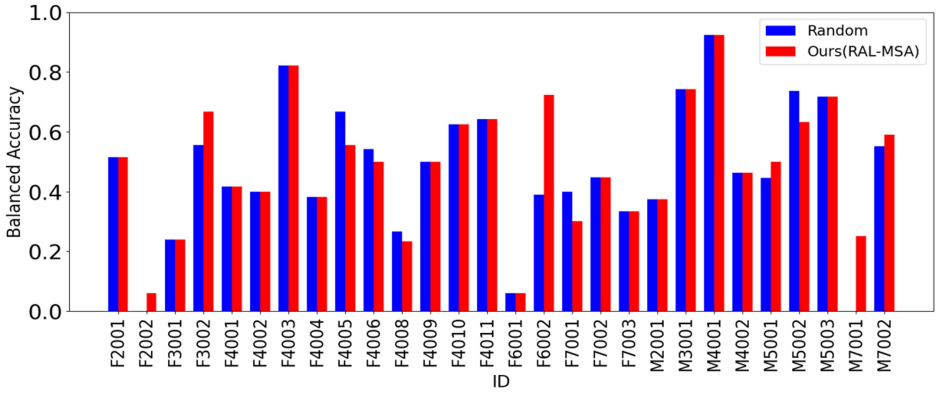
Threshold）が RAL-MSA をわずかに上回る結果となった．しかしながら，RAL-MSA が 10-shot 条件において最良の性能を示した点を踏まえると，極めて初期の学習段階においては，利用可能なデータ量が少ないことに起因して不確実性推定が十分に安定せず，閾値適応の効果が顕在化しにくかった可能性が考えられる．

さらに，RAL-MSA と w/o Weight は両データセットにおいて完全に同一の結果を示した．このことは，少数ショット条件の短い適応期間においては，モダリティ重みの調整が性能に寄与する程度が比較的限定的であること，すなわち，使用可能なラベル数が少ない状況では重み更新の影響が顕著に反映されにくいことを示唆している．

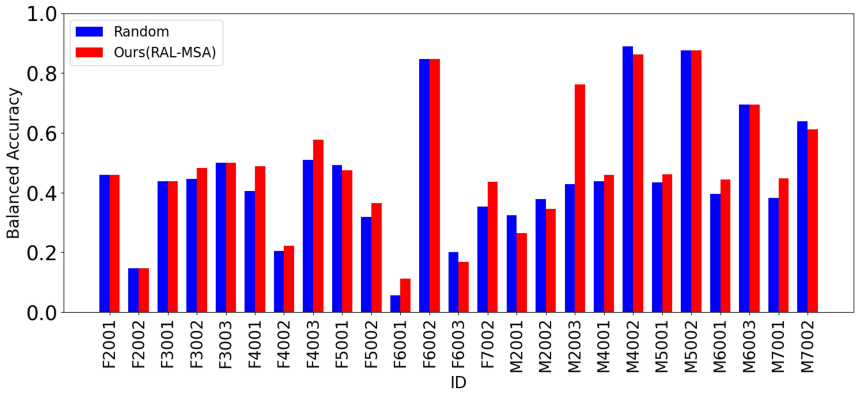
また，Hazumi1902 では RAL-MSA の性能が 5-shot から 10-shot へと低下する一方で，Hazumi1911 では RAL-MSA とベースラインの双方で類似した低下傾向が確認された．これらの挙動の明確な原因は現時点では特定できていないものの，短期的なサンプル選択のばらつきや，極少データ条件に起因する不安定性など，さまざまな要因が複合的に作用した可能性が考えられる．この点については，より大規模かつ長期的な対話データセットを用いた追加検証や，ラベルの多様性を考慮した獲得戦略の導入などを通じて，より詳細に明らかにすることが今後の重要な課題である．

プール型能動学習である Confidence(Pool) と Entropy(Pool) に着目すると，Hazumi1902 においては，プール型能動学習手法が全体として提案手法を上回る傾向が確認される．これは，各学習ステップにおいて未ラベルデータ全体を俯瞰し，不確実性の高いサンプルを選択できるというプール型能動学習の特性によるものと考えられる．さらに Hazumi1911 においては，予測の不確実性に基づいてサンプル選択を行うプール型能動学習手法が，必ずしも高い性能向上につながっていない点が確認される．この結果は，Hazumi1911 においては，推定器が算出する不確実性が必ずしも情報量の高いサンプル選択を適切に反映していない可能性を示唆している．

さらに，本研究全体にわたって被験者間で極めて大きな性能差が観察されており，これは本研究で採用した検証手法が個人差を反映する性質を持つことを示している．各被験者の詳細な性能分析とその背景にある要因については，章 6.3 においてさらに掘り下げて議論する．



(a)Hazumi1902



(b)Hazumi1911

図 6.1: 被験者ごとの精度.

6.3 被験者別精度分析

図 6.1 には，Hazumi1902 および Hazumi1911 の各被験者における Random Sampling と RAL-MSA の 5-shot 条件における Balanced Accuracy の比較結果を示す．横軸は被験者の ID，縦軸はそれぞれの被験者の Balanced Accuracy を表している．また，ID の先頭につく F，M は実験参加者が女性または男性であることを示し，それに続く 2 桁の数字は年齢を表している．例えば，F2001 は 20 代の女性である．

まず Hazumi1902 に着目すると，全 28 名の被験者のうち，RAL-MSA が Random Sampling よりも高い精度を示したのは 6 名にとどまり，5 名では逆に性能が下回り，さらに 17 名では両手法が同一の精度を示した．この結果は，被験者全体としては両手法の性能が概ね近い値となる一方で，被験者によっては両者の差が顕著になるケースが存在することを示している．特に F6002 や M7001 のように，

RAL-MSA が 20%を超える大幅な性能向上を達成した例も確認されている．また，Random Sampling では精度が 0% であった F2002 や M7001 といった被験者において，RAL-MSA は非ゼロの精度を達成しており，情報量の高いサンプルの選択的ラベル取得が，データ不足やモダリティ表出特性の個体差によって生じる困難をより効果的に克服し得る可能性を示唆するものである．

続いて Hazumi1911 における結果をみると，25 名中 12 名で RAL-MSA が Random Sampling を上回り，6 名では下回り，7 名では同等の性能を示した．すなわち，全体として約半数の被験者が RAL-MSA によって性能向上を達成しており，個人適応型の不確実性推定にもとづくストリーム型能動学習が一定の効果を発揮していることが示唆される．特に M2003 においては約 30%という顕著な改善が得られており，RAL-MSA が個々の被験者ごとに異なる不確実性分布や表出特性に対して効果的に適応できる可能性を裏付ける結果となっている．

以上の結果から，RAL-MSA はすべての被験者に対して一様に優れた性能を示すわけではないものの，特定の被験者に対しては極めて大きな改善効果をもたらし，個人差の大きいマルチモーダル感情推定において有効な戦略となり得ることが明らかである．また，本研究で採用した評価手法が個人差を反映するものであることを踏まえると，本手法の効果をより正確に検証するためには，長期的データや多様な対話条件における追加検証が今後の重要な課題となる．

6.4 モダリティ重み (ω) の推移

本節では，RAL-MSA におけるモダリティ重み適応機構がどの程度性能向上に寄与したのかを検証するために，学習過程におけるモダリティ重みの変動を詳細に分析する．各モダリティの重みの推移は 図 6.2 に示すとおりであり，これらの推移を通して，対話データの逐次処理中にシステムが各モダリティをどのように重要視したかを明らかにする．横軸はラベル要求を行ったサンプル数，縦軸はそれぞれのモダリティ重みの値を表している．また，グラフの色はそれぞれの被験者を表している．本研究では，各データセットに含まれる被験者を五つのグループに分割し，章 5.3 で述べた手順に従ってグループ単位で事前学習を行った．その結果として，各データセットに対して五種類の初期の重み（0-shot 値）が生成されることになり，この初期の重みは 表 6.3 に記載している．

Hazumi1902 においては，映像モダリティに割り当てられた重みが学習初期から一貫して相対的に高い値を示しており，視覚的手がかりが本データセットにおける不確実性推定の主要な根拠として機能していたことが示唆される．この傾向は Hazumi1911 においても観察され，映像モダリティの重みが他モダリティと比較して少し高めの状態で推移した．これは，表情変化や身体動作といった視覚的特徴が，対話時の不確実性推定において重要な情報源になっていた可能性を示すものである．

表 6.3: それぞれの検証グループにおける 0-shot 時のモダリティ重み，不確実性閾値の値.

Group	Hazumi1902		Hazumi1911	
	threshold θ	modality weight $\{\omega[\text{text}], \omega[\text{audio}], \omega[\text{visual}]\}$	threshold θ	modality weight $\{\omega[\text{text}], \omega[\text{audio}], \omega[\text{visual}]\}$
Group 1	0.615	{0.303, 0.349, 0.349}	0.520	{0.348, 0.300, 0.352}
Group 2	0.590	{0.260, 0.359, 0.381}	0.516	{0.336, 0.332, 0.332}
Group 3	0.549	{0.300, 0.331, 0.369}	0.514	{0.284, 0.308, 0.408}
Group 4	0.614	{0.315, 0.334, 0.351}	0.570	{0.314, 0.350, 0.336}
Group 5	0.565	{0.313, 0.339, 0.349}	0.479	{0.329, 0.310, 0.360}

一方で，Hazumi1911 においては，クエリ数の増加に伴い言語モダリティの重みが大きく変動し，その分散が拡大する傾向が確認された．これは，言語的特徴の重要度が被験者ごとに大きく異なっていたことを意味しており，モダリティ重みの適応機構が，個人差に起因する不確実性の源泉を捉える上で一定の役割を果たしていた可能性を示す．特定の個人では発話内容に不確実性推定に役立つ情報が含まれていた一方，他の個人では非言語的手がかりが相対的に重要であったなど，個人適応的な推定に寄与した側面がうかがえる．

ただし，RAL-MSA では，いずれかのモダリティの重みが 0.5 を超えると，その単一モダリティの判断のみでラベル要求条件 $s_t \geq \tau$（本研究では $\tau = 0.5$）を満たすことが可能になるが，図 6.2 に示されるように，学習全体を通していずれのモダリティ重みも 0.5 を超えることはなかった．このことから，本研究の few-shot 条件においては，モダリティ重みの適応が直接的にラベル要求決定を支配するほどの影響力を持たず，最終的な性能に対する寄与がなかったと推察される．

しかし，より長期の対話データや，より大きなクエリ予算を設定したシナリオでは，重み適応の影響が強まる可能性がある．多様なサンプルが蓄積されることでモダリティごとの信頼度の反映能力が向上し，個人ごとのモダリティの不確実性重要度をより精密に反映できるようになるためである．したがって，モダリティ重み適応は短期的な few-shot 学習では顕著な影響を示さなかったものの，長期的な個人適応シナリオにおいては重要な構成要素となり得る点に留意する必要がある．

6.5 不確実性閾値 (θ) の推移

本節では，不確実性閾値の変動に着目することで，閾値適応機構が果たす役割およびその影響を分析する．各クエリに対する不確実性閾値の推移は，図 6.3 に示されている．ここで，横軸はラベル要求を行ったサンプル数，縦軸は不確実性閾値の値を表している．また，グラフの色はそれぞれの被験者を表している．本研究では，各データセットの被験者を 5 つのグループに分割して事前学習を行ったため（章 5.3），各データセットに対して 5 種類の 0-shot 初期値が存在する．これら初期閾値については 表 6.3 に明示されている．

両データセットにおいて，クエリが進むにつれて fold 内での閾値の分散が増大する傾向が確認された．これは，閾値適応機構が，各被験者が示す不確実性分布

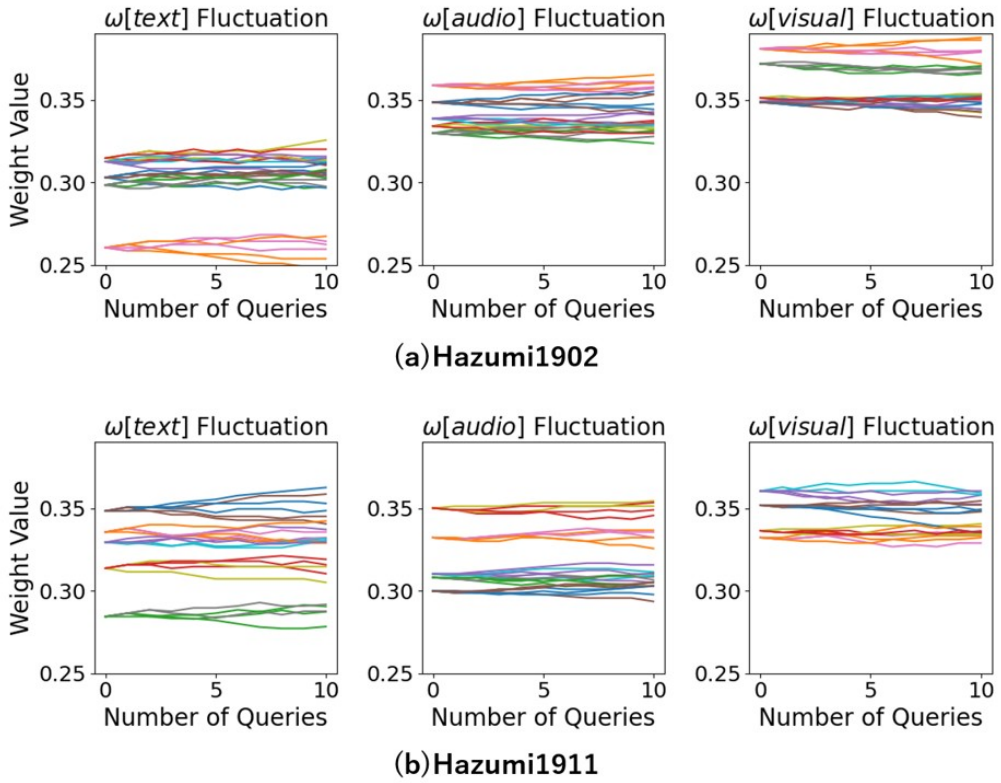


図 6.2: モダリティ重み (ω) の推移.

の違いに応じて動的に変化していることを示唆している．すなわち，被験者特性に基づく柔軟な調整が自動的に行われていると解釈できる．

さらに，w/o Threshold ベースラインの結果と併せて考察すると，閾値適応は個人適応に対して一定の効果をもたらしていた可能性が高いことが示唆される．特に，学習初期の極めて少ないサンプルしか観測できない段階では十分に機能しない場合があったものの，ある程度のサンプルが蓄積されて以降は，被験者固有の不確実性特徴に合わせて閾値が調整されることで，より安定したラベル要求戦略，さらには推定性能の向上に寄与していたと考えられる．

6.6 今後の展望

本研究には，対話ベースの個人適応研究に一般的に見られるいくつかの限界が存在する．第一に，本研究で提案した RAL-MSA は，類似した条件下で収集された Hazumi1902 および Hazumi1911 という 2 つのコーパスのみを用いて評価を行っており，より多様なユーザ層や幅広い対話状況に対する一般化可能性には制約がある．そのため，将来的には，長期的な時間変化への適応能力や，幅広いユーザ

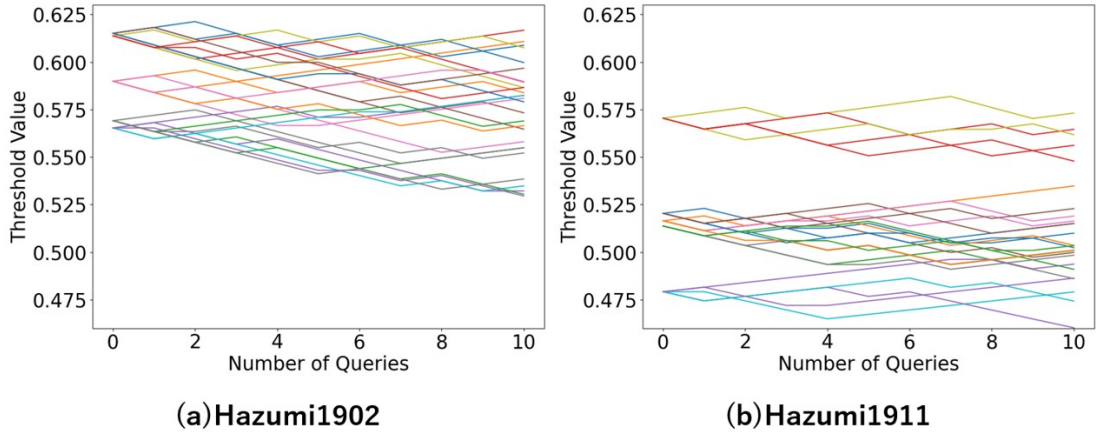


図 6.3: 不確実性閾値 (θ) の推移.

層，さらには異なる言語環境を含む多様なデータセットを対象とした検証を行い，より包括的な評価を進める必要がある．

第二に，本研究の実験はすべてシミュレーション環境で実行されており，実際のユーザを用いた評価は行っていない．このため，リアルタイム推定の挙動やユーザ体験への影響といった実運用上の側面が未検討のまま残されている．特に，ユーザに過度な負担を与えずにどの程度の頻度で問い合わせを行えるのか，また，ユーザ自己申告による感情ラベルの信頼性をどのように保証するかといった課題は，現実環境でシステムを導入する際に極めて重要であり，今後の検討が不可欠である．

第三に，本フレームワークでは，各発話後にラベル要求を行うという前提を置いているが，より自然な対話の流れに基づくタイミング，例えば話題の切れ目や沈黙の挿入箇所など，ユーザの負担が少なくかつ情報的価値が高いと考えられる場面で選択的にラベルを取得する戦略については十分に検討されていない．加えて，サンプルの多様性を積極的に促すようなラベル獲得戦略についても，さらなる探索の余地が残されている．

以上のような限界は存在するものの，本研究は，ストリーム型能動学習に基づく個人適応的なマルチモーダル感情推定に向けた初期的な試みとして位置付けられ，ラベル取得コストを抑えつつ精度向上を実現するための有望な方向性を示している．今後の研究において，本研究の枠組みを発展させ，より実践的かつ汎用性の高い対話システムの設計へとつなげていくことが期待される．

第7章　結論

本研究では，対話システムにおける個人適応型マルチモーダル感情分析を実現するために，強化学習に基づくストリーム型能動学習手法である RAL-MSA を提案した．本手法は，対話の進行中に逐次的に観測される音声・言語・映像といったマルチモーダル行動に基づき，各時点で感情ラベルを要求するか否かを自律的に判断する枠組みであり，感情推定精度の向上とユーザ負担の低減を同時に達成することを目的として設計されている．

RAL-MSA の特徴は，単に不確実性の高いサンプルに対して機械的にラベル要求を行うのではなく，ラベル要求の結果が実際に推定性能の改善に寄与したかどうかを報酬として評価し，その結果に基づいてラベル要求方策そのものを強化学習によって最適化する点にある．さらに，ユーザごとに異なる感情表出特性へ適応するために，モダリティ重みおよび不確実性閾値を対話の進行とともに動的に更新する仕組みを導入することで，不要な問い合わせを抑制しつつ，限られたラベル取得機会を最大限に活用できる設計となっている．

本研究の主な貢献は，以下の点にまとめられる．第一に，個人ごとに異なるマルチモーダル感情表出の多様性を考慮し，個人適応型マルチモーダル感情分析をストリーム型能動学習問題として定式化した点である．第二に，マルチモーダルな不確実性推定と強化学習を統合し,「いつ感情ラベルを要求すべきか」を学習によって獲得する実践的なフレームワークを構築した点である．第三に，人間–エージェント対話コーパスを用いた評価を通じて，少数サンプル条件下における提案手法の有効性を実証した点である．

Hazumi1902 および Hazumi1911 の 2 つの対話コーパスを用いたシミュレーション実験の結果，RAL-MSA は few-shot 条件において Random Sampling を一貫して上回る性能を示し，特に一部の被験者においては顕著な性能向上が確認された．この結果は，限られた問い合わせ回数のもとでも，強化学習に基づく能動的なラベル要求が個人適応を効果的に促進できることを示している．また，閾値適応は学習初期を過ぎた段階において個人適応に寄与していることが示唆された．一方で，推定精度への直接的な影響は限定的であったものの，モダリティ重みの適応は，被験者ごとに異なる情報源の重要度を反映する形で機能しており，個人差を捉える上で有用な指標となり得る可能性が明らかとなった．

以上の結果から，強化学習に基づく能動的なラベル要求戦略は，マルチモーダル感情分析において，効率的な個人適応を実現する有望なアプローチである可能性が示唆される．

一方で，本研究にはいくつかの課題も残されている．本研究の評価はシミュレーション環境に基づくものであり，実際のユーザを対象としたリアルタイム対話実験は実施されていない．そのため，実運用環境におけるユーザ体験や，感情ラベル要求が対話の自然性に与える影響については，今後の検証が必要である．また，評価に用いたコーパスは日本語かつ類似した収録条件下で構築されたものであるため，異なる言語や対話スタイルに対する一般化性能についても今後の課題として挙げられる．今後は，より多様なユーザを対象とした長期的かつリアルタイムな対話環境へ RAL-MSA を拡張し，実ユーザとの実験を通じてユーザ負担やシステムの応答性の評価を行うことが重要である．さらに，サンプルの多様性を考慮した報酬設計や，感情以外への応用を検討することで，適応的な人間とエージェントとの対話の実現に向けた発展が期待される．

謝辞

本修士論文の執筆にあたり，多くの方々から多大なるご指導とご支援を賜りました．ここに深く感謝の意を表します．

まず，本研究を進めるにあたり，ご指導を賜りました指導教員の岡田将吾教授に心より感謝申し上げます．また，論文執筆に際し，構成や表現について多くの貴重なご助言と添削をしていただきました，大阪大学の駒谷和範教授に深く感謝申し上げます．さらに，研究方針の検討から論文執筆，添削に至るまで，親身にご相談に乗ってくださり，多くの有益な助言をいただいた林貴斗さんに心より感謝いたします．

最後に，本研究に関わり支えてくださったすべての皆様に，厚く御礼申し上げます．

参考文献

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

[2] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, Vol. 27, No. 12, pp. 1743–1759, 2009. Visual and multimodal analysis of human spontaneous behaviour:.

[3] Nicola Binetti, Nadejda Roubtsova, Christina Carlisi, Darren Cosker, Essi Viding, and Isabelle Mareschal. Genetic algorithms reveal profound individual differences in emotion recognition. *Proceedings of the National Academy of Sciences*, Vol. 119, No. 45, p. e2201380119, 2022.

[4] Demet Özer and Tilbe Göksun. Gesture use and processing: A review on individual differences in cognitive resources. *Frontiers in Psychology*, Vol. 11, , 2020.

[5] Jangwon Kim, Asterios Toutios, Sungbok Lee, and Shrikanth S. Narayanan. Vocal tract shaping of emotional speech. *Computer Speech & Language*, Vol. 64, p. 101100, 2020.

[6] Jialin Li, Alia Waleed, and Hanan Salam. A survey on personalized affective computing in human-machine interaction, 2023.

[7] Joe Li and Peter Washington. A comparison of personalized and generalized approaches to emotion recognition using consumer wearable devices: Machine learning study. *JMIR AI*, Vol. 3, , 2024.

[8] Kazunori Komatani and Mikio Nakano. User impressions of questions to acquire lexical knowledge. In *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 147–156, 2020.

[9] Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[10] Issei Waki, Ryu Takeda, and Kazunori Komatani. Learning to ask efficiently in dialogue: Reinforcement learning extensions for stream-based active learning. In *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 431–440, August 2025.

[11] Sahisnu Mazumder, Bing Liu, Shuai Wang, and Nianzu Ma. Lifelong and interactive learning of factual knowledge in dialogues. In *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 21–31, September 2019.

[12] Lingrui Mei, Shenghua Liu, Yiwei Wang, Baolong Bi, and Xueqi Cheng. SLANG: New concept comprehension of large language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 12558–12575, November 2024.

[13] Kohei Ono, Ryu Takeda, Eric Nichols, Mikio Nakano, and Kazunori Komatani. Lexical acquisition through implicit confirmations over multiple dialogues. In *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 50–59, August 2017.

[14] Jiwei Li, Alexander H. Miller, Sumit Chopra, Marc'Aurelio Ranzato, and Jason Weston. Learning through dialogue interactions by asking questions, 2016.

[15] Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. Learning from dialogue after deployment: Feed yourself, chatbot! In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 3667–3684, July 2019.

[16] Akira Taniguchi, Tadahiro Taniguchi, and Tetsunari Inamura. Spatial concept acquisition for a mobile robot that integrates self-localization and unsupervised word discovery from spoken sentences. *IEEE Transactions on Cognitive and Developmental Systems*, Vol. 8, No. 4, pp. 285–297, 2016.

[17] Jesse Thomason, Aishwarya Padmakumar, Jivko Sinapov, Nick Walker, Yuqian Jiang, Harel Yedidsion, Justin Hart, Peter Stone, and Raymond J. Mooney. Improving grounded natural language understanding through human-robot dialog. In *International Conference on Robotics and Automation (ICRA)*, pp. 6934–6941, 2019.

[18] Benjamin Kane, Felix Gervits, Matthias Scheutz, and Matthew Marge. A system for robot concept learning through situated dialogue. In *Proceedings of*

the Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 659–662, September 2022.

[19] Xinyu Li, Wenqing Ye, Yueyi Zhang, and Xiaoyan Sun. Grace: Gradient-based active learning with curriculum enhancement for multimodal sentiment analysis. In *Proceedings of the ACM International Conference on Multimedia*, MM '24, p. 5702–5711, 2024.

[20] Mohammed Abdelwahab and Carlos Busso. Active learning for speech emotion recognition using deep neural network. In *International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 1–7, 2019.

[21] Santiago A. Moreno-Acevedo, Juan Camilo Vasquez-Correa, Juan M. Martín-Doñas, and Aitor Álvarez. Stream-based active learning for speech emotion recognition via hybrid data selection and continuous learning. In *Text, Speech, and Dialogue*, pp. 105–117, 2024.

[22] Thus Karnjanapatchara, Sixia Li, Candy Olivia Mawalim, Kazunori Komatani, and Shogo Okada. Incremental multimodal sentiment analysis for hais based on multitask active learning with interannotator agreement. In *International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 72–79, 2024.

[23] Sarah Wassermann, Thibaut Cuvelier, and Pedro Casas. RAL - Improving Stream-Based Active Learning by Reinforcement Learning. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD) Workshop on Interactive Adaptive Learning (IAL)*, 2019.

[24] Kazunori Komatani, Shogo Okada, Haruto Nishimoto, Masahiro Araki, and Mikio Nakano. *Multimodal Dialogue Data Collection and Analysis of Annotation Disagreement*, pp. 201–213. Springer Singapore, 2021.

[25] Kazunori Komatani and Shogo Okada. Multimodal human-agent dialogue corpus with annotations at utterance and dialogue levels. In *International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 1–8, 2021.

[26] Shun Katada, Shogo Okada, and Kazunori Komatani. Effects of physiological signals in different types of multimodal sentiment estimation. *IEEE Transactions on Affective Computing*, Vol. 14, No. 3, pp. 2443–2457, 2023.

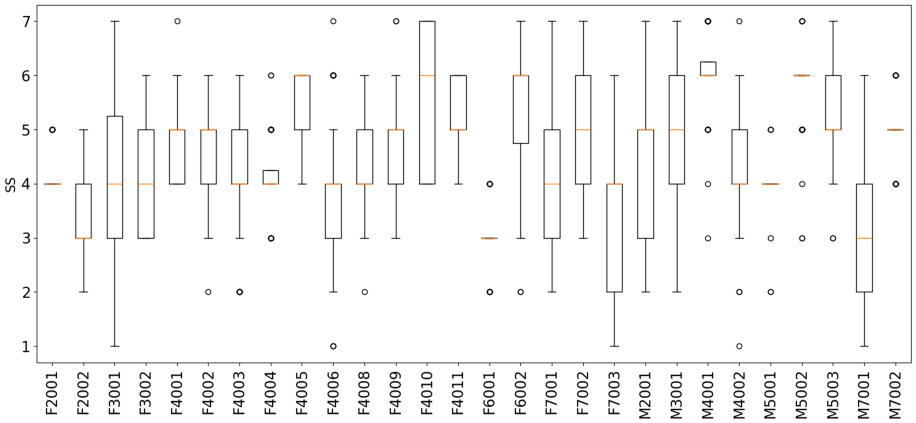
[27] Taku KUDO. Applying conditional random fields to japanese morphological analysis. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 230–237, 2004.

[28] Björn W. Schuller, Stefan Steidl, and Anton Batliner. The INTERSPEECH 2009 emotion challenge. In *the Annual Conference of the International Speech Communication Association*, pp. 312–315, 2009.

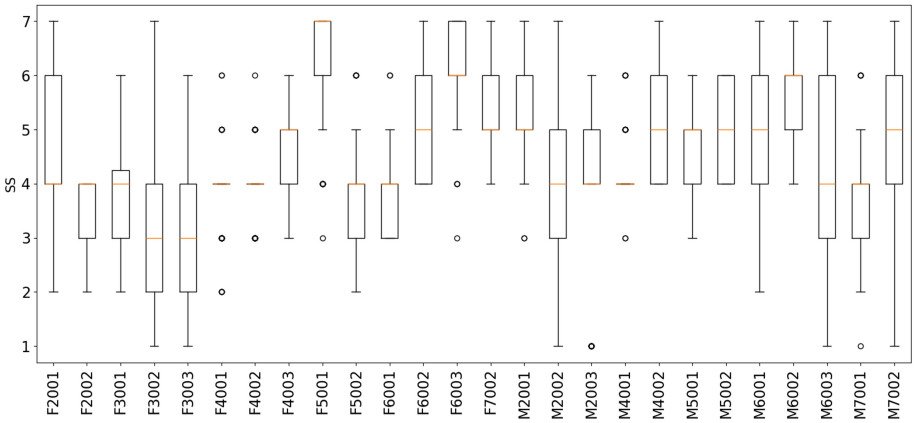
[29] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: An open source facial behavior analysis toolkit. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–10, 2016.

付 録A　データセットの分布

本付録ではデータセットの詳細な分布について記載する．1-7 の 7 クラスである主観的感情状態 SS の分布を図 A.1(訓練データ)，図 A.2(テストデータ) に示す．また，3 クラス変換後の被験者全体のラベル分布を図 A.3，被験者それぞれのラベル分布を図 A.4(訓練データ)，図 A.5(テストデータ) に示す．



(a)Hazumi1902



(b)Hazumi1911

図 A.1: 訓練データにおけるそれぞれの被験者での SS 分布．

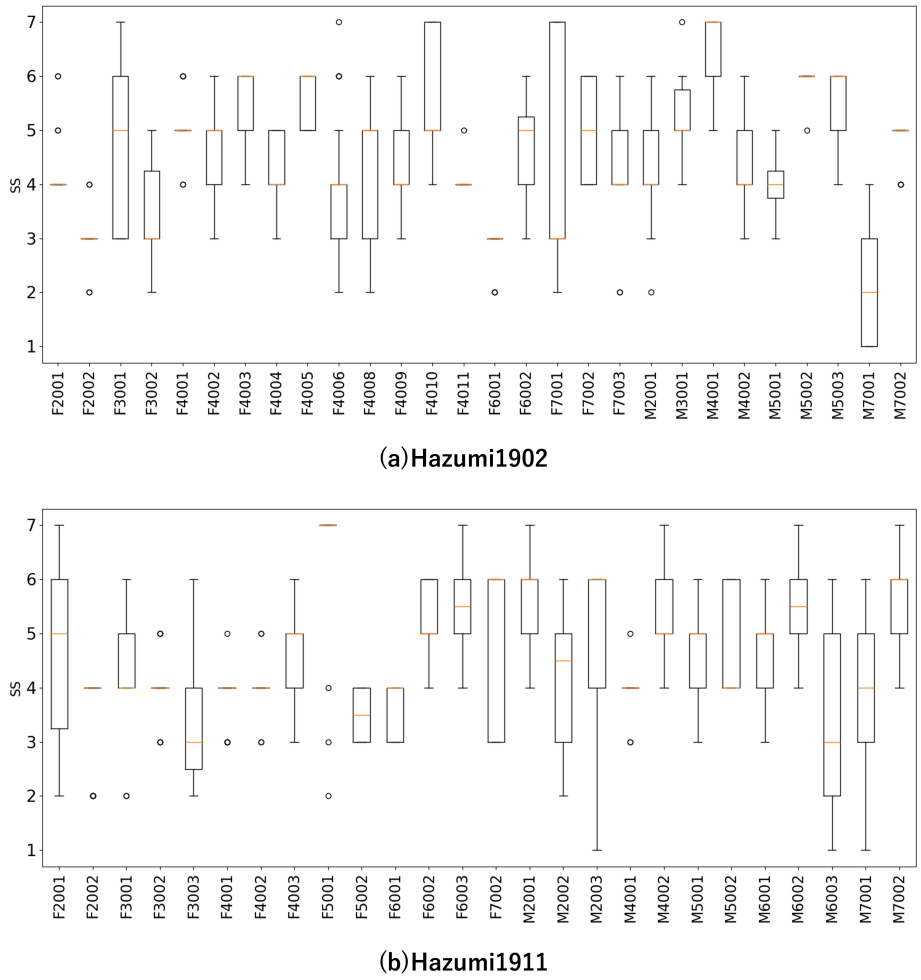


図 A.2: テストデータにおけるそれぞれの被験者での SS 分布.

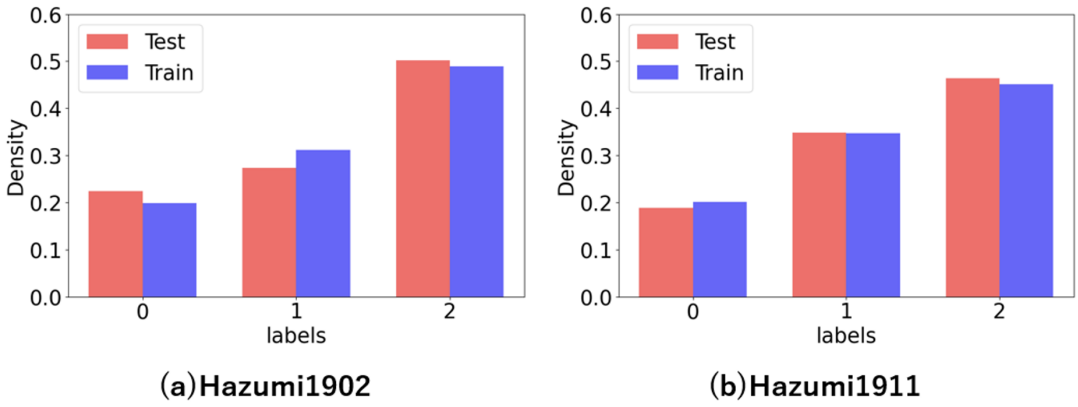
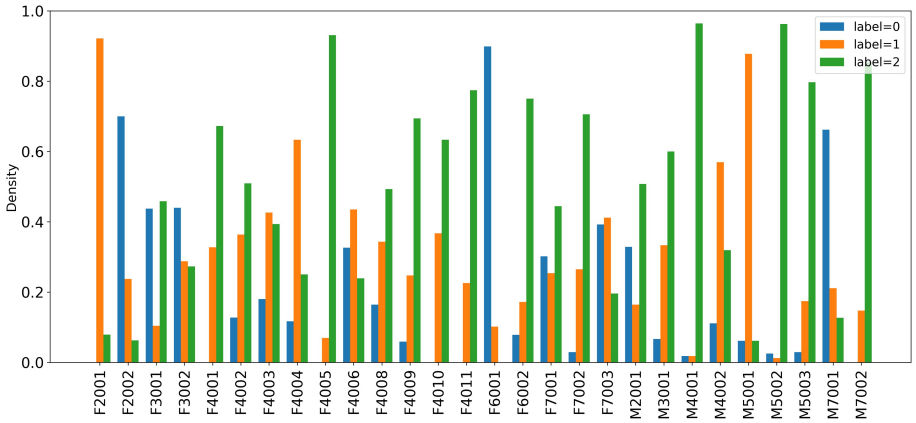
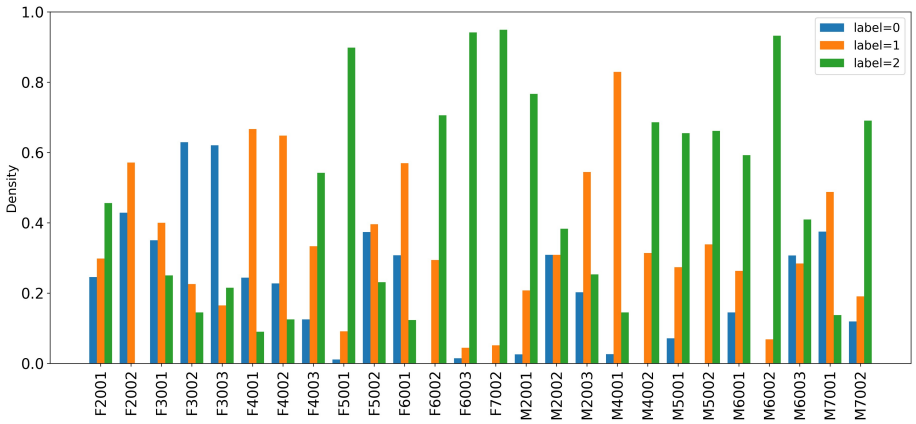


図 A.3: Hazumi1902 と Hazumi1911 のラベル分布.

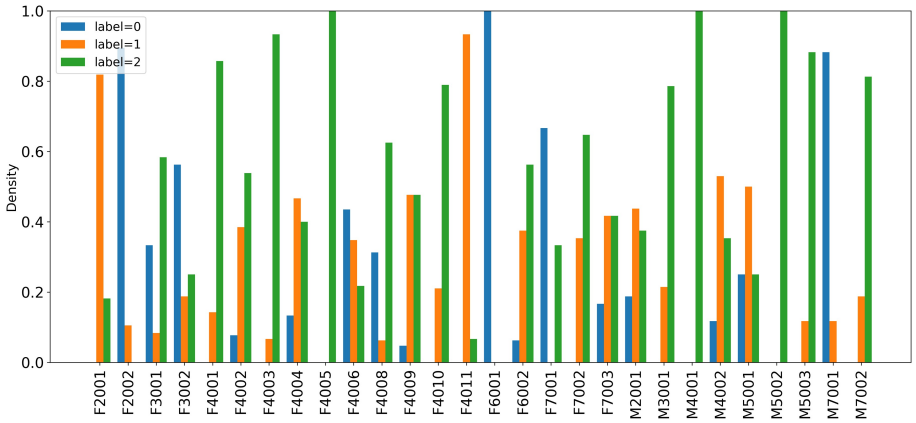


(a)Hazumi1902

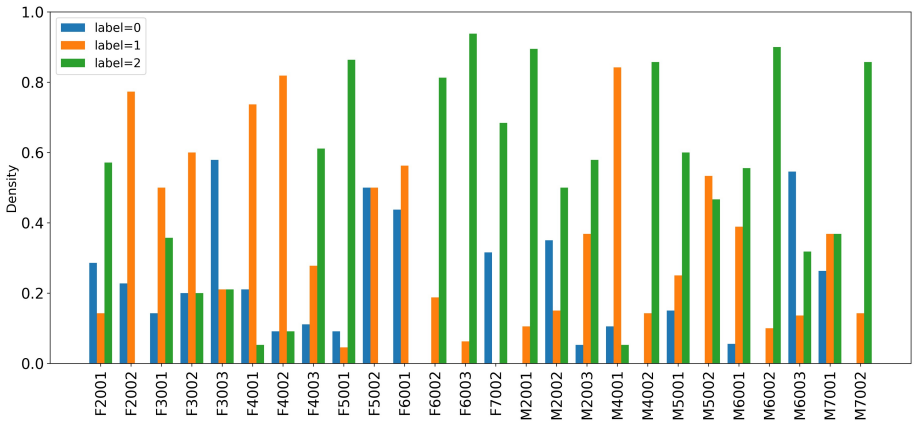


(b)Hazumi1911

図 A.4: 訓練データにおけるそれぞれの被験者でのラベル分布.



(a)Hazumi1902



(b)Hazumi1911

図 A.5: テストデータにおけるそれぞれの被験者でのラベル分布.

付 録B　特徴量重要度の分析

本節では，各被験者について 15 サンプルの追加学習用データを用いて学習を行った後 に得られた，各モダリティの推定器における特徴量重要度について分析する．特徴量重要度の算出には Permutation Importance を用いた．Permutation Importance は，各特徴量の値をランダムに置換した際の推定性能の低下量に基づいて重要度を評価する手法であり，モデルの予測性能に対する各特徴量の寄与度を直接的に反映できる点が特徴である．

Hazumi1902 における特徴量重要度を，言語特徴については図 B.1，音声特徴については図 B.2，映像特徴については図 B.3 に示す．同様に，Hazumi1911 における特徴量重要度を，言語特徴，音声特徴，映像特徴の順に，それぞれ図 B.4，図 B.5，図 B.6 に示す．各図において，赤色で大きく表示されたプロットはグループ内における特徴量重要度の平均値を表し，それ以外の色のプロットは グループ内の各被験者の値を示している．なお，表示している特徴量は，グループ内平均の重要度が高い上位 10 個に限定し，重要度の高い順に上から配置している．

分析の結果，両データセットに共通して，被験者ごとに特徴量重要度が大きくばらつく傾向が確認された．特に，Hazumi1902 における言語特徴および Hazumi1911 における映像特徴では，ある被験者において感情推定性能の向上に大きく寄与した特徴量が，他の被験者では逆に性能低下に寄与するといった，重要度の変動が顕著な特徴量が観察された．さらに，同一特徴量であってもグループ間で重要度の順位が変動する傾向がみられ，特徴量の有効性が被験者ごとに異なることが示唆された．

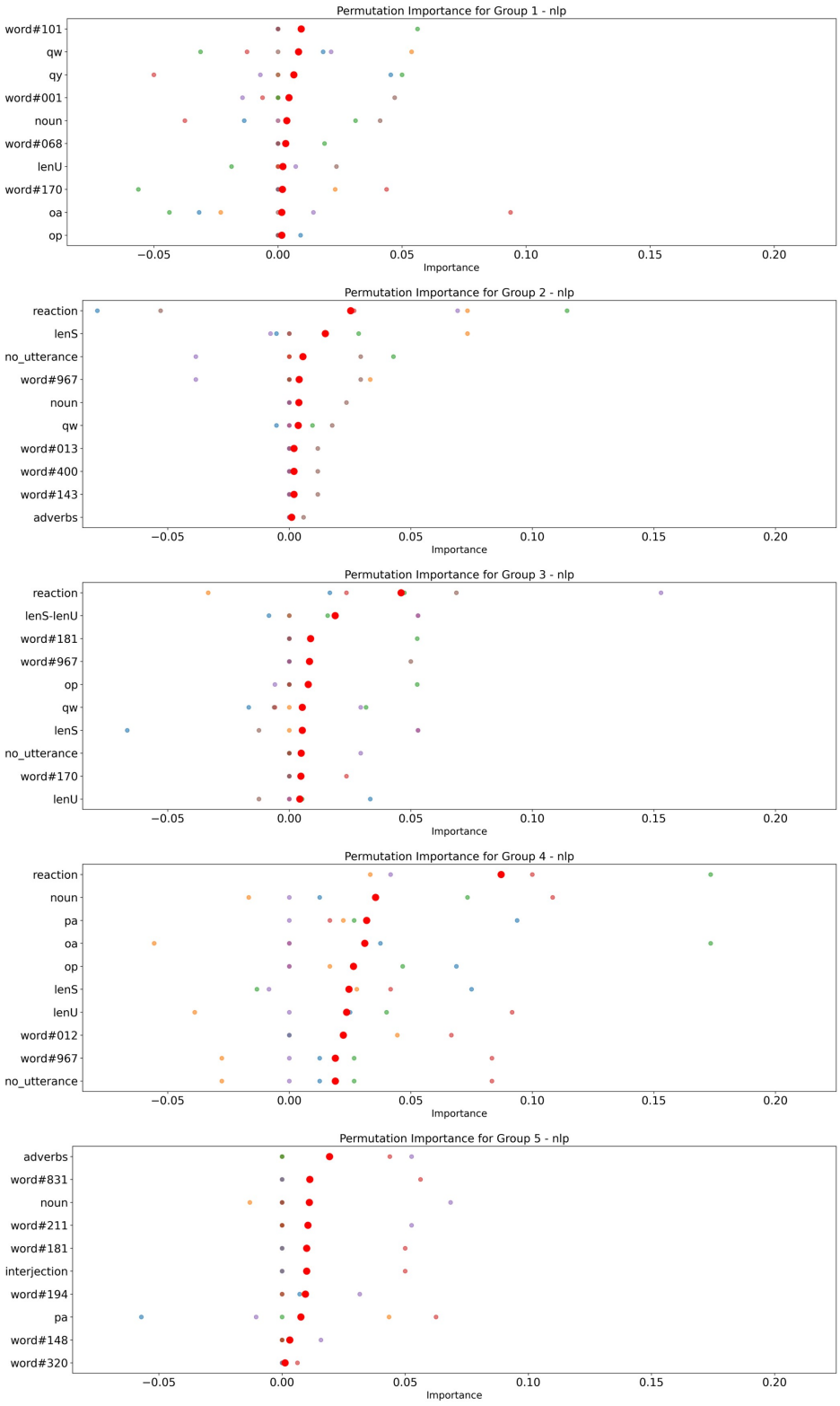


図 B.1: Hazumi1902 における言語特徴の重要度.

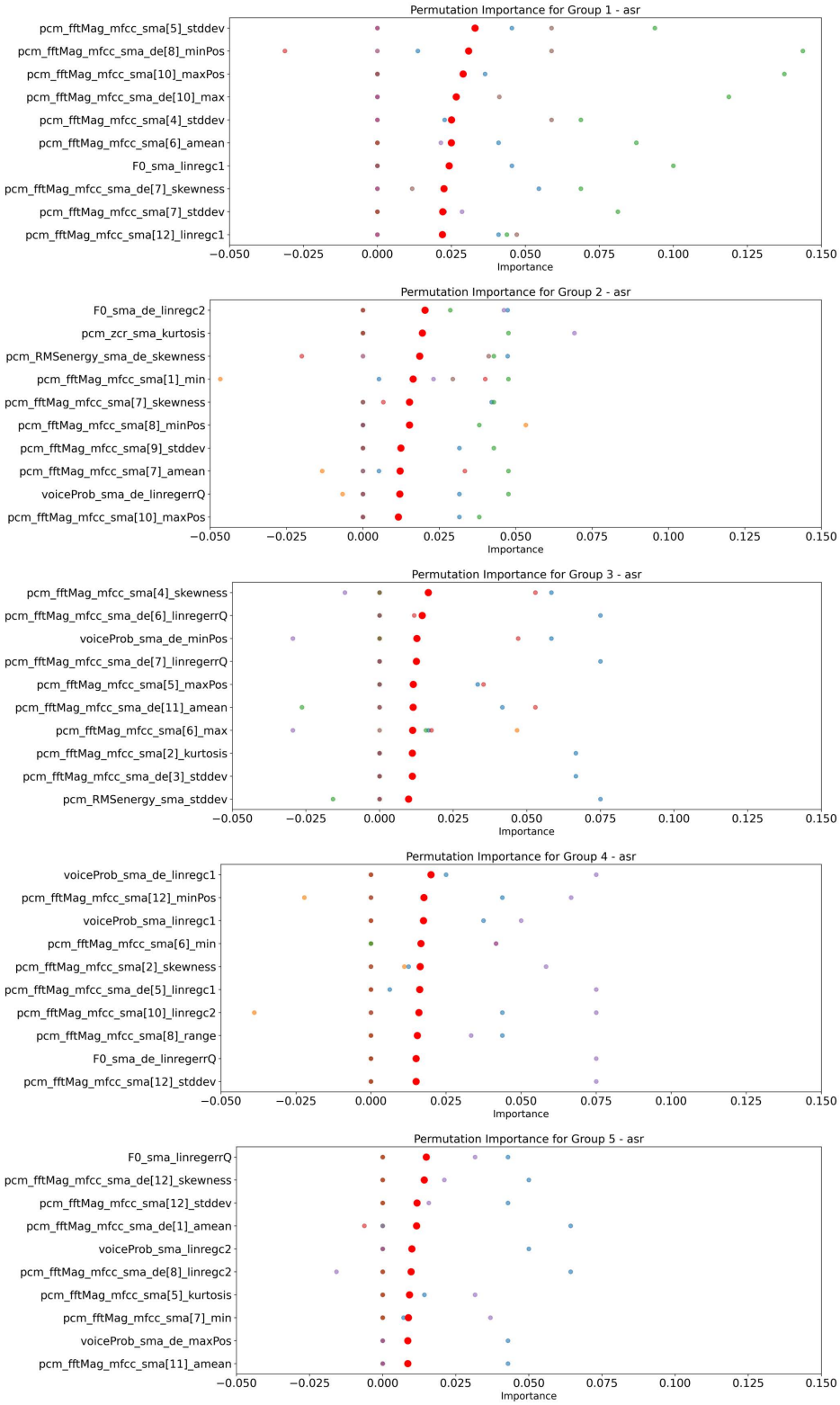


図 B.2: Hazumi1902 における音声特徴の重要度.

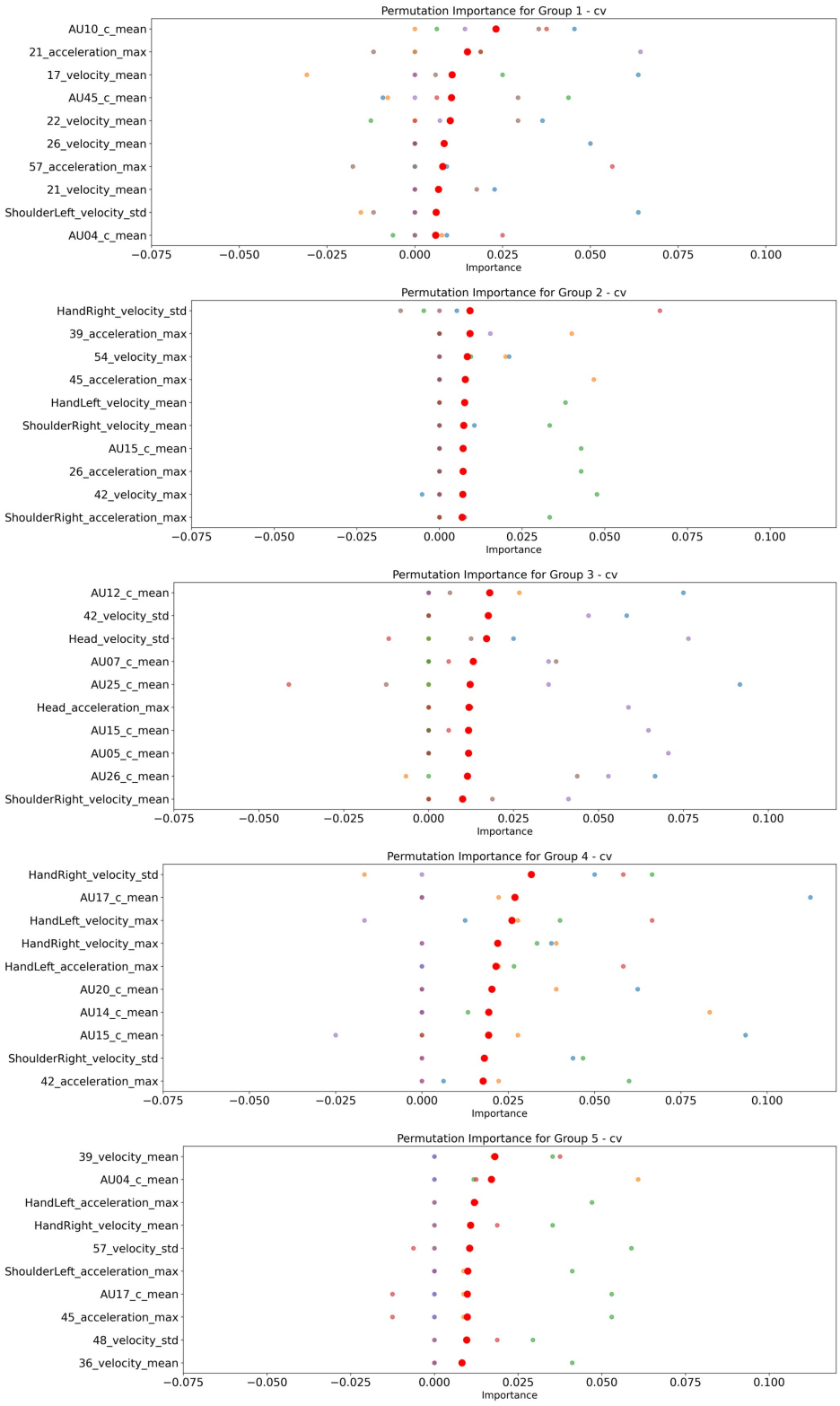


図 B.3: Hazumi1902 における映像特徴の重要度.

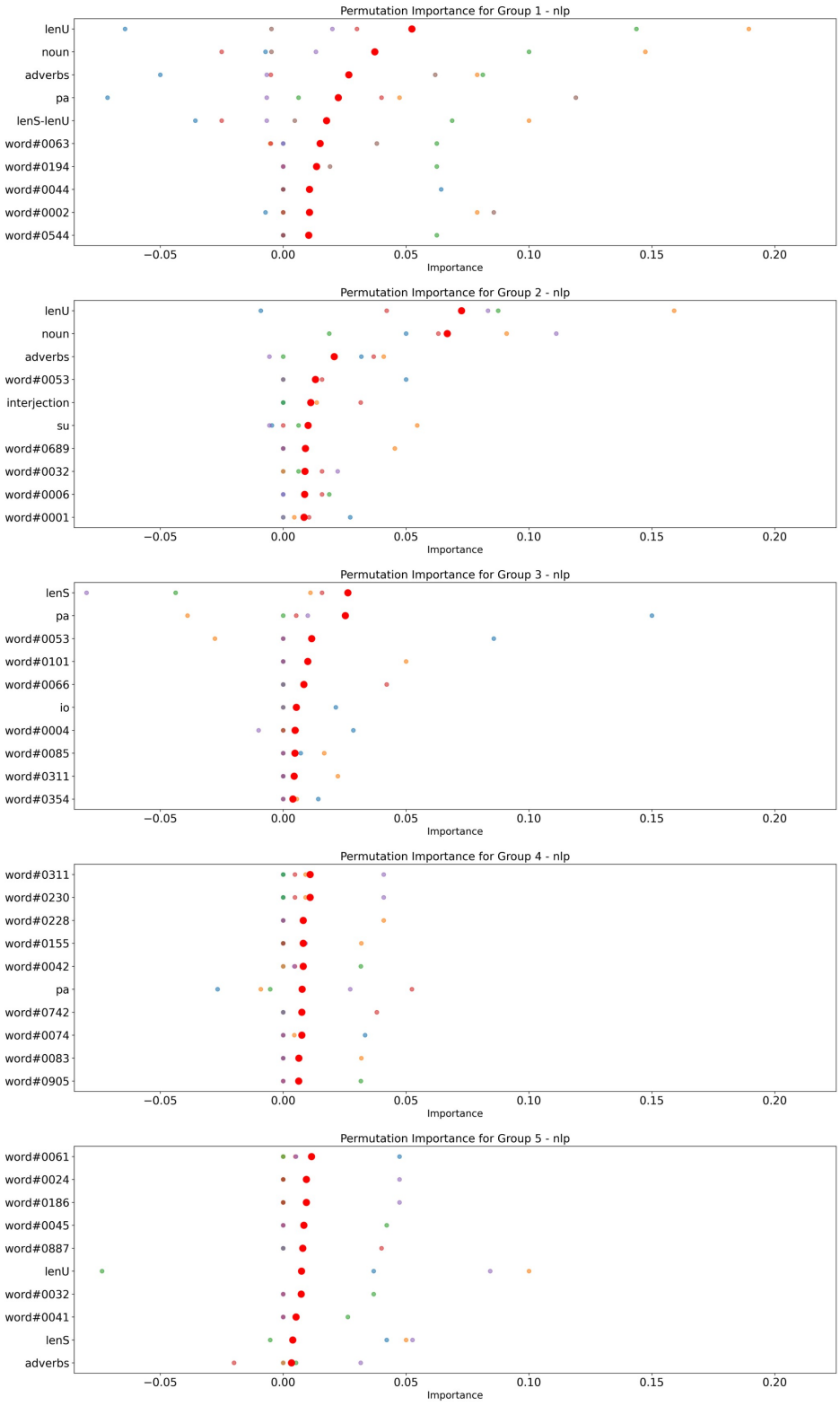


図 B.4: Hazumi1911 における言語特徴の重要度.

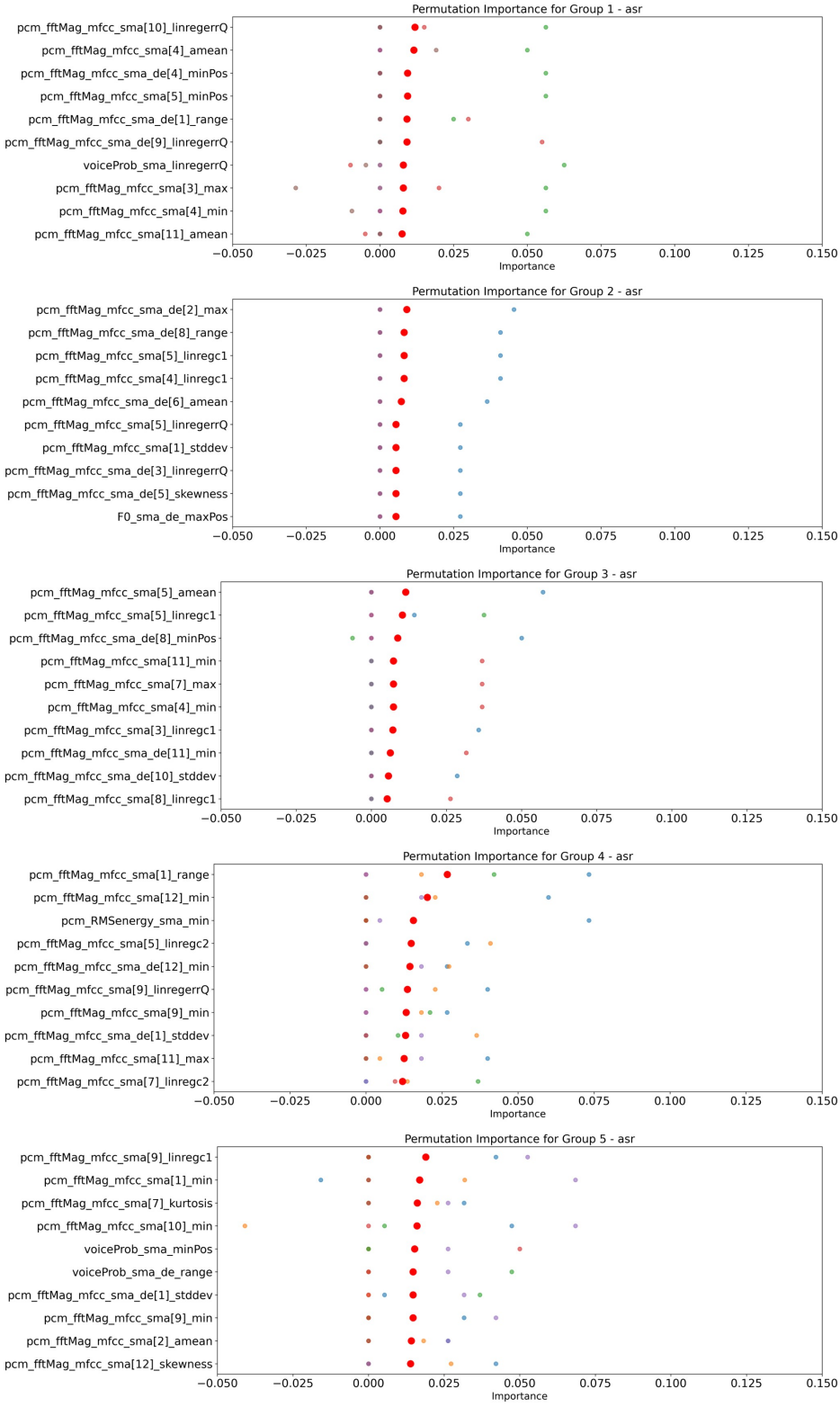


図 B.5: Hazumi1911 における音声特徴の重要度.

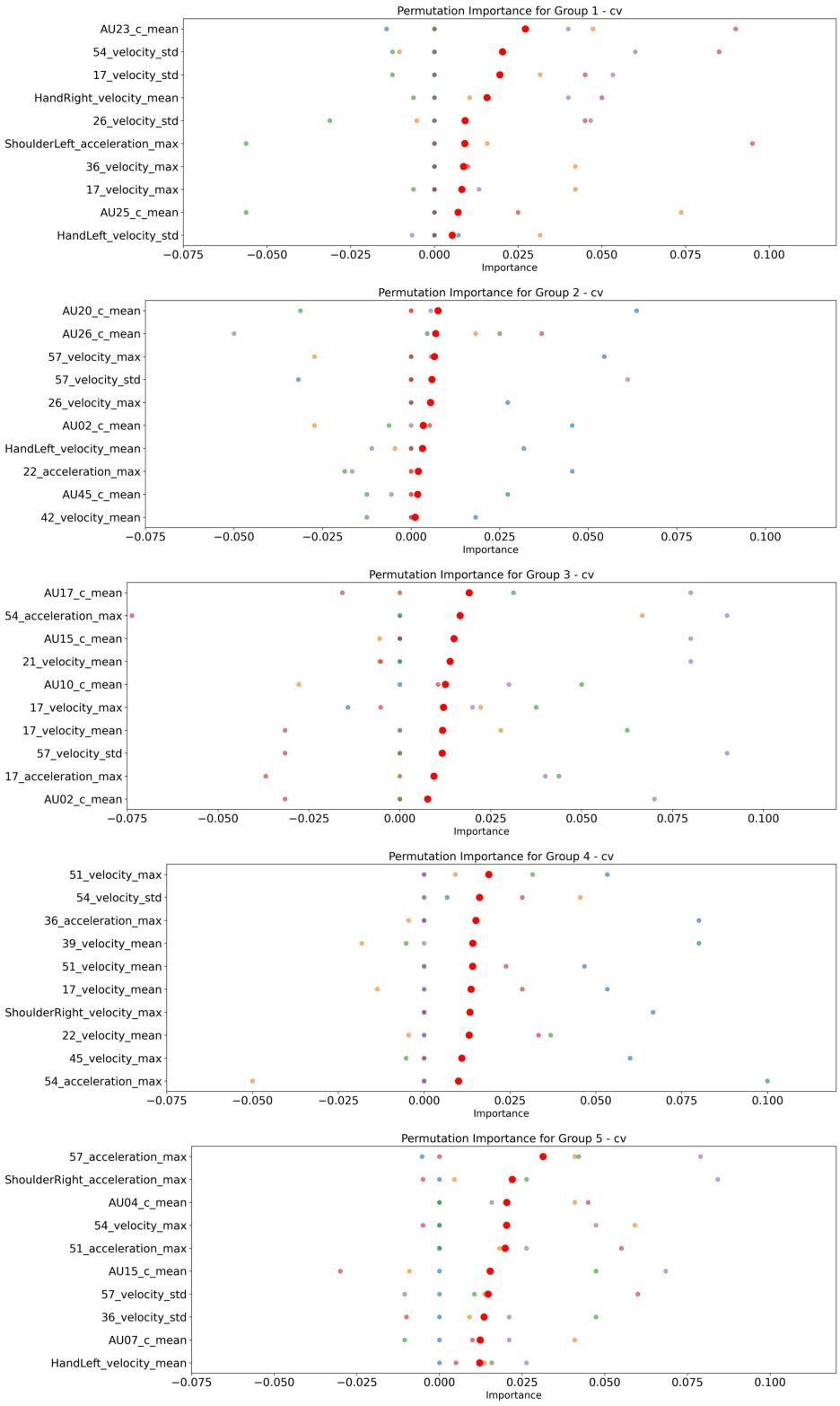


図 B.6: Hazumi1911 における映像特徴の重要度.