OUKA
The University of Osaka
Institutional Knowledge Archive

Title	共同注意が不完全なマルチモーダル環境における主観的整合性に基づく対応学習とカテゴリ化
Author(s)	笹本, 勇輝
Citation	大阪大学, 2015, 博士論文
Version Type	VoR
URL	https://doi.org/10.18910/53992
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

https://ir.library.osaka-u.ac.jp/

The University of Osaka

博士学位論文

共同注意が不完全なマルチモーダル環境における主観的整合性に基づく対応学習とカテゴリ化

笹本 勇輝

2015年 4月

大阪大学大学院工学研究科

タイトル: 共同注意が不完全なマルチモーダル環境における
主観的整合性に基づく対応学習とカテゴリ化

主査: 浅田 稔　教授
副査: 石黒 浩　教授
細田 耕　教授
吉川 雄一郎　准教授

Copyright © 2015 by 笹本 勇輝
All Rights Reserved.

要旨

近年のロボット技術の発展に伴い，人とロボットが共存する社会の実現に期待がよせられている．これに対し，ロボットがユーザとの相互作用を通じて得たマルチモーダルな情報をカテゴリ化あるいは対応付けることでユーザの振舞や環境中の物体を自律学習する仕組みがいくつか提案されている．しかし，それらの手法では，ユーザとロボットが注意を共有し，ロボットが観測するマルチモーダルな情報が特定の事物を表すことが想定されており，その想定が崩れる際にロボットがどのように学習を進めるべきかには焦点は当てられていなかった．人の生活環境などの実環境においては，ユーザが常にロボットに注意を向けるとは限らない，すなわち共同注意が不完全な場合が起こりうる．そのような場合，ロボットが観測したマルチモーダルな情報すべてが必ずしも特定の事物を表すとは限らない．そこで本論文では，ユーザとロボットの共同注意が不完全なマルチモーダル環境におけるロボットの学習課題を扱い，そのような環境でもロボットが自律的にカテゴリ化や対応学習が可能な手法の構築に取り組む．

共同注意が不完全なマルチモーダル環境における問題は，ロボットの観測に，特定の事物とは対応しないデータが含まれることである．例えば，人がロボットが見ている物体の特定に失敗する場合，ロボットが見ている物体と聴取した人の発話音声は，それぞれ別の事物を表す．このような場合，それら観測を単純に学習に利用する手法では，誤ったカテゴリや対応関係を学習してしまう可能性がある．一方，人の乳児は，必ずしも注意を共有しているとは限らない親とのやり取りを通じて，すなわち，共同注意が不完全なマルチモーダル環境において音韻や物体などのカテゴリ化や模倣などの対応学習を達成していると考えられ，乳児の発達を参考にした学習メカニズムをロボットに実装できれば，上記の問題を解決できると考えられる．そこで本論文では，乳児がその発達過程で示す整合性に基づく学習戦略に着想を得た主観的整合性の概念

を導入し，これを用いたカテゴリ化および対応学習手法を提案した．

ロボットのカテゴリ化と対応学習の典型的な課題に対して主観的整合性のアイデアを適用し，その有効性を確認した．具体的には，まず，共同注意が不完全な対応学習の課題として，教示者が必ずしも学習者を模倣するとは限らない状況における音声模倣学習に対し，主観的整合性に基づく語彙学習との相補的対応学習の手法を提案した．実ロボットを用いた実験および計算機シミュレーションにより，提案手法を用いることで，教示者がロボットを模倣しない場合であっても，音声模倣に必要な対応学習が可能であることを示した．次に，カテゴリ化を扱った課題として，ロボットとユーザが必ずしも共同注意しているとは限らない状況における語意カテゴリ学習に対し，主観的整合性を導入したマルチモーダルカテゴリゼーションの手法を提案した．実ロボットおよび人工データを用いた実験により，提案手法を用いることで，ユーザがロボットの意図の汲み取りに失敗し，注意を共有できなかった場合でも，ロボットが語意カテゴリを学習可能であることを示した．

謝辞

本研究を進めるにあたり，多くの方々の御支援を賜りました．ここに感謝の意を示します．まず，指導教員である大阪大学 大学院工学研究科 知能・機能創成工学専攻 浅田稔教授に感謝致します．修士の頃から多大なる御指導，御鞭撻を賜りました．ここに深く感謝致します．

御多忙の中，博士論文の副査をして下さった，大阪大学 大学院基礎工学研究科 システム創成専攻 石黒浩教授，大阪大学 大学院基礎工学研究科 システム創成専攻 細田耕教授，大阪大学 大学院基礎工学研究科 システム創成専攻 吉川雄一郎准教授に感謝致します．石黒教授には，修士の頃の研究から気にかけていただき，多くの御助言をいただきました．細田教授には，研究への御助言以外でも，細田教授自身の研究に対する姿勢から，研究生活の送り方，研究の楽しみ方など学ばせていただきました．吉川准教授には，修士の頃から研究に関して手取り足取り教えていただき，論文に至っては一字一句御指導いただきました．本論文を執筆することができたのも，吉川准教授の御指導，御鞭撻あってこそです．心より感謝致します．

大阪大学 大学院工学研究科 知能・機能創成工学専攻 杉原知道准教授に感謝致します．研究を遂行する上で常に勉強し続ける姿勢やその大切さなど多くのことを学ばせていただきました．また，大阪大学 大学院工学研究科知能・機能創成工学専攻 長井志江特任准教授，関西大学 総合情報学部 大学院総合情報学研究科 荻野正樹教授に感謝いたします．研究室の先輩でもあるお二人からは，日頃より本研究に関して的確な御助言をいただきました．

(株) 国際電気通信基礎研究所 住岡英信氏，(株) 富士通研究所 三浦勝司氏，大阪大学 大学院工学研究科 知能・機能創成工学専攻 石原尚助教，大和ハウス工業 (株) 中野吏氏に感謝致します．同じ研究グループで数々の有益な議論をして頂き，多くの御

協力及び御助言をいただきました.

大阪大学 大学院工学研究科 知能・機能創成工学専攻 浅田研究室の学生各位，特に堀井隆斗氏，河合祐司氏に感謝致します．本研究を遂行するために，時には深夜まで密に議論してくださりました.

最後に，自由気ままに研究生活を送っていた私を温かく見守り，精神的・経済的に支えてくれた両親に深い感謝の意を示します.

*2015*年 *03*月 *03*日

笹本 勇輝

目 次

第 1 章　緒論 **1**

1.1　背景 1

1.2　従来研究 2

1.2.1　ロボットのカテゴリ化に関する研究 2

1.2.2　ロボットの対応学習に関する研究 3

1.2.3　従来研究における問題点 4

1.3　本論文の目的と内容 5

第 2 章　主観的整合性の概念 **7**

2.1　乳児のカテゴリ化と対応学習 7

2.2　人の認知や脳処理における整合機構 9

2.3　主観的整合性の基本アイデア 10

第 3 章　音声模倣と語彙獲得のための主観的整合性に基づく相補的対応学習 **13**

3.1　はじめに 13

3.2　乳幼児の認知発達と主観的整合性による共発達 16

3.3　音声模倣と語彙獲得の共発達メカニズム 17

3.3.1　問題設定 17

3.3.2　観測信号ベクトル 20

3.3.3　マッピングの学習 21

3.3.4　主観的整合機構：複数信号の整合性に基づく統合 23

3.4　実ロボットを用いた実験 25

3.4.1　実験設定 25

3.4.2 実験結果 29
3.5 シミュレーション実験 32
3.5.1 基本設定 32
3.5.2 予備実験: 主観的整合機構による誤った対応学習の抑制効果の確認 35
3.5.3 実験 1: 主観的整合機構による促進効果の検証 36
3.5.4 実験 2: 実環境での養育者の応答を模した状況での検証 41
3.6 考察 42
3.6.1 学習手法としての位置づけ 42
3.6.2 乳児の発達との関連性 45

第 4 章 語意のカテゴリ化のための主観的整合性に基づくマルチモーダルカテゴリゼーション 49

4.1 はじめに 49
4.2 語意学習状況 51
4.3 主観的整合性に基づくマルチモーダルカテゴリゼーション 54
4.3.1 基本的な学習手法: Multimodal LDA 54
4.3.2 主観的整合性に基づく拡張 57
4.4 実験 60
4.4.1 実験 1：実データを用いた検証 60
4.4.2 実験 2: 人工データを用いた詳細な検証 67
4.5 考察 76
4.5.1 ロボットの語意学習手法としての位置づけ 76
4.5.2 提案手法のスケーラビリティ 77

第 5 章 結論 81

関連図書 85

表 目 次

3.1 “Synchy” hardware specifications 27
3.2 Phoneme data for simulation 34
3.3 Object data for simulation 35

図 目 次

3.1 An example scene when the learning for vocal imitation is facilitated by acquired lexicon 15
3.2 Assumed environment of caregiver-robot inreraction 18
3.3 Mutually associated multimodal mapping model 20
3.4 Notations for learning rule 21
3.5 Humanoid robot “Synchy” 26
3.6 A sample scene for human-robot interaction 28
3.7 Behaviors of robot 30
3.8 Transitions of learning performance of (a) Articulation-Sound Mapping, (c) Sound-Word one, and (e) Word-Articulation one without the proposed subjective consistency, in turn (b) (d) (f) with it 31
3.9 Observation probability with respec to combinations of three input variables (a) without integrated signal generated by proposed mechanism and (b) with it 37
3.10 Average probability of predicting corresponding vector by acquired mappings until 100,000 step with respect to dependency on subjective consistency(η) and maternal teaching rate (p_a) 39
3.11 Average probability of predicting corresponding vector by acquired mappings until 100,000 step with respect to low maternal teaching rate (p_a) under the proposed method ($\eta = 1.0$) 40

3.12 Average of average subjective consistencies among different three mappings during final 100 steps of learning with respect to maternal teaching rate p_a: subjective consistency for external signal (filled squares with solid line), direct prediction (blank triagle with broken line), and indirect prediction (blank diamond with dash-dotted line). 41
3.13 Average transitions of learning performance of (a) articulation-sound mapping, (b) sound-word one, and (c) word-articulation one with the proposed subjective consistency ($\eta = 1.0$) and without it ($\eta = 0.0$) 43
3.14 Average transition of subjective consistency calculating in learning (a) articulation-sound mapping, (b) sound-word one , (c) word-articulation one: subjective consistency for external signal (filled squares with solid line), direct prediction (blank triagle with broken line), and indirect prediction (blank diamond with dash-dotted line) 44
4.1 Assumed environment of robot that learns words from human 52
4.2 An example of caluculation process of feature vectors 53
4.3 Graphical models of (a) previous method and (b) proposed method for the word learning of robot. Some notations in (a) are changed from original ones. 55
4.4 The environment in experiment 1 61
4.5 Examples of image data of learning target and non-learning target . 62
4.6 Average F-value in 10 steps with respect to corresponding rate P_c . . 65
4.7 Input distributions of artificial data in experiment 2 66
4.8 Average performance of clacssification in 100 steps with respect to correponding rare P_c 67
4.9 An example of feature distribution of each category formed under the $P_c = 1.0$ 70
4.10 Transitions of sampled categories in learning process (a) without proposed subjective consistency and (b) with it under $P_c = 0.5$ 71

4.11 The final subjective consistency with respect to corresponding rate . . 73
4.12 Average performance of clacssification in 100 steps with respect to correponding rare P_c (a) with data in modalities of v_o and s, and (b) with ones in v_s 75
4.13 Transitions of subjective consistencies of data in each modality under $P_c = 0.5$ 76

第1章　緒論

1.1　背景

近年のロボット技術の発展に伴い，従来は産業分野に閉じていたロボットの活動範囲が人の生活環境へと広がりつつある．実際に，エンターテイメント用あるいは家庭内でのパートナー用として様々なロボットが開発され[1; 2; 3; 4]，人とロボットが共存する社会の実現への期待も高まっている．しかしながら，それらロボットの多くは，特定の状況でのみ動作するようにロボットの振舞が設計されている．そのため，想定内の状況では適切に動作するが，そうでない場合は動作できない．特に，AIBO などのコミュニケーションロボットにおいては，動作パターンが単調であるために，人が飽きてしまい，本来の目的である人とのコミュニケーションを継続できないなどの問題に繋がる．様々な状況が起こりうる人の生活環境においてロボットが人とコミュニケーションし適切に動作できるためには，ロボットは人の振舞を観察しそれに示されている指示を理解できる必要があると考えられる．また，ロボットが観察する人の振舞やその指示は，対面するユーザあるいは環境に存在する事物などに応じて異なることが予想されるため，ロボットは自律的にそれらを学習できることが望ましい．そのようにロボットにユーザの振舞や指示対象となる環境中の事物を自律学習させる仕組みに関していくつか研究があるが，基本的には以下の問題が扱われている．

- 音声や画像など，ロボットが取得した様々なセンサ情報を，どのように人の特定の発話や動作，あるいは物体として認識させるかのカテゴリ化の問題
- 人の発話とロボットの発話あるいは人の動作と同様のロボットの動作など，観測した人の振舞とロボット自身の振舞をどのように対応付けるかの対応学習の問題

上記問題に対して，ロボットがユーザとの相互作用を通じて得られるマルチモーダルな情報に基づいて振舞や環境中の事物を自律学習する仕組みが提案されている．例えば，ユーザに提示された物体について，ロボットが同時に観測した視覚や聴覚，触覚などの複数のセンサ情報をカテゴリ化することで物体カテゴリを学習する仕組み[5; 6]やロボットの構音運動と，それと同期して観測された人の発話を対応づける対応学習の仕組みなどが提案されている[7; 8]．しかしながら，それらの研究では，ユーザがロボットが注目している物体についての提示や発話を行うあるいはロボットが行った動作を真似るなど，ロボットが観測するマルチモーダルな情報がロボットが注目している特定の事物を捉えていることが想定されており，この想定が崩れる際に，どのようにロボットが学習を進めるすべきかには焦点はあてられていない．実環境においては，常にユーザとロボットが注意を共有しているとは限らない．すなわち，ロボットがユーザとの相互作用を通じて観測したマルチモーダルな情報すべてが必ずしもロボットが注目している事物を表すとは限らない．従って，ロボットは，ユーザとの相互作用から得られた情報すべてを単純に学習に利用するのではなく，学習に必要な情報を取捨選択できることが望ましい．本論文では，そのように必ずしもユーザとロボットが同じ事物に注目しているとは限らない状況，すなわち，共同注意が不完全なマルチモーダル環境におけるロボットの自律学習の実現を目指す．

1.2 従来研究

ロボットがユーザとの相互作用を通じて自律的に振舞や環境中の事物を学習する仕組みがいくつか提案されている．ここでは，主に，カテゴリ化と対応学習の枠組みに分けてそれらの研究を概観する．

1.2.1 ロボットのカテゴリ化に関する研究

ロボットが，人の指示を理解する上で，人の発話の意味（それが示す事物）を特定できることは重要であり，ロボットが取得した音声や画像などのセンサ情報を一つの

物体や語として認識できるようになる語意カテゴリ学習は，その典型的な課題であると考えられる．これに対して，Roy and Pentland は，カメラから取得した物体画像とマイクから聴取した発話音声の相互情報量に基づいて語意カテゴリを学習するシステム，CELL を提案している[9]．Yu は，事前に物体画像のみから物体プロトタイプを学習し，その物体プロトタイプに関して聴取した単語情報に基づいて物体プロトタイプをカテゴリ化することで語意を学習する手法を提案している[10]．これらの手法により，ロボットは，同時に観測した音声や画像などの視聴覚情報を，一つの語意としてカテゴリ化することができる．一方，視聴覚以外の情報を用いた手法がいくつか提案されている．田口らは，発話音声と指示対象に加え，文法情報である単語の並びを考慮した語意学習手法を提案している．この手法では，発話とその指示対象の関係が，音響，意味，文法情報の共起確率として表現され，MDL 原理に基づいて，少ない単語列で記述できるように，それら関係が学習される．これにより，連続音声からの単語の抽出と指示対象との関係（意味）の学習が可能である[11]．Nakamura et al. は，事前に，視覚，聴覚，触覚情報の共起確率からそれらの元となる潜在変数として指示対象（物体）のカテゴリを学習し，その後，指示対象とその名称として聴取した単語情報の共起確率からそれらを分類して，語意を学習する Multimodal LDA を提案している[12]．

上記一連の研究で提案された手法により，ロボットは，複数のセンサから得られるマルチモーダルな情報の共起確率に基づいてカテゴリを学習できる．例えば，ロボットは，ある物体を見たり振ったりした時に観測した視覚（物体の見た目），聴覚（物体が発する音），触覚（物体表面の手触り）情報とその物体について人が発した音声情報（名称）の共起確率に基づいて，それらを分類することで発話とそれが示す物体の関係（意味）として，語意カテゴリを学習できる．

1.2.2 ロボットの対応学習に関する研究

ロボットが，人とコミュニケーションするためには，人の振舞を理解し，それと同様の振舞を表出できる必要がある．これに対して，ロボットが人の行動と自身の行

動の対応関係を学習する模倣学習を扱った研究がいくつかある．Alissandrakis et al. は，ロボット (エージェント) に自身と他者の身体部位間の対応を記述したマップを与え，そのマップに基づき対応する身体部位間での状態（関節角度）の誤差を最小化するように動作することで，ロボットが他者とは異なる身体構造であっても他者の動きを模倣できることを示している[13]．Hafner and Kaplan は，ロボットが動作（歩行）し，それと同時に教示者役のロボットが同じ動作を実演する場合に，その時の同期性 (カップリング) に基づいて身体部位間の対応関係を学習できる仕組みを提案している [14]．音声の模倣学習を扱った研究もいくつかあり，聴覚フィードバックを利用した音声と構音運動の対応付け[15; 16]，あるいは，教示者役の人との相互模倣を通じて，明瞭な発声を学習する仕組み[7; 8]が提案されている．

上記一連の研究により，ロボットは同期性に基づいて，すなわち，自身の行動とそれと同時に観測した人の行動を対応付けることで，それらの対応関係を学習することができる．

1.2.3 従来研究における問題点

これまでの研究により，ロボットが教示者役の人やエージェントとの相互作用を通じて得られる，物の見えや動き，発声などのマルチモーダルな情報に基づいて振舞や環境中の事物を自律学習できる仕組みが提案されている．しかしながら，カテゴリ化の従来研究においては，ロボットが同時に観測した視覚や聴覚，触覚などの複数のセンサ情報がすべて特定の事物を表すものであることが想定されている．また，対応学習の従来研究においては，ロボットが振舞を学習する上で，どこ (どの身体部位) に注目すれば良いかを予め知っている，あるいは，教示者がロボットの動作を真似することが想定されている．すなわち，これまでの研究では，ユーザとロボットが共同注意し，ロボットが観測する複数の情報が特定の事物を表す環境でのロボットの学習が扱われており，その想定が崩れる際に，どのようにロボットが学習を進めるすべきかには焦点はあてられていない．実環境においては，常にユーザとロボットが注意を共有しているとは限らない．すなわち，ロボットがユーザとの相互作用を通じて観測し

たマルチモーダルな情報すべてが必ずしもロボットが注目している事物を表すとは限らない．例えば，人がロボットが見ている物体あるいは行っている行動の特定に失敗し対応しない物体の提示や行動表出を行う，あるいは，ロボットの見ている物とは関係なく呼びかけや指示を行う場合などには，ロボットが観測したマルチモーダル情報の中には，ある特定の事物とは対応しない情報も含まれることになる．このような場合，観測の共起確率や同期性のみに基づいて学習するだけでは，誤ったカテゴリや対応関係を学習してしまう可能性がある．そのため，ロボットは，ユーザとの相互作用から得られたデータすべてを単純に学習に利用するのではなく，学習に必要なデータを取捨選択できることが望ましい．

1.3 本論文の目的と内容

本論文では，ユーザとロボットが必ずしも同じ事物に注目しているとは限らない状況，すなわち共同注意が不完全なマルチモーダル環境においてロボットが自律的に振舞や環境中の事物を学習する仕組みを提案する．共同注意が不完全なマルチモーダル環境における問題は，ロボットの観測に，ロボットが注目している事物とは対応しないデータが含まれることである．例えば，人がロボットが見ている物体の特定に失敗する場合，ロボットが見ている物体と聴取した人の発話音声は，それぞれ別の事物を表す．このような場合，それら観測を単純に分類するあるいは対応付けるだけでは，誤ったカテゴリや対応関係を学習してしまう可能性がある．

これに対して，人の乳児の発達が参考になると考えられる．人の乳児は，親との相互やりとりを通じて，音韻や模倣などのコミュニケーション能力を発達させていることが示唆されている[17; 18]．また，親は乳児と常に注意を共有しているわけではないため，その振舞は，必ずしも乳児の意図を汲み取ったものではないと考えられる．実際に，乳児に対する親の振舞を観察した研究により，親は乳児に対して物体の提示や模倣といった教示的な振舞以外の振舞も行うことが示されている[19; 20]．従って，乳児は，必ずしも注意を共有しているとは限らない親とのやり取りを通じて，すなわち，共同注意が不完全なマルチモーダル環境において音韻などのカテゴリ化や模倣な

どの対応学習を達成していると考えられ，乳児の発達を参考にした学習メカニズムをロボットに実装できれば，上記の問題を解決できると考えられる．

発達心理学の知見により，乳児が観測刺激の一貫性に基づいてカテゴリを形成すること[21; 22] や，既に獲得した経験を利用して対応学習すること[23] が示唆されている．本論文では，そのような乳児の学習戦略に着想を得て，マルチモダールな観測のモダリティ間での整合性を評価することで学習に利用すべきモダリティあるいはデータを決定する主観的整合性の概念を導入し，これを用いることで，共同注意が不完全なマルチモーダル環境においてもロボットがカテゴリ化および対応学習可能であることを示す．

本論文は，5 つの章から構成される．1 章では，背景と従来研究の問題，本論文の目的について述べた．以下では，まず，2 章において，人の発達の知見とそれから着想を得た主観的整合性の基本アイデアについて説明する．3，4 章では，ロボットの対応学習とカテゴリ化の典型的な課題に対して主観的整合性のアイデアを適用する．具体的には，3 章では，教示者が必ずしも学習者を模倣するとは限らない状況における音声模倣学習の課題に対して，主観的整合性に基づく語彙獲得との相補的対応学習の手法を提案する．4 章では，ユーザとロボットが必ずしも共同注意しているとは限らない状況における語意カテゴリ学習の課題に対して，主観的整合性を導入したマルチモーダルカテゴリゼーションの手法を提案する．最後に，5 章においてまとめと今後の展望について述べる．

第2章　主観的整合性の概念

本章では，乳児のカテゴリ化や対応学習に関する発達心理学の知見を紹介し，それら学習過程における観測刺激の一貫性や既に獲得した経験を利用した乳児の学習戦略について説明する．そのような戦略と同様の処理が，大人の認知や脳処理においても見られることが社会心理学や認知科学，脳科学の研究からも示唆され，それらの知見についても紹介する．その後，それら知見から着想を得た主観的整合性のアイデアについて説明する．

2.1　乳児のカテゴリ化と対応学習

乳児は発達初期から，感覚刺激に基づいてカテゴリを形成できることが報告されている．3ヶ月児は，提示された一連の物体や聴取した音声から，動物カテゴリ[24]や母音カテゴリ[25]，すなわち物体や音韻のカテゴリを形成することができる．乳児は，知覚した刺激の統計的性質に敏感[26]であり，これに基づいて刺激間の統計的類似性を判断しカテゴリを形成していると考えられてきた[27; 28; 29]．しかし，最近の研究により，乳児はそのように単一モダリティにおける刺激の統計情報だけでなく，複数のモダリティにおいて提示された刺激の一貫性を利用して，個々のモダリティのカテゴリ化を促進させることが報告されている．12ヶ月児は，視覚的に (外見的に) 異なる物体であっても，一貫した音声ラベルと共に提示されることで，それら物体を同じ物体カテゴリとして認識できる[21]．また，9ヶ月児は，非母国語の音韻対比 (通常は，一つのカテゴリとして認識する二つの音韻) が，異なる物体と共に提示された場合，それらを異なる音韻カテゴリとして認識できる[22]．すなわち，乳児は，観測刺激の一貫性に基づき，より整合性のとれる情報を利用してカテゴリを形成しているようであ

る．このような，観測刺激の一貫性に基づいたカテゴリ化は，3, 4ヶ月児においても同様に見られることが示唆されており[30; 31]，発達初期から，複数のモダリティ間で相互促進的にカテゴリ化していることが伺える．

一方，乳児の模倣能力は，古典的には，他者や環境との相互作用によって得られる感覚運動経験に基づいて発達的に獲得されていくと考えられている．乳児は，生後8ヶ月頃までに，手の運動や言語の基本単位と呼べる母音の模倣を示すようになる．さらに 14ヶ月頃までには，表情や複数母音などの複雑な模倣ができるようになる[32; 33]．これら能力の発達に並行して，親との相互模倣やりとり[34; 35] が見られることが報告されており，乳児の模倣発達は，そのようなやりとりを通じた感覚運動間の連合学習，すなわち観測した他者の行動と実際に動作させた自身の行動とを対応付ける対応問題を解く過程として捉えられる[36; 18]．新生児模倣の現象により模倣能力は生得的に存在するとの主張があるものの[37; 38; 39] ，ある月齢を過ぎると新生児模倣が消失すること[40; 41]や模倣のバリエーションが少ないことから反射や新生児特有の反応と解釈されている[42; 43]．すなわち，乳児は，生後間もない時期からある種の反射的な模倣を表出できるものの，それを様々な感覚運動経験を通じて発達させ，他者と自身の行動間の対応関係を学習していくと考えられる．そのような模倣の発達過程において，乳児は，既に獲得した経験を利用して対応学習を促進させることが示唆されている．例えば，乳児は，”frown” という単語を聞いた時に顔をしかめている他者の顔を観測し，その後，同じ単語を自分自身が顔をしかめているときにも観測した場合に，その”frown” という単語を介して他者の渋面の顔と自身の顔とを対応付けることができると考えられる[23]．実際に，音声と同時に動作が提示された場合に，動作のみの場合に比べて模倣しやすくなると言われている[44; 33]．すなわち，乳児は，既に獲得した経験を参照して，それに基づいて模倣に必要な対応関係を学習していることが伺える．

上記のように，乳児は他者や環境との相互作用を通じて得られる感覚運動刺激などの観測情報に基づいて，カテゴリ化や対応学習を進めている．また，乳児に接する養育者の行動に関する知見[20; 19] を考慮すると，乳児は，必ずしも単純に観測情報をそのまま学習に利用しているだけではなく，上述の観測刺激の一貫性や既に獲得した

経験などを利用して，すなわち複数の情報の中から学習に必要なデータを取捨選択して，学習を進めていると考えられる．

2.2　人の認知や脳処理における整合機構

前節で説明した乳児の学習戦略と同様の処理が，大人の認知や脳処理においても見られることが社会心理学や認知科学，脳科学の研究からも示唆されている．

人は，矛盾する二つの認知があると不快な状態に陥り，そのような状態を解消または低減するために矛盾している不整合な認知を変え，不快な緊張状態から回復しようとする．例えば，単純でつまらない作業の” 楽しさ” を次に同じ作業を行う者に伝える際，作業についての報酬が少ない場合の方が多い場合よりもその楽しさを伝える度合いが強いと言われている．すなわち，割安な報酬に対して「(実際にはつまらない作業であるが) 本当は面白い作業であったに違いない」と認知を修正し，その不協和を解消しようとする心理が働く．これは社会心理学の分野で認知的整合性理論と呼ばれ，代表的な例として，認知的不協和理論[45]やバランス理論[46]が提唱されている．また，より低次な知覚レベルでも，同様の現象が見られる．人は，複数の感覚運動刺激がある場合，それらを統合して知覚する．例えば,「ガ」と発声している時の口の動きと「バ」と発声している音声を同時に観測した場合,「ダ」と知覚する．これは，マガーク効果[47]と呼ばれ，矛盾している視聴覚情報を統合して知覚を修正していることを示す．他にも，定位に関して口の動きによって音源の位置の知覚が変わる腹話術効果[48] や，視覚と触覚の間でのラバーバンドイリュージョン[49]などが報告されている．これらは，人は恒常的に整合した認知や知覚を好み，それが崩れると，整合性がとれるように自身の認知や知覚を変えようとする傾向にあることを示唆する．

一方，人は観測した感覚刺激を知覚する際に，既に獲得した経験を参照しているようである．代表的な例として，音声知覚の運動理論がある[50]．これは，音声の知覚は，単純に観測した聴覚刺激のみに基づいて行われるわけではなく，構音運動の表象あるいは知識を参照しながら実現されるとする理論である．従来これは概念的な理論に留まっていたが，近年のイメージング技術の進歩により，その裏付けが取られつつ

ある[51]．例えば，霊長類において音声の知覚時に音声を生成 (構音運動) する時の脳部位が賦活することが示されている[52; 53]．人においても同様の結果が示されており[54]，口唇や舌の動きが関わる発話音声を知覚する際に，その動きの生成に関わる部位が賦活すると言われている[55]．最近では，乳児でも同様に，音声を知覚する際に発話に関わる脳部位が関係することが示唆されている[56]．また，音声の知覚においては，構音などの運動情報だけでなく，語彙などの意味的な情報も関わることが示唆されている．音声知覚には，発話時などの構音に関わる音素の表象から運動の表象へのマップを示す Dorsal pathway と語彙的意味処理，すなわち，意味の導出や知識を得るための学習に関わる Ventral pathway の二つの経路が関わることが脳機能イメージングの実験により示されている[57]．このような知見から，人は音声を知覚する際に，聴覚，語彙，構音運動，のそれぞれの表象を繋ぐマッピング[58]を利用していることが伺える．劣化雑音音声[59]を用いた実験により，努力 (学習) すれば聞き取れるようになる音を聞く際には，文生成や発話などの能動的に言語を創出する処理経路が関わることが示唆されており[60]，曖昧な音声を聞く際などには，より顕著に複数の経路を利用した知覚の修正あるいは促進が起こると考えられる．すなわち，人は，単純に観測した刺激をそのまま知覚 (処理) するのではなく，複数の処理経路を利用して，既知の経験あるいは知識とのマッチングにより，知覚を修正あるいは促進していると考えられる．

2.3 主観的整合性の基本アイデア

上記，発達心理学や社会心理学，認知科学，脳科学の知見により，人は乳児期から様々なレベルで情報の整合性に敏感であり，その整合性が崩れる際に，それを解消しようとしていることが伺える．さらに，そのような整合性は，観測のみに対してではなく，既知の経験や知識なども含めて評価されているようである．乳児が，生後わずかな期間で様々なカテゴリ化や対応学習を実施できていることを考えると，そのような学習戦略が，曖昧で様々な情報が溢れる実世界での効率的な学習や行動決定において重要であると考えられる．

本論文では，そのような整合性に基づく学習戦略をロボットの学習に応用し，共同注意が不完全なマルチモーダル環境の問題に取り組む．共同注意が不完全なマルチモーダル環境の問題は，前述のように，ロボットの観測に，注目している事物とは対応しないデータが含まれることであり，例えば，人がロボットが見ている物体の特定に失敗する場合，ロボットが観測する物体と人の発話音声がそれぞれ別の事物を表す．しかし，その場合でも，人が見ている物体と発話音声は一貫して同じ事物を表す，あるいは，聴取した人の発話音声と対応する物体をロボットが既に知っている，場合があると考えられる，すなわち，観測の一貫性や既知の知識を利用し，部分的には特定の事物に対応するデータを判断できる場合があると考えられる．本論文では，マルチモダールな観測のモダリティ間の整合性を評価することで注目すべきモダリティを決定する主観的整合性を導入し，部分的なデータを利用した対応学習およびカテゴリ化の手法を提案する．後章で詳しく説明するように，主観的整合性によって注目すべきモダリティが判断できれば，そのモダリティのデータ，すなわち部分的に対応するデータのみに基づいた学習が可能となり，対応しないデータを排除した効率的な学習が実現できると期待される．

第3章　音声模倣と語彙獲得のための主観的整合性に基づく相補的対応学習

ロボットの対応学習の従来研究では，ロボットが振舞を学習する上で，どこ (どの身体部位) に注目すれば良いかの対応を予め知っている，あるいは，教示者がロボットの動作を真似することが想定されており，その想定が崩れる際に，どのようにロボットが学習を進めるすべきかには焦点はあてられていなかった．これに対し，本章では，教示者が必ずしも学習者を模倣するとは限らない状況における音声模倣学習の課題を扱う．具体的には，認知発達ロボティクスのアプローチにより[61]，乳児が音声模倣と語彙獲得を共発達させている現象に注目し，そこに内在するメカニズムとして，前章で説明した主観的整合性のアイデアを導入した音声模倣と語彙獲得との相補的対応学習の手法を提案する．これにより，教示者がロボットを模倣しない場合であっても，音声模倣に必要な対応学習が可能であることを示す．

3.1　はじめに

人間社会への導入が期待されるロボットにとって，人間とのコミュニケーション能力は重要である．特に音声言語は，人にとって最も自然で馴染みのあるコミュニケーション手段の一つであり，それが可能な会話ロボットの開発が進められている．しかし，工学的な観点から設計製作されたロボットの多くは，技術的な限界も含めて，人間と同等レベルにはほど遠い．これは，音声信号解析，音声合成などの個別の技術課題に加えて，言語という極めて困難かつ大きな研究課題を内包しているためと考えら

れる．これは，人間自身が，いかにして，そのような能力を獲得できたかというミステリーを脳神経科学，認知科学，心理学，言語学など多くの分野も共有していることを意味している．

このような背景に対し，まずは人間の初期，すなわち乳幼児がいかにして，言語を始めとする様々な認知能力を獲得するかの課題に注目し，それをロボットを発達させる課題を扱うことを通じて，構成論的に迫ると同時に，発達する人工物の設計論の確立を目指した認知発達ロボティクスと呼ばれるアプローチが注目されている[61]．認知発達ロボティクスの従来研究では，音声模倣や語彙といった音声言語コミュニケーションの基礎となる機能の発達が個別に研究されている．音声模倣の従来研究では，音声情報と構音運動の対応学習[15; 16; 8]が，語彙に関しては，視覚情報と物体ラベルの対応学習[9; 62; 63]が，モデル化されてきた．しかしながら，実際の乳児は，これら音声模倣と語彙を同時に発達させていることが発達心理学の研究から示唆されており，これらの発達の並行性や相互促進的な発達メカニズムについて議論する必要があると考えられる．

そこで本章では，どのようなメカニズムにより音声模倣と語彙獲得の相互促進的な学習が可能であるかを検討することを通じて，これらの共発達過程の構成的理解を目指す．これまでの研究では，乳児の音声模倣のための対応学習の手がかりとして，養育者が乳児の発話を高頻度で模倣するという傾向[34; 35] が仮定されることが多かった．しかしながら，実場面での養育者と乳児のインタラクションを観察した実験結果から，そのような養育者の模倣の傾向は非常に低いことが報告されている[20; 19]．そのため，従来研究で考えられているような同期して観測される乳児自身の音声と養育者の音声とを結びつける単純な対応学習だけでは，音声模倣のための対応学習は容易ではないと考えられる．これに対し，乳児がある程度語彙を獲得しており，養育者の音声と語彙との対応及び乳児自身の音声と語彙との対応が分かると仮定すると，語彙を介することで，間接的に養育者の音声と乳児自身の音声との対応が想起できると考えられる．そのため，聴取した養育者の音声が発話した乳児自身の音声と対応するものであるかを乳児が直接知らない場合であっても，語彙を介して想起することで，その判断が可能になると考えられる (Fig.3.1 参照)．これにより，乳児が発話した音声に対して，それに対応しない音声を養育者が発話した場合でも，乳児が単純

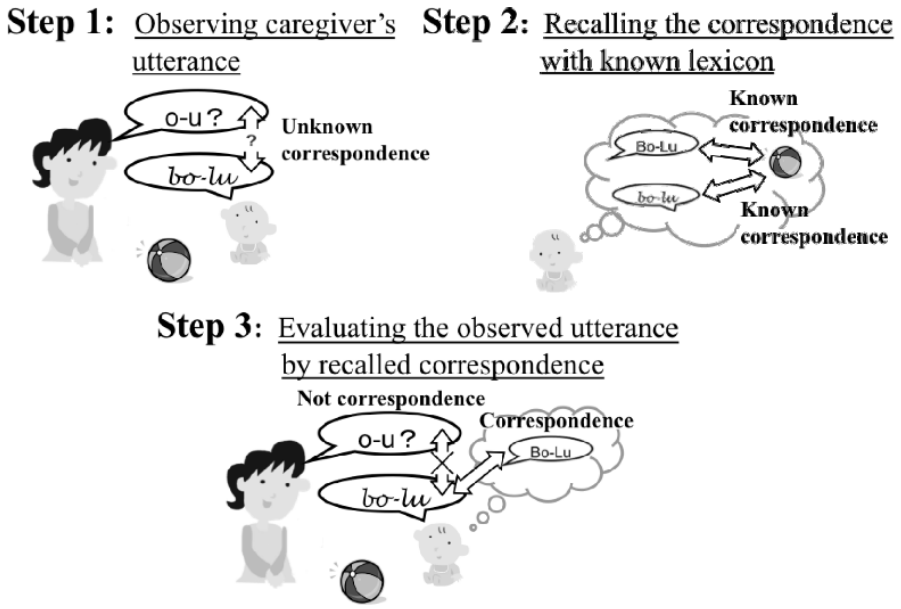


Figure 3.1: An example scene when the learning for vocal imitation is facilitated by acquired lexicon

にそれらの対応を結ぶといった誤りが抑制され，音声模倣のための対応の学習が促進される．このような既知の対応を利用した促進は，語彙の学習においても同様に当てはまる．すなわち，乳児が，聴取した養育者の音声（ラベル）と注目している物体との対応が正しいものであるかを知らない場合でも，養育者の音声と乳児自身の音声の対応及び物体と乳児自身の音声の対応を知っていれば，それが対応するものであるか判断できる．近年の脳科学において，音声処理のための聴覚，語彙概念，構音運動，のそれぞれの表象を繋ぐマッピングが示唆されている[58]ことからも，音声模倣と語彙の学習は相互のマッピングを利用し合いながら発達していくことが察せられるが，それがどのように相互促進的に形成されるかは明らかではない．

そこで本章では，音声模倣と語彙の共発達のための対応学習に焦点をあて，これを，養育者の音声，物体，ロボット自身の音声の 3 つの表象の相互マッピングの学習過程としてモデル化する．以下では，まず最初に，乳幼児の認知発達に対する知見の概略を示し，その上で，本研究の位置づけを 2 節で明らかにする．3 節では，想定する母

子相互作用と学習の具体的なメカニズムを示し，4，5節では，ロボットを用いた簡易的な実験結果およびシミュレーションによる詳細な結果に基づき，手法の有効性を評価する．そして，最後に実験結果を考察し，提案手法の限界と今後の展望に関して議論する．

3.2 乳幼児の認知発達と主観的整合性による共発達

人の乳児は，8ヶ月頃から聴取した語彙への理解を示し， 12ヶ月頃から自身も語彙を発し始める[64]．また，ほぼ時を同じくして， 8ヶ月頃には，言語の基本単位と呼べる母音の模倣を示すようになり，さらに 14ヶ月頃には，それらが連なった複数母音についても模倣できるようになる[33]．このような語彙と音声模倣の能力は，これらが出現する時期が重複していることや，どちらも基本的には音韻の聴取と発声を必要とすることから，互いに影響を及ぼし合いながら発達していくものと察せられる．また，模倣の経験がその後の語彙の発達を促進すること[65]や語彙の知識が模倣に必要な音韻対比の形成を可能とすること[66]が示唆されていることからも，これらの能力は相互促進的に発達していると考えられるが，どのようなメカニズムで，それらの相互促進的発達が可能となるのかについての理解は十分ではない．

このような乳児の発達に関する問いに対して，これまで発達心理学の分野において盛んに研究されてきた．しかしながら，倫理的な問題あるいは言葉の通じない乳児を扱っているため，実験の統制が容易ではなく，発達様相の記述には及んでもその過程の裏にあるメカニズムを理解するには限界があった．そこで，認知発達ロボティクスのアプローチで，この問題に取り組むが，その基本的な考え方は以下である：音声模倣の発達を計算論的に扱うためには，連続的に聴取した音声情報を音韻列として認識できるかのカテゴリ化の問題，その音韻と構音運動との対応付けのマッピングの問題に取り組む必要がある．同様に，語彙の発達では，視覚情報を一つの対象（物体）として認識できるかのカテゴリ化の問題，その対象と音声ラベルとの対応付けのマッピングの問題に取り組む必要がある．カテゴリ化およびマッピングの課題では，ともに，教師信号の選択が大きな問題となる．これは本章の実験で確認していくように，信号

の主観的整合性と呼ぶ指標を導入することで，取扱いが可能である．本章では，主観的整合性の対応学習における有効性の検証に焦点を当てるため，カテゴリ化の課題は次章で扱うことにし，以後では，マッピング課題に集中して議論する．

ここで導入する信号の主観的整合性とは，ある事物をある方法で捉えた信号が，他の複数の方法で捉えた信号とどれほど一致しているかを表すものである．本章では，これを観測信号や，獲得されたマッピングから想起される信号に対して適用し，各信号がどの程度正しい対応関係を表しているかを示すものとみなして，対応関係の学習に用いることを考える．つまり観測された信号をそのまま対応関係の教師信号とするのではなく，それまでに学習された対応関係やそれらの推移的な関係から，観測されるべき信号を想起し，観測信号と複数の想起信号の中に一貫して現れる信号を教師信号とみなす．これにより，対応しない信号が観測される場合，あるいはマッピングの学習が途上である場合，観測信号と想起信号はお互いに異なる信号となり，いずれの主観的整合性も低くなるため，それらの平均的な信号が教師信号となる．この平均的な信号に従い，徐々にマッピングの学習が進み，入力に対して一貫した信号がマッピングから想起されるようになると，対応しない信号が観測された場合に，観測信号の想起信号との主観的整合性が低くなることで，教師信号と見なされにくくなっていくと期待される．従って，複数のマッピングの学習は相互促進的に進んでいくことになると期待される．

3.3 音声模倣と語彙獲得の共発達メカニズム

3.3.1 問題設定

本章では，発達心理学の知見を参考にし，以下を想定する．

- **言語的音声模倣：発声時の構音運動とその後聞こえてきた言葉との対応学習**
  乳児は，構音運動の未熟さや養育者とは異なる構音器官を持つことから，養育者と音響的に同一の音声を発声することは困難である．これに対して，乳児は，養育者に乳児自身の発話を模倣される経験を通じて，音声模倣の学習をしてい

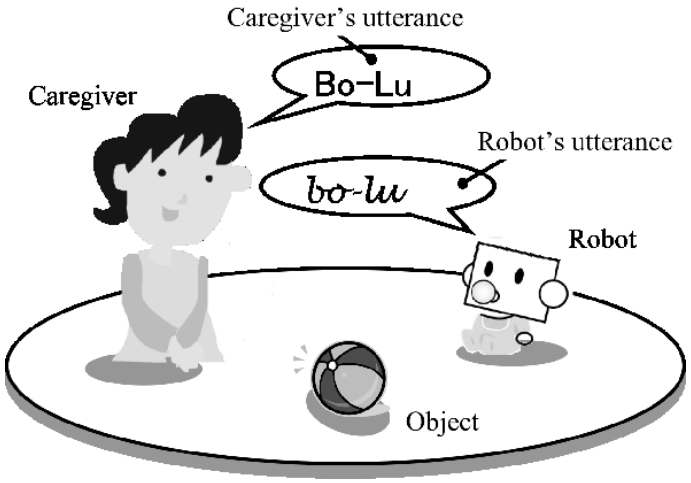


Figure 3.2: Assumed environment of caregiver-robot inreraction

ると考えられている[18]．すなわち，既に乳児の発声に対応する言葉を知っている養育者が，乳児が発した音声をある一つの言葉として言語的に模倣して発声することで，乳児は，その時の構音運動と聞こえてきた発声とを対応付けることができ，音声模倣の能力を獲得していると考えられる．

- **語彙獲得：物体とそのラベルとの対応学習**

  乳児の語彙の学習にとって，養育者から物体を提示され，また同時にその名前を聞かされる経験は重要であるとされている[67]．養育者が，環境中にある多くの物体の中から，乳児が注目している物体を特定し，また同時にそのラベルを発話することで，乳児はそれらを対応付け，語彙を獲得していると考えられる．

- **常に理想的な教師ではない養育者**

  乳児に応じる養育者の行動を分析した研究[20; 19]が示しているように，養育者が乳児に対して，乳児の発した音声を模倣したり，物体とそのラベルを同時に提示したりする教師的な行動で応じることは，現実世界では頻度が低い．

本章では，上記の状況を想定して，Fig.3.2 に示すロボットと養育者と物体が存在する環境を考える．ロボット，養育者の順に交互に行動し，養育者の行動が終了した時点までを 1 ステップとする．各ステップで，ロボットは養育者と物体のどちらかを見る．その時，同時に発声するかしないかを選択する．養育者は，ロボットの行動に対して，音声提示，物体提示，物体呼称の 3 つの何れかの行動で応答する．ただし，養育者の行動が常にはロボットの行動に対応するものであるとは限らないことを考慮し，それぞれの行動の結果は確率的に定まるとする．各行動の詳細は以下の通りである．

音声提示: ロボットの発声に応じて発声する．ただし，養育者がロボットの発声を模倣する言葉 (対応する言葉) を発する場合と，ロボットの発声と関係なく言葉 (この場合は，ロボットの発声とは対応しない他の言葉) を発する場合もあることを考慮して，養育者の発した言葉がロボットの発声の模倣である確率を p_V とする．

物体提示: ロボットの発声に応じて物体を見せる．ただし，養育者がロボットの発声に対応する物体を提示するだけでなく，ロボットの発声に関係なく物体を提示する場合もあることを考慮して，養育者の提示した物体がロボットの発声した物体に対応している確率を p_S とする．

物体呼称: ロボットに物体の名前を教える，すなわち，物体を見せて，その名前を発声する．あるいは，ロボットが見ている物体の名前を発声する．ただし，養育者が物体の名前を教えるため以外に，ロボットに対して言葉を発する場合もあることを考慮して，養育者の発した言葉が物体の名前を教えるためのものである確率を p_D とする．

ロボットは，このような養育者とのインタラクションを通じて，ロボット自身の音声と養育者の音声の対応（音声模倣の能力），養育者の音声と物体及び物体とロボット自身の音声の対応（語彙理解及び生成の能力）を学習する．ただし，養育者は，行動ごとに定められた確率 (p_V, p_S, p_D) に応じてロボットに対応を示し，ロボットは，その確率を事前に知ることはなく，養育者の行動を観測する．これらの確率は，養育

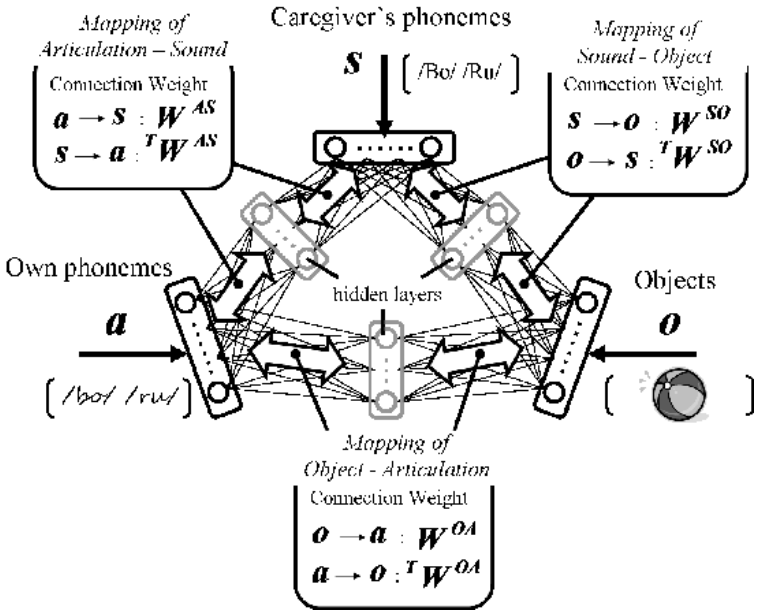


Figure 3.3: Mutually associated multimodal mapping model

者がロボットに対応を与える，つまり，ロボットに対して教示的に振る舞う確率を表し，以後総称して，教示率と呼ぶ．

本章では，教示率を変えて母子相互作用シミュレーションを行うことで，養育者が必ずしもロボットにとって理想的な教師であるとは限らない状況を想定する．そして，そのような状況下でも，提案する主観的整合性に基づいた対応学習により，ロボットが音声模倣と語彙獲得のための対応を学習可能であることを示す．つまり，養育者がロボットに対してより教示的である状況では，それをより信頼して対応を学習し，養育者がロボットに対して全く教示的でない状況では，他のマッピングから想起されるものを信頼して学習する相互促進的学習を実現する．

3.3.2 観測信号ベクトル

前節の母子相互作用によって，ロボットは以下の 3 種の信号を観測する．

1. ロボット自身の発声によって形成されるロボットの音韻系列ベクトル ($\boldsymbol{a} \in \Re^{M_i}$

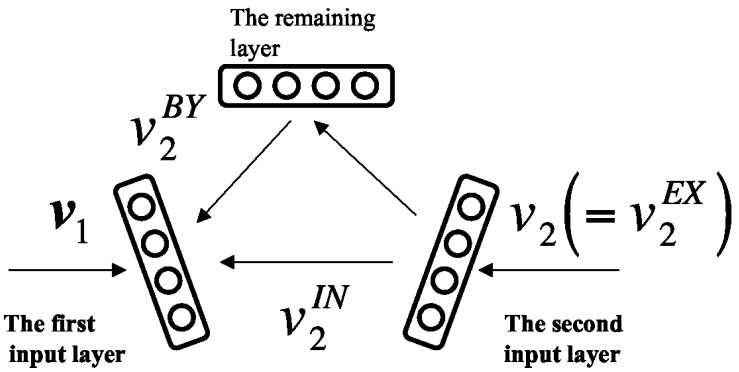


Figure 3.4: Notations for learning rule

)

2. 養育者の発声を観測することで形成される養育者の音韻系列ベクトル ($\boldsymbol{s} \in \Re^{M_c}$)

3. 物体を観測することで形成される物体ベクトル ($\boldsymbol{o} \in \Re^{N}$)

例えば，ロボットが音韻系列ベクトル $\boldsymbol{a_c}$ による発声を行い，養育者が音声提示を行う場合，養育者は確率 p_V で音韻系列ベクトル $\boldsymbol{s_c}$ による発声を，確率 $\boldsymbol{1-p_V}$ で $\boldsymbol{s_{\bar{c}}}$ による発声を行い，これらがネットワークに入力される．ここで，$\boldsymbol{a_i}$，$\boldsymbol{s_i}$ は i 番目の物体のラベルに対する音韻系列ベクトルであり，$\bar{c}$ は，c 番目以外の物体のラベルを表す．

3.3.3 マッピングの学習

本章では，マッピングの学習則として，相互想起型ボルツマンマシンの学習則[68]を適用する．相互想起型ボルツマンマシンは，確率的ニューラルネットワークの一つであり，観測される二つの変数間の確率的関係を学習することができる．そのため，本章で想定する入力に対して出力が一意に定まらない非決定的な状況での対応学習手法として有効である．

ボルツマンマシンでは，ある入力 $\boldsymbol{x}$ に対する出力 $\boldsymbol{y}$ ，及びその逆方向の想起は，共通の結合強度行列 $\boldsymbol{W}$ を用いて，

$$\begin{aligned} \Pr\left(y_m = 1 | \boldsymbol{W}, \boldsymbol{x}\right) &= \frac{1}{1 + \exp\left(-\sum_n w_{nm} x_n\right)}, \\ \Pr\left(x_n = 1 | \boldsymbol{^T W}, \boldsymbol{y}\right) &= \frac{1}{1 + \exp\left(-\sum_m w_{nm} y_m\right)}, \end{aligned} \tag{3.1}$$

の確率に従って実行される．ただし，y_m は $\boldsymbol{y}$ の m 番目の要素であり，x_n は $\boldsymbol{x}$ の n 番目の要素である．w_{nm} は入力層の n 番目のノードと出力層の m 番目のノードの結合の強さを表し，値が大きいほど，それらのノードが対応していることを表す．

本章では，前節で説明した 3 種の観測から，マッピングを学習することを想定し，Fig.3.3 に示すように異なる 3 つの表象が相互に結合したニューラルネットワークモデルを考える．このモデルでは，3 種のベクトルの各要素は 3 つの異なる層のそれぞれ，M_i, M_c, N 個のノードに対応付けられ，ロボットはそれらノード同士の結合強度を要素とする以下の 3 種の行列を学習する．

1. ロボット自身の音韻と養育者の音韻との対応を表す構音-聴覚マッピング $\boldsymbol{W}^{AS}$
2. 養育者の音韻系列と物体との対応を表す聴覚-単語（語彙）マッピング $\boldsymbol{W}^{SO}$
3. 物体とロボット自身の音韻系列との対応を表す単語（語彙）-構音マッピング $\boldsymbol{W}^{OA}$

ここで，Fig.3.4 に示すように， 3 つの層のうち何れかに入力があり，続いて別の層に入力があった場合について考える．初めに入力があった層 (以下，第 1 入力層) への入力が $\boldsymbol{v_1}$ であり，次に入力あった層 (以下，第 2 入力層) への入力が $\boldsymbol{v_2}$ であったとき，それら二つの入力の対応関係は以下の手順に従って学習される．

1. 第 1 入力層及び第 2 入力層をそれぞれ， $\boldsymbol{v_1}$， $\boldsymbol{v_2}$ で固定した状態で，マッピング (式 (3.1)) により，隠れ層の状態を更新する．全てのノードの状態更新を行った時点を 1 サイクルとし，十分なサイクル数で状態を更新する．そして，各結合に関して，それらを繋ぐノードが同時に 1 になる頻度を計算する．ノード n と ノード m が 結合強度 w_{nm} で結合している場合，それらが同時に 1 になる頻度を K_{nm} として計算する．上記の操作を二度繰り返し行い，K_{nm} を計算する．

2. 上記の計算を，今度は，第 1 入力層のみ $\boldsymbol{v_1}$ で固定した状態で行う．同様に各結合に関して，それらを繋ぐノードが同時に 1 になる頻度を計算する．ノード n と ノード m が 結合強度 w_{nm} で結合している場合，それらが同時に 1 になる頻度を ${}^1K'_{nm}$ として計算する．また，両方向の対応を学習するため，第 2 入力層のみ $\boldsymbol{v_2}$ で固定した状態でも同様に計算し，ノード n と ノード m が同時に 1 になる頻度を ${}^2K'_{nm}$ とする．そして，入力層を固定した状態で，ノード n と ノード m が同時に 1 になる頻度 K'_{nm} を，第 1 及び第 2 入力層を固定した状態で計算した，${}^1K'_{nm}$，${}^2K'_{nm}$ の和として，以下のように計算する．

$$K'_{nm} = {}^1K'_{nm} + {}^2K'_{nm}. \tag{3.2}$$

3. ノード n と ノード m の結合強度 w_{nm} を上記計算で求めた K_{nm}， K'_{nm} により，以下のように更新する．

$$w_{nm} = w_{nm} + \alpha\left(K_{nm} - K'_{nm}\right). \tag{3.3}$$

ここで，α は学習係数である．

上記の手順を繰り返し行い，結合強度を更新していくことで，二つの入力間の対応がとれるようにマッピングを学習する．例えば，ロボットが発声したのち，養育者が発声した場合，$\boldsymbol{v_1}$ はロボット自身の音声 $\boldsymbol{a}$，$\boldsymbol{v_2}$ は観測される養育者の音声 $\boldsymbol{s}$ となり，ロボットが養育者の発声を模倣できるようになるためには，聞こえてきた養育者の音声 $\boldsymbol{s}$ からそれに対応するロボット自身の音声 $\boldsymbol{a}$ をマッピングを介して正しく推定できるように，結合強度行列 $\boldsymbol{W^{AS}}$ を更新していくことが必要となる．

3.3.4 主観的整合機構：複数信号の整合性に基づく統合

前節の学習則により，例えば，ロボットの発声を養育者が高頻度で模倣する場合，ロボットはロボット自身の音声とそれに対応する養育者の音声の関係を学習することができる．しかし，養育者が高頻度で模倣ではない発声をロボットに行う場合は，そのような単純な対応学習のみでは，ロボットは誤った対応関係を学習する問題がある．

これに対して，Fig.3.3 に示す学習モデルでは，複数のマッピングが相互に結合しており，学習時に養育者から与えられる信号以外にマッピングを介して想起される信号も同時に利用することができる．そのため，マッピングが成熟していれば，そこから想起される信号を学習に利用できる，すなわち，養育者からの信号ではなく，想起される信号を対応する信号とみなして学習することができる．そのため，養育者が高頻度で対応しない信号をロボットに与える場合であっても，ロボットが誤った対応を学習する問題を抑制できると考えられる．しかしながら，それをロボット自身が判断するとなれば，ロボットが主観的に観測・計算しうる形で，その判断の仕組みが構成される必要がある．本節では，そのような複数の信号を基に主観的に対応する信号を生成するための整合性に基づいた統合手法を提案する．

前節と同様に，3 つの層のうち何れかに入力があり，続いて別の層に入力があった場合について考える．この時，第 2 入力層への入力 $\boldsymbol{v}_{\boldsymbol{2}}^{EX}(=\boldsymbol{v_2})$ (以下，外部信号) ，第 1 入力層 から 第 2 入力層への直接のマッピングを介して想起される信号 $\boldsymbol{v}_{\boldsymbol{2}}^{IN}$ (以下，直接予測信号) 及び 第 1 入力層から 残っているもう一つの層を介した第 2 入力層への間接のマッピングを介して想起される信号 $\boldsymbol{v}_{\boldsymbol{2}}^{BY}$ (以下，間接予測信号) を統合した統合信号 $\boldsymbol{v}_{\boldsymbol{2}}'$ を，次式のように計算する．

$$\boldsymbol{v}_{\boldsymbol{2}}' = f(\boldsymbol{v}_{\boldsymbol{2}}^{EX}, \boldsymbol{v}_{\boldsymbol{2}}^{IN}, \boldsymbol{v}_{\boldsymbol{2}}^{BY}) = \lambda_{EX}\boldsymbol{v}_{\boldsymbol{2}}^{EX} + \lambda_{IN}\boldsymbol{v}_{\boldsymbol{2}}^{IN} + \lambda_{BY}\boldsymbol{v}_{\boldsymbol{2}}^{BY}. \tag{3.4}$$

ここで，$\lambda_n\,(n \in \{EX, IN, BY\})$ は 外部信号，直接予測信号，間接予測信号，それぞれの主観的整合性を表し，

$$\lambda_n = \frac{\exp\left(-e_n/\sigma^2\right)}{\sum_{m\in\{EX,IN,BY\}} \exp\left(-e_m/\sigma^2\right)}, \tag{3.5}$$

と計算される．ここで，σ は e_n に対する感度パラメータである．また，e_n は信号 $\boldsymbol{v}_{\boldsymbol{2}}^{n}$ が他の二つの信号と比べて，どれ程かけ離れたものであるかを表し，信号間の距離の和として，

$$e_n = \sum_{l\notin n} \left\| \boldsymbol{v}_{\boldsymbol{2}}^{n} - \boldsymbol{v}_{\boldsymbol{2}}^{l} \right\|, \tag{3.6}$$

と計算される．式 (3.6) より，信号間の距離の和を計算し，それに基づき，式 (3.4) を用いて統合することで，信号間の近さに応じて統合信号を求める．これにより，あ

る一つの信号が他の二つの信号と近ければ，統合時にその信号が反映される．逆に，ある一つの信号が他の二つの信号と遠ければ，統合時にその信号が反映されないように統合信号が計算される．提案手法では，外部信号，直接予測信号，間接予測信号，それぞれについて，主観的整合性を計算し，式 (3.4) で重み付け統合した信号を，入力とみなしてマッピングを学習する．すなわち，前節の学習則の手順 (1)，(2) において，第 2 入力層を外部からの入力 $\boldsymbol{v}_2^{EX}$ ではなく，統合信号 $\boldsymbol{v}_2'$ で固定して結合強度を更新する．

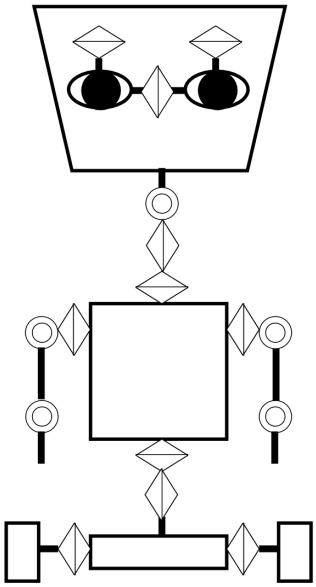
以上のように，外部信号だけでなくマッピングを介して想起される信号も含めて統合し，それを対応する信号とみなして学習することで，養育者の行動だけに依存して学習することを防ぐことができる．これにより，養育者が正しい対応を示さない場合に誤った対応を学習する問題を抑制できると考えられる．また，それらを相互の近さによって重み付け統合することで，より一貫している信号がより正しい対応を示す信号であるとロボット自身が主観的に判断できる．

3.4 実ロボットを用いた実験

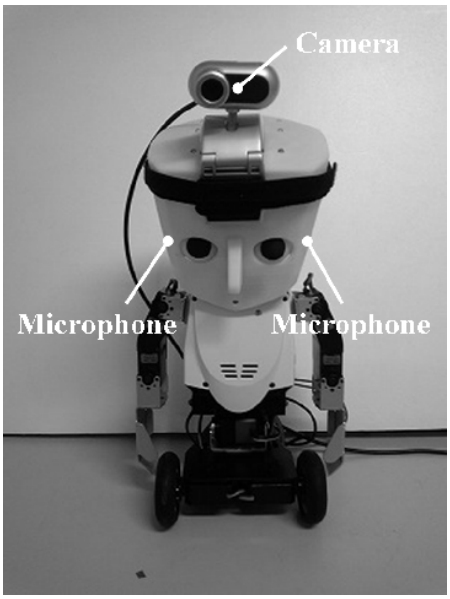
本章では，実例として，比較的簡単な環境を想定した実ロボットと人とのインタラクションにおいて，提案手法を適用した結果について述べる．

3.4.1 実験設定

本実験では，実ロボットとして小型ヒューマノイドロボット Synchy を用いた (Fig.3.5 及び Table 3.1 参照)．Fig.3.6 に示すように，ロボットと人が互いに向き合い，間に物体が置かれている状況を考える．インタラクションに使用する物体及びラベルの数は 3 (aka, ao, midori) とし，ラベルを構成する音韻（モーラ）は，6 (/a/, /ka/, /o/, /mi/, /do/, /ri/) とする．このような状況で，ロボットは，ランダムに物体や人に視線をうつしたり，発声したりすることで，人の発声や提示，教示を誘発し，それによって与えられる音声と画像情報から，自身と人のモーラ同士の対応，自身のモーラ系列



(a) Overview of the actuators



(b) Photo from the front view of robot

Figure 3.5: Humanoid robot “Synchy”

と物体の対応及び人のモーラ系列と物体の対応付けを行う．ロボットの行動は，具体的には，Fig.3.7 に示す 3 つの行動とした．ロボットの各行動は，以下のように 3 つの動作で構成される．

発声誘発行動 (Fig.3.7 (a)):
1. 人の顔を見る．
2. 発声を行う．
3. 人の発声による音声入力を待つ．

提示誘発行動 (Fig.3.7 (b)):
1. 物体を見る．
2. 発声を行う．
3. 人の提示による画像入力を待つ．

教示誘発行動 (Fig.3.7 (c)):
1. 環境にあるすべての物体を見る．
2. 人の顔を見る．
3. 環境にある一つの物体を見ることで画像入力を得た状態のまま，人の発声による音声入力を待つ．

Table 3.1: “Synchy” hardware specifications

Size (mm)		360(H) × 190(W) × 133(D)
Weight (kg)		1.3
DOF		16
Motor Type		Vstone Servo VS-S092J
Computing unit	CPU	Vstone VS-RC003 ARM
	ROM	512 KB
	RAM	40 MB
Camera unit	Type	ELECOM UCAM-DLT30H
	Pixel count	300K pixels
	Photoreceptor	1/6 inch CMOS sensor
	Maximum resolustion	640 × 480 pixels
	Maximum frame rate	30 fps

自身の音声を表すベクトル $\boldsymbol{a}$ は，ステップ毎に物体のラベルを構成するモーラの組み合わせをランダムに選ぶことで生成される．そして，生成したベクトルから，あらかじめ用意した各モーラに対応する音声を組み合わせて発声するようにした．相手の音声を表すベクトル $\boldsymbol{s}$ は，便宜的に聴取した人の音声を汎用大語彙連続音声認識エンジン Julius を用いて認識し，それを構成するモーラの組み合わせから生成されるようにした．また，物体を表すベクトル $\boldsymbol{o}$ は，取得した画像をあらかじめ作成した物体毎の色に関するルックアップテーブルを用いて ID を認識し求めた．提案手法における各パラメータは経験的に $\alpha = 1.0$，$\sigma = 1.0$ とした．

提案手法のパフォーマンスを，観測された信号をそのまま教師信号とみなす単純な同期性に基づく手法のみに従って学習した場合，すなわち，式 (3.4) を，

$$\boldsymbol{v_2'} = \boldsymbol{v_2^{EX}} \tag{3.7}$$

とする方法 (以下，direct) と比較することで，提案手法の有効性の検証を行う．

パフォーマンスは，入力ベクトルを一通り入力することで測定する．例えば，構音-聴覚マッピングのパフォーマンスを測定する場合，環境中の物体ラベルを表す 3 通り

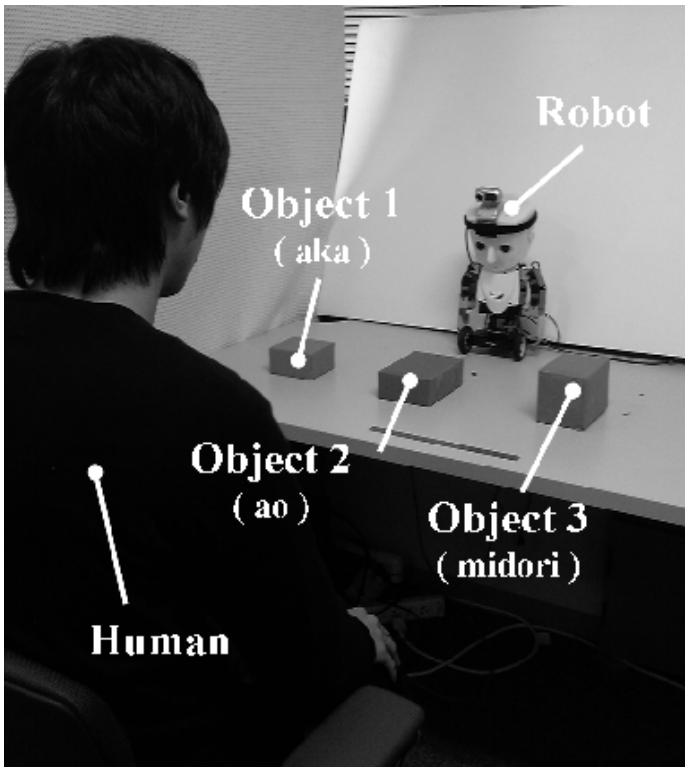


Figure 3.6: A sample scene for human-robot interaction

(/a//ka/, /a//o/, /mi//do//ri/) の音韻系列ベクトルを，自身の音声を表すベクトル $\boldsymbol{a}_i$ $(i \in N)$ とし，これを構音-聴覚マッピングに入力する，すなわち式 (3.1) の $\boldsymbol{x}$ とすることで，相手の音声の予測ベクトル $\boldsymbol{s}_i^{in}$ を想起する．この予測ベクトルが，正しく対応する相手の音声ベクトル $\boldsymbol{s}_j$ $(j = i)$ に最も近くなった場合は， 1 ，それ以外の対応しない相手の音声ベクトル $\boldsymbol{s}_j$ $(j \neq i)$ に最も近くなった場合は 0 となるように，評価点 R_i を，

$$R_i = \begin{cases} 1 & \text{if } \left|\left|\boldsymbol{s}_i^{in} - \boldsymbol{s}_i\right|\right| = \min\limits_{j \in N} \left(\left|\left|\boldsymbol{s}_i^{in} - \boldsymbol{s}_j\right|\right|\right) \\ 0 & \text{otherwise} \end{cases} \tag{3.8}$$

のように計算する．評価点は，3 通りのベクトルすべてについて計算し，また，相手の音声から自身の音声の予測ベクトルを想起した場合の両方向について計算する．最

終的な評価値は，評価点が 1 となる割合として，

$$E = \frac{\sum_{i \in N} R_i}{2N} \tag{3.9}$$

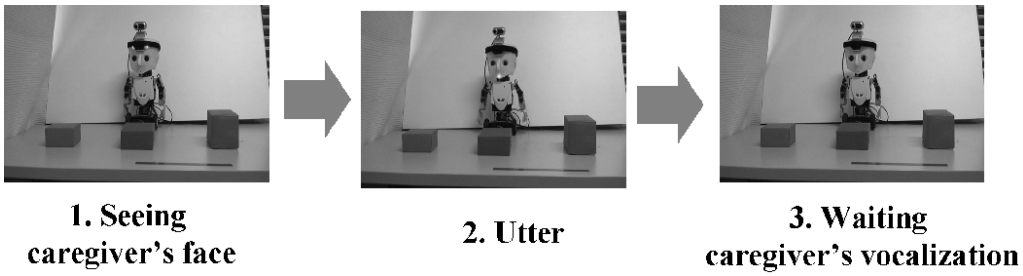
のように計算する．語彙聴取マッピング，語彙生成マッピングについても同様にして，パフォーマンスを計算する．

3.4.2 実験結果

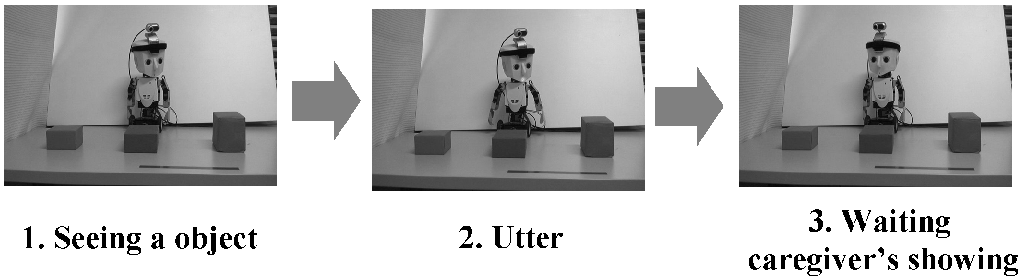
Fig.3.8 に 500 ステップのインタラクション実験を行った場合の，各マッピングの学習パフォーマンスの遷移を示す．500 ステップのインタラクションにおける教示率は 音声提示では 74.2 %，, 物体提示では 76.9 %,，物体呼称では 75.1 %，であった．

Fig.3.8 (a) (c) (e) より，単純な同期性に基づく手法のみに従って学習した場合，高いパフォーマンスに留まらず，収束していないことが伺える．それ対して，Fig.3.8 (b) (d) (f) より，提案手法に従って学習した場合では，高いパフォーマンスで飽和しており，頑健に対応の学習が可能であることがわかる．本実験では， 1 ステップの所要時間は 15 秒程度であるため，本実験の設定のように，教示率が 75 % 程度であるような，人の行動が完全には制限されない場合でも，提案手法により， 2 時間程度のインタラクションで，3 種類の音韻間の対応及び物体と複数の音韻で表されるラベル間の対応を学習可能である．

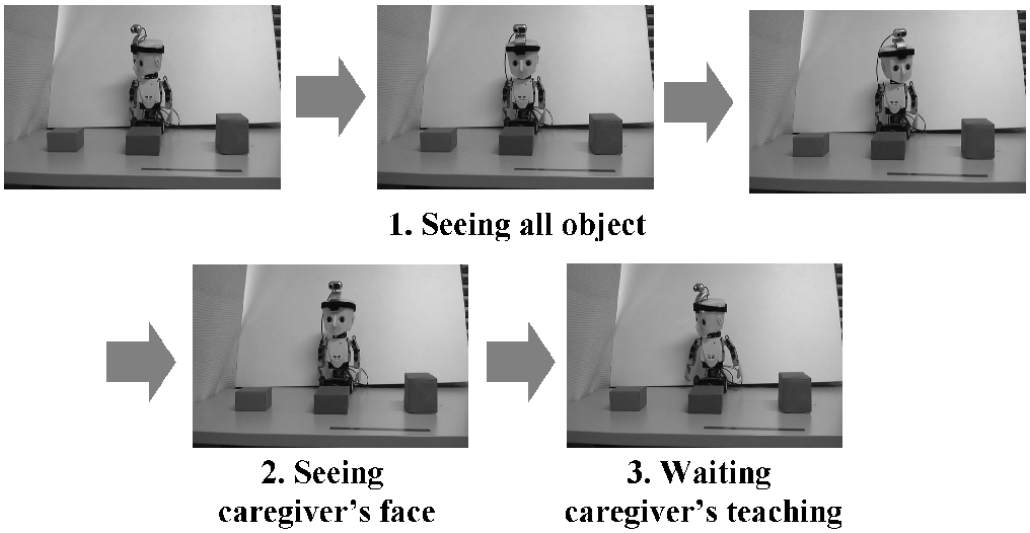
しかしながら，乳児の音声模倣と語彙の共発達メカニズムとしての提案手法の有効性を確かめるためには， 3 種類の物体だけでなく，実際の乳児が置かれているような多種の物体が存在する環境において，どの程度までの教示率ならば学習可能か，または教示率に偏りがある場合でも，相互促進的に学習可能かについて検証を行う必要があると考えられる．そこで，次節において，様々な環境を想定したシミュレーション実験を実施し，提案手法の有効性を詳細に検証する．



(a) Behavior for inducing human to vocalize

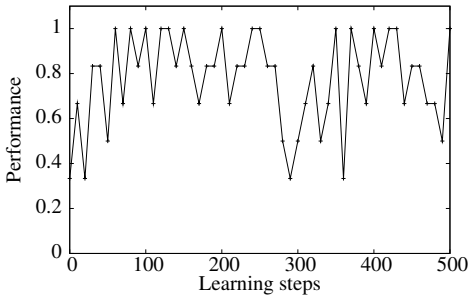


(b) Behavior for inducing human to show a object

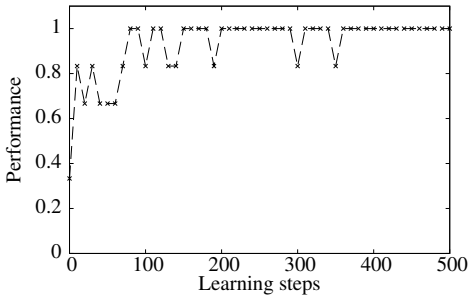


(c) Behavior for inducing human to share a name

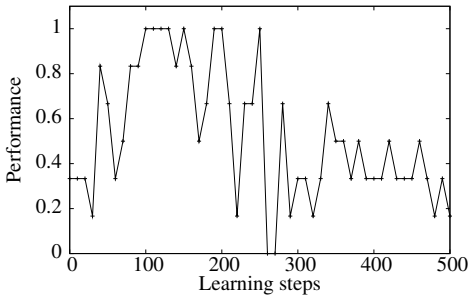
Figure 3.7: Behaviors of robot



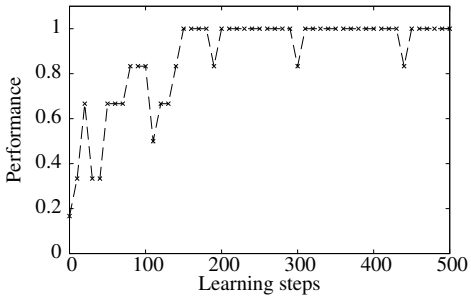
(a) Articulation-Sound Mapping (direct)



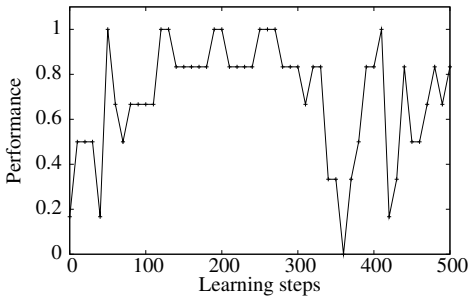
(b) Articulation-Sound Mapping (proposed)



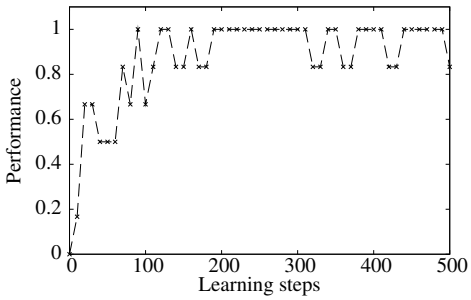
(c) Sound-Word Mapping (direct)



(d) Sound-Word Mapping (proposed)



(e) Word-Articulation Mapping (direct)



(f) Word-Articulation Mapping (proposed)

Figure 3.8: Transitions of learning performance of (a) Articulation-Sound Mapping, (c) Sound-Word one, and (e) Word-Articulation one without the proposed subjective consistency, in turn (b) (d) (f) with it

3.5 シミュレーション実験

提案手法の有効性を詳細に確かめるために，計算機シミュレーションを用いて以下の 3 つの実験を実施した．

予備実験: 簡単な状況において，提案手法により，誤った対応が与えられる場合に，それを抑制できるかを確認する．

実験 1: 様々な教示率で養育者がロボットに応答する状況を想定し，提案手法の頑健性及び限界を検証する．

実験 2: 乳児に応じる養育者の行動を分析した実験[20]から，実場面では，養育者が乳児に対して模倣やラベル付けを行う割合は非常に低いことが示されている．そこで，そのように養育者が乳児にほとんど対応を与えない状況においても，提案手法により，相互促進的な対応の学習が可能であるか検証する．

3.5.1 基本設定

乳児は，まず聴取した音声を母国語の音韻としてある程度認識できるようになり[69]，その後，模倣できる[33]ように発達していると考えられる．すなわち，乳児は，聴取した音声から既にある程度排他的に音韻を認識し，それと乳児自身が発話した音韻との対応がとれるように音声模倣能力を発達させていると考えられる．そこで，ここでは簡単のため，ロボットと養育者の発声はお互いが共通に持ついくつかの音韻（モーラ）で構成されるものとした．具体的には，$\boldsymbol{a}$，$\boldsymbol{s}$ は，それぞれが M 種類あるうちのどのモーラで構成されているかを表わすベクトルであるとした．例えば，ロボットの発声が $/a_2\ a_8/$ のように，2 番目のモーラと 8 番目のモーラの組合せで構成される音であった場合，$\boldsymbol{a}$ は，2 番目と 8 番目の要素が 1，それ以外の要素が 0 である M 次元ベクトルとなる．また，乳児は語彙の獲得以前から，ある程度物体の認識が可能である[70]ことが報告されている．そこで，物体は，N 種類のどの物体について注目しているかを表わすベクトル $\boldsymbol{o}$ で表記できるとした．例えば，ロボットが i 番目の物体に注目している場合，$\boldsymbol{o}$ は i 番目の要素が 1，それ以外が 0 である N 次

元ベクトルとなる．これらのベクトルは，それぞれの表象に入力される特徴ベクトルに相当する．例えば，養育者の音韻系列ベクトル $\boldsymbol{s}$ は，聴取した人の音声データを音声認識器を用いて対応するモーラ毎に分解し，ベクトル化 (あるモーラが含まれていれば 1，含まれていなければ 0) したものに相当する．

予備実験では，主観的整合機構により，どのように誤った対応学習が抑制されるかについて注目するため，簡単な状況を想定する．具体的には，環境中の物体及びラベルの数は，$N = 2$ 個とし，すべてのラベルを構成するためのモーラ数も $M = 2$ 個とした．また，すべての変数が同じ値 ($\boldsymbol{a} = \boldsymbol{s} = \boldsymbol{o} = [0, 1]^T$) 及び $\boldsymbol{a} = \boldsymbol{s} = \boldsymbol{o} = [1, 0]^T$))) となる場合を正しい対応とした．

実験 1，2 では，より複雑な状況として，実際の乳児が置かれている状況を想定して，2009 年 2 月 22 日時点で goo ベビー[71]に記載されていた乳児が 10ヶ月から 18ヶ月までに獲得する語彙を実際の母子相互作用場面において頻出する語彙とし，その中の名詞単語をシミュレーションで使用するデータとした．抽出したデータから，環境中の物体及びラベルの数は，$N = 39$ 個であり，すべてのラベルを構成するためのモーラ数は $M = 37$ 個であった (Table 3.2, 3.3 参照)．また，実験 1，2 では，提案手法の有効性を確かめるために，観測された信号を対応する信号とみなす，通常の相互想起型のボルツマンマシンの学習手法 (3.3.3 節参照) と，主観的整合性に基づき統合した信号を対応する信号とみなす提案手法 (3.3.4 節参照) を比較する．具体的には，後者に従う程度を表すパラメータ η を導入し，対応の学習のための信号 $\boldsymbol{v}'$ を，

$$\boldsymbol{v}' = (1 - \eta)\boldsymbol{v}^{EX} + \eta(\lambda_{EX}\boldsymbol{v}^{EX} + \lambda_{IN}\boldsymbol{v}^{IN} + \lambda_{BY}\boldsymbol{v}^{BY}), \tag{3.10}$$

と決定する．すなわち，η の値が大きいほど，提案手法に頼る割合が増え，低いほど，通常の学習手法に頼る割合が増えることを意味する．

Table 3.2: Phoneme data for simulation

No	Word	Phoneme (mora)	a	i	u	ka	ku	ko	si	ta	ti	tsu	to	na	ni	ne	ma	mi	me	n	ga	gi	go	ji	de	ba	bi	bu	bo	pa	pu	po	tya	tyu	tyo	bya	ju	ttu	-
1	manma	/ma/ /n/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	ne-ne	/ne/ /-/	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
3	mi-	/mi/ /-/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
4	ga-	/ga/ /-/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
5	tyattya	/tya/ /ttu/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
6	jijji	/ji/ /ttu/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
7	anpanpan	/a/ /n/ /pa/	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
8	u-u-	/u/ /-/	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
9	anan	/a/ /n/	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	panpan	/n/ /pa/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
11	tyotyo	/tyo/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
12	bai	/i/ /ba/	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	bobo	/bo/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
14	nya-nya-	/nya/ /-/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0
15	gogog	/go/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	kutsu	/ku/ /tsu/	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	ka-ka-	/ka/ /-/	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
18	natto	/to/ /na/ /ttu/	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
19	poppo	/po/ /ttu/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0
20	bu-	/bu/ /-/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
21	baba	/ba/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	konkon	/ko/ /n/	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	me-me-	/me/ /-/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
24	pu-	/pu/ /-/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
25	niitan	/i/ /ta/ /ni/ /n/	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	pa	/pa/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
27	bi-bi-	/bi/ /-/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
28	si	/si/	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	bo-	/bo/ /-/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
30	tyuntyun	/n/ /tyu/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
31	nma	/n/ /ma/	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	iti	/i/ /ti/	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	pu-n	/n/ /pu/ /-/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
34	kankankan	/ka/ /n/	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	popo	/po/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
36	dcntya	/n/ /dc/ /tya/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
37	gi-	/gi/ /-/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
38	bunbunbun	/n/ /bu/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
39	jutyu	/tyu/ /ju/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0

Table 3.3: Object data for simulation

No	Word	Object	food	sister	orange	power shovel	tea	grandfather	Anpan-man	firetruck	dog	bread	batterfly	Baikin-man	hat	cat	strawberry	shoe	crow	natto
1	manma	food	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	ne-ne	sister	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	mi-	orange	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	ga-	power shovel	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	tyattya	tea	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
6	jiji	grandfather	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
7	anpanpan	Anpan-man	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
8	u-u-	firetruck	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
9	anan	dog	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
10	panpan	bread	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
11	tyotyo	batterfly	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
12	bai	Baikin-man	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
13	bobo	hat	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
14	nya-nya-	cat	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
15	gogog	strawberry	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
16	kutsu	shoe	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
17	ka-ka-	crow	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
18	natto	natto	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
19	poppo	railway train	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	bu-	car	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	baba	grandmather	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	konkon	fox	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	me-me-	sheep	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	pu-	Winnie-the-Pooh	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	nitan	brother	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	pa	leaf	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	bi-bi-	TV	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	si	bug	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	bo-	ball	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	tyuntyun	bird	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	nma	horse	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	iti	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	pu-n	spoon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	kankankan	railroad crossing	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	popo	dandelion	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36	dentya	train	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
37	gi-	door	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
38	bunbunbu	bee	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
39	jutyu	juice	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

railway train	car	grandmather	fox	sheep	Winnie-the-Pooh	brother	leaf	TV	bug	ball	bird	horse	1	spoon	railroad crossing	dandelion	train	door	bee	juice
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

3.5.2 予備実験: 主観的整合機構による誤った対応学習の抑制効果の確認

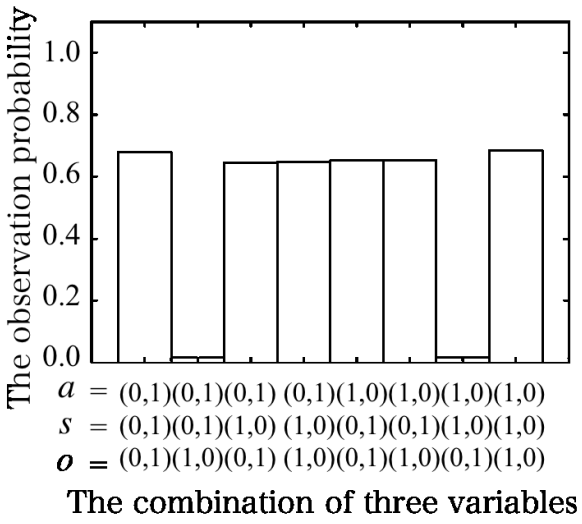
提案手法によって，誤った対応が与えられる場合に，それを抑制できるかを確認するために，あえて誤った対応が与えられる状況を想定して実験した．具体的には，教

示率を $p_V = 0.05$, $p_S = 1.0$, $p_D = 1.0$ とし，100,000 ステップの母子相互作用シミュレーションを実施した．Fig.3.9 に，100,000 ステップの学習中に観測された 3 つの変数の組み合わせとそれを教師信号として学習に用いた頻度を示す．Fig.3.9(a) は，入力された信号をそのまま教師信号とした用いた場合であり，(b) は，提案手法により統合した信号を教師信号として用いた場合である．本実験では，$\boldsymbol{a} = \boldsymbol{s} = [0,1]^T$ 及び $\boldsymbol{a} = \boldsymbol{s} = [1,0]^T$ を正しい対応としているので，それらの観測頻度が，3 つの変数の他の組み合わせに比べて相対的に高い場合，観測の共起関係のみから，それら正しい対応を学習できる．しかしながら，$p_V = 0.05$ としたため，$\boldsymbol{a}$ と $\boldsymbol{s}$ の間の正しい対応 ($\boldsymbol{a} = \boldsymbol{s} = [0,1]^T$ 及び $\boldsymbol{a} = \boldsymbol{s} = [1,0]^T$) が入力される割合がチャンスレベルよりも非常に少く，結果として，Fig.3.9(a) に示すように，3 つの変数の正しい対応 ($\boldsymbol{a} = \boldsymbol{s} = \boldsymbol{o} = [0,1]^T$ 及び $\boldsymbol{a} = \boldsymbol{s} = \boldsymbol{o} = [1,0]^T$) を教師信号として用いる頻度が他の組み合わせの観測頻度とほぼ同程度となっている．そのため，単純に入力の共起関係をみるだけでは，正しい対応の学習が困難である．これに対して，提案手法によって統合した信号を教師信号とすることで，Fig.3.9(b) に示すように，正しい対応を教師信号とする頻度が，他の組み合わせよりも相対的に高くなっていることがわかる．この結果から，提案手法により，誤った対応が与えられる場合は，それを教師信号として用いないようにするという抑制が可能となることが確認された．

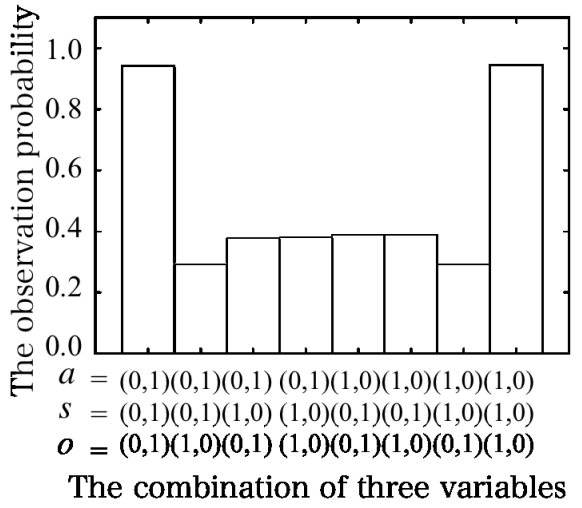
3.5.3 実験 1: 主観的整合機構による促進効果の検証

教示率毎の学習パフォーマンス

対応学習の目的においては誤った対応となる信号が観測される状況であっても，提案手法により頑健に対応の学習が可能であるかを確かめるために，養育者の教示率 p_V, p_S, p_D を同じ値 p_a に固定し，p_a を 0.0 から 1.0 まで 0.1 ずつ変化させたそれぞれの場合について，η を 0.0 から 1.0 まで 0.1 ずつ変化させ，100,000 ステップの母子相互作用シミュレーションを 10 回実施した．Fig.3.10 は，p_a と η の違いによる，対応学習のパフォーマンスの違いを示している．但し，明暗は 100,000 ステップ経過時点での学習パフォーマンスを表しており，明るい程高く，暗い程低いことを表す．



(a) Without integrated signal



(b) With integrated signal

Figure 3.9: Observation probability with respec to combinations of three input variables (a) without integrated signal generated by proposed mechanism and (b) with it

パフォーマンスは，入力ベクトルを一通り入力することで測定する．具体的には，39 通りの可能な入力ベクトルをそれぞれ各表象に入力し，マッピングを介して想起されるベクトルが 39 通りの可能な出力ベクトルの中で正しく対応するものに最も近くなった場合の割合の，各マッピングに関する平均値として評価した．また，提案手法における各パラメータは経験的に　$\alpha = 0.2$，$\sigma = 1.0$ とした．Fig.3.10 より，教示率が高い場合 ($p_a = 1.0$)，すなわち，外部信号として直接対応を示す信号が観測される場合では，通常の学習手法のみでも高いパフォーマンスが得られていることがわかる．一方，教示率が比較的低い場合 ($0.2 <= p_a < 1.0$)，すなわち，外部信号として対応を示さない信号も観測される場合では，主観的整合性に基づく統合に従う程度 (η) が高いほど，正しい対応学習が可能になっていることがわかり，提案手法の有効性が確認できる．しかしながら，教示率がかなり低い場合 ($p_a < 0.2$) では，たとえ主観的整合性に基づく統合に従う程度が高かったとしても，最終パフォーマンスは低く，$\eta = 0.8$ 付近でのパフォーマンスが最も高い．これは，教示率がかなり低い場合では，誤った対応を経験する割合が多くなり，複数のマッピングが誤った対応を学習する割合も増えるため，主観的整合性が高いことで，他のマッピングもその誤った対応で固定される結果であると考えられる．

極端に低い教示率に対する提案手法の効果

教示率が低い場合での提案手法の効果を詳細に分析するために，$\eta = 1.0$ と固定し，p_a を 0.0 から 0.2 まで 0.01 ずつ変化させた場合の学習パフォーマンスを Fig.3.11 に示す．これより，p_a が低くなるにつれ，パフォーマンスが急激に下がっていることがわかる．$0.05 < p_a < 0.15$ の範囲では，分散が大きく，場合によっては，ある程度高いパフォーマンスが得られているものの，$p_a <= 0.05$ の範囲では，パフォーマンはゼロ付近に留まっている．これは，主観的整合性に基づく統合では，一貫した信号であれば，それがたとえ誤った対応を示すものであっても，信頼してマッピングの学習が進められるためであると考えられる．つまり，提案手法により，学習者は，パフォーマンスの良し悪しとは関係なく，あるマッピングが一貫した信号を想起していれば，その信号を正しい対応を示すものであると主観的に判断し学習する．しかしながら，

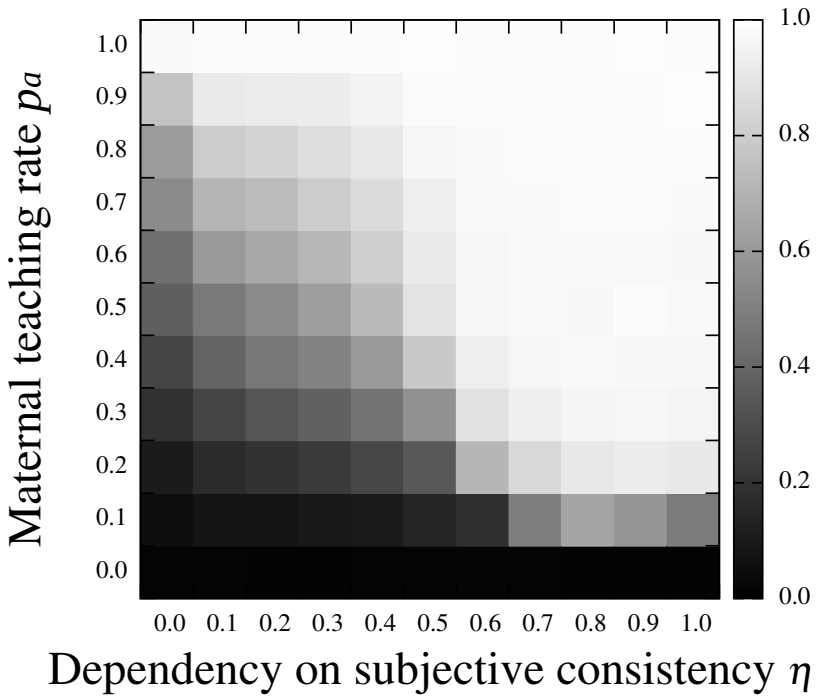


Figure 3.10: Average probability of predicting corresponding vector by acquired mappings until 100,000 step with respect to dependency on subjective consistency(η) and maternal teaching rate (p_a)

その一貫した信号が誤った対応を示すものであった場合，その誤った対応を示す信号によって他のマッピングの学習が拘束される．そして，最終的には，対応を何も学習しないのではなく，すべてのマッピングで一貫した対応ではあるものの，誤った対応を学習してしまい，結果として，最終的なパフォーマンスが低くなっている．そのため，養育者が対応を与える割合がかなり低い場合 ($p_a < 0.2$) では，提案手法のような主観的整合性に基づく統合により，すべてのマッピングで一貫した対応を学習するものの，それが正しい対応となる保証はなく，注意する必要がある．

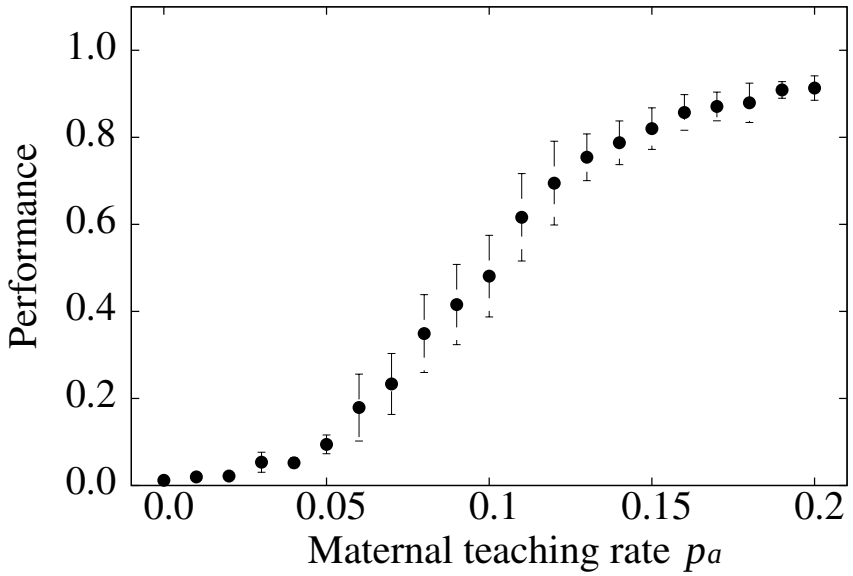


Figure 3.11: Average probability of predicting corresponding vector by acquired mappings until 100,000 step with respect to low maternal teaching rate (p_a) under the proposed method ($\eta = 1.0$)

教示率に対する学習の適応性

提案手法により，教示率に対してどのように適応的に学習が可能となったかを見るために，提案手法にのみ頼って学習した場合 ($\eta = 1.0$) の教示率と学習終了時 (99,900 〜 100,000 ステップ) における各信号に対する主観的整合性の関係を Fig.3.12 に示す．Fig.3.12 から分かるように，$0.2 <= p_a < 1.0$ の範囲では，教示率が低くなるにつれ，外部信号に対する主観的整合性のみが下がっていることが見てとれる．すなわち提案手法により，外部信号が常に対応する信号ではない場合には，それが対応する信号としてあまり反映されないようにすることで，教示率によらない頑健な対応学習が可能になっていると考えられる．しかしながら，教示率がかなり低い場合 ($p_a < 0.2$) では，外部信号がほとんど対応を示すものではないにも関わらず，その主観的整合性が $p_a = 0.2$ の場合と同程度であることがわかる．これが上記で説明した，対応を何も学

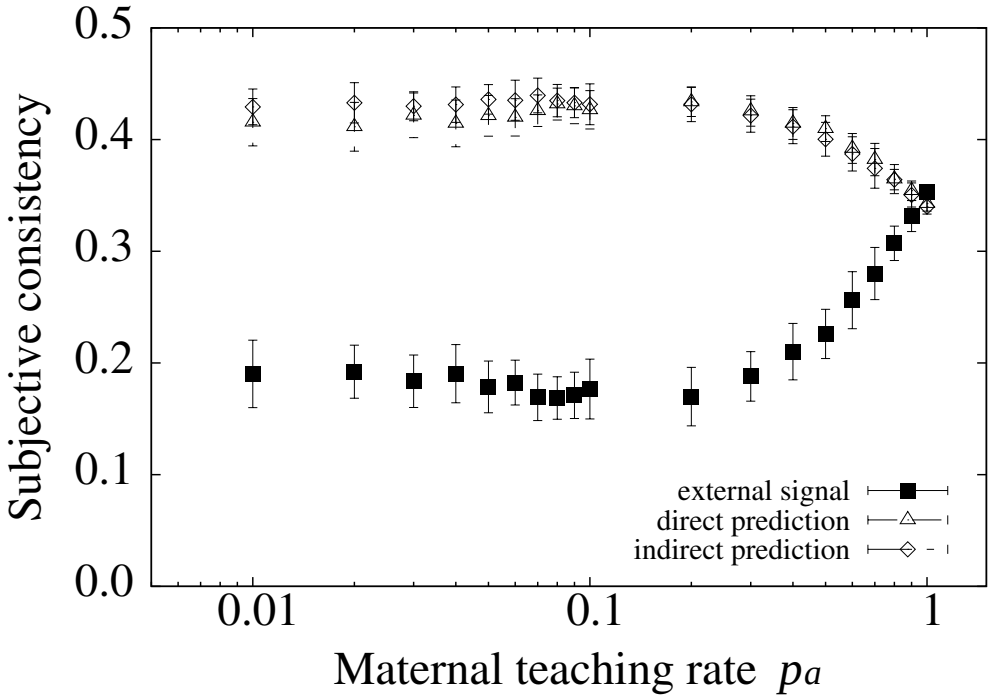


Figure 3.12: Average of average subjective consistencies among different three mappings during final 100 steps of learning with respect to maternal teaching rate p_a: subjective consistency for external signal (filled squares with solid line), direct prediction (blank triagle with broken line), and indirect prediction (blank diamond with dash-dotted line).

習しないのではなく，誤った対応を学習する原因であると考えられる．

3.5.4 実験 2: 実環境での養育者の応答を模した状況での検証

前節の実験では，3 種の教示率について同じであると仮定して，学習パフォーマンスを検証した．しかし，実際の乳児に応じる養育者の行動を分析した実験[20]によれば，養育者の模倣の頻度は非常に低く約 5 %程度，ラベル付けについては約 35 %程度と，行動によってその頻度が異なる．そこで，この観察実験結果に倣い，p_V を 0.05，p_S と p_D を 0.35 とした場合について，100,000 ステップの母子相互作用シミュレー

ションを 10 回実施した．η は 1.0 と 0.0 の 2 通りに設定し，パフォーマンスは前節と同様の方法により評価した．

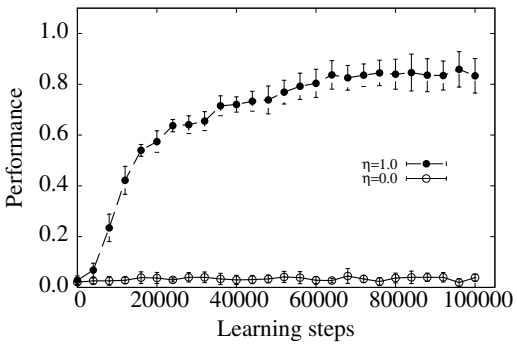
各マッピングの学習パフォーマンスの遷移を Fig.3.13 (a) (b) (c) に示す．通常の学習手法に従って対応を学習した場合 ($\eta = 0.0$)，語彙の学習 ((b), (c)) に関しては，教示率程度の想起が可能になるまでにとどまっており，模倣の学習 ((a)) に関しては，ほとんど対応を学習できていない．一方，提案手法に従って対応を学習した場合 ($\eta = 1.0$)，養育者の模倣をほとんど経験できない状況であっても，構音-聴覚マッピングの最終パフォーマンスが維持されていることがわかる．

$\eta = 1.0$ の場合の，各信号に対する主観的整合性の 100 ステップ毎の移動平均値の遷移 (Fig.3.14 (a) (b) (c)) をみてみると，外部信号に対する主観的整合性が学習が進むにつれて減少していくこと，すなわち養育者から与えられる外部信号をあまり信頼しないように学習が進んでいることがわかり，これにより，提案手法では高いパフォーマンスが維持されている．また，模倣の学習 (Fig.3.14(a)) は，語彙の学習 (Fig.3.14(b)(c)) に比べ，間接予測信号をより信頼して，つまり語彙を介した想起をより利用して進んでいることがわかり (Fig.3.14 (a) ◇の一点鎖線)，模倣の学習において，より対応を経験できる語彙の学習の経験が利用され，相互促進的な学習が実現されていることがわかる．この結果は，他のいずれかのマッピングについての教示率が非常に低い場合についても再現されると考えられる．すなわち，ロボットに対応を与える養育者の行動に偏りがあり，ロボットがあるマッピングにおいてはほとんど対応を学習できない状況であっても，提案手法により，他のマッピングの学習経験を利用することで，相互促進的に対応の学習が可能となる．

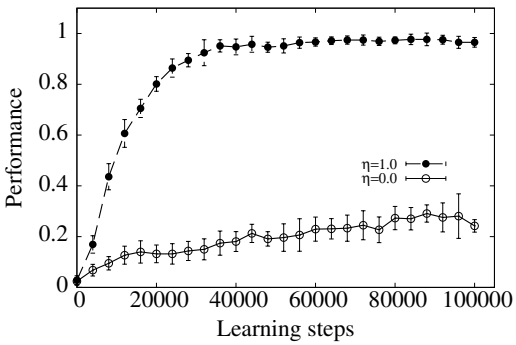
3.6 考察

3.6.1 学習手法としての位置づけ

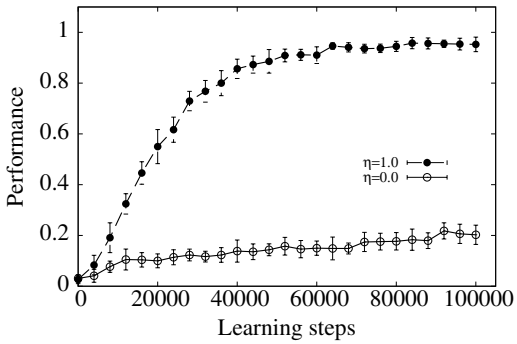
機械学習の一つに半教師有り学習[72] と呼ばれる学習手法がある．これは，データの識別器の学習問題において，データベースとしてラベル付けされたデータとともに



(a) Articulation-Sound Mapping



(b) Sound-Word Mapping

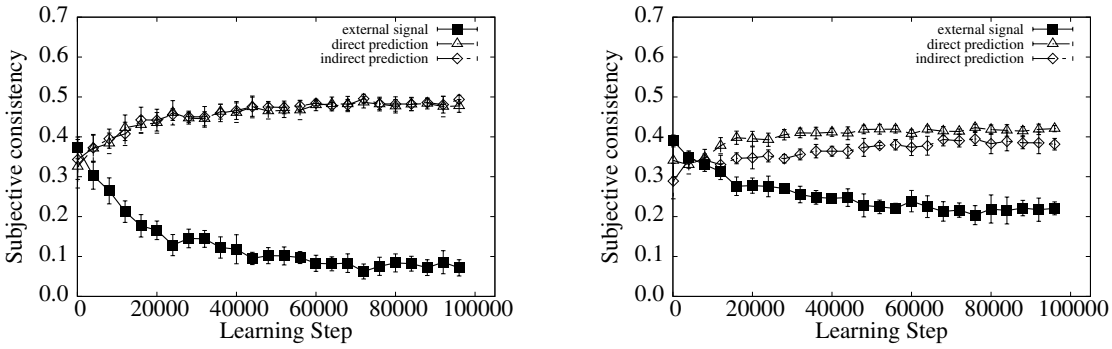


(c) Word-Articulation Mapping

Figure 3.13: Average transitions of learning performance of (a) articulation-sound mapping, (b) sound-word one, and (c) word-articulation one with the proposed subjective consistency ($\eta = 1.0$) and without it ($\eta = 0.0$)

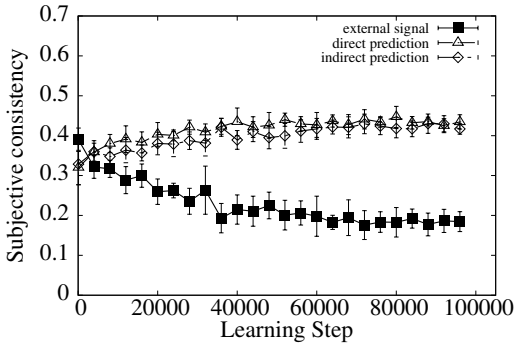
ラベル付けされないデータも与えられる時に，いかにして，高い汎化性能を得るかに注目した研究である[73; 74]．Nigam et al. や Ando and Zhang は，文書のカテゴリを識別する問題において，部分的にしかそのカテゴリのラベルが付与されていない文書のデータベースが与えられた時に，残りのラベル付けされていないデータについては，与えらえている部分的な対応関係から推定したラベルを用いる方法を提案している．これに対して本章で実施した実験では，教師データが完全でないことを想定している点で類似しているが，問題として，間違ってラベル付けされたデータが与えられてしまう，ということを想定している点が異なる．

また，本章では，音声模倣と語彙の二つの機能を同時に学習するためのメカニズム



(a) Articulation-Sound Mapping

(b) Sound-Word Mapping



(c) Word-Articulation Mapping

Figure 3.14: Average transition of subjective consistency calculating in learning (a) articulation-sound mapping, (b) sound-word one , (c) word-articulation one: subjective consistency for external signal (filled squares with solid line), direct prediction (blank triagle with broken line), and indirect prediction (blank diamond with dash-dotted line)

を提案したが，そのような複数機能の同時学習メカニズムは，これまでいくつか提案されている．Wolpert and Kawato が提案している MOSAIC [75]では，複数の学習モジュールの予測性能を評価し，環境の状態や文脈に対して適応的にモジュール割当と学習の重み付けを行うことで，各モジュールの効率の良い学習が実現されている．これに対して，提案手法では，学習すべきモジュールの割り当てには着目せず，音声模倣と語彙のそれぞれのモジュールの学習において，いかにして他のモジュールと相互促進的に学習を進められるかということに着目している点で異なると考えられる．

また，Uchibe and Doya が提案している CLIS [76]では，複数モジュールがある中で，各モジュールの状態価値関数に従って学習すべきモジュールを選択することで，単純なモジュールから学習が進み，さらにその学習における経験が複雑なモジュールの学習の手助けとなり，複雑なモジュールのみで学習させた場合よりも効率の良い学習が実現されている．提案手法も，一つのマッピングの学習に他のマッピングの学習経験を利用している点では同じである．しかし，CLIS では，状況に応じて適応的に選択されるモジュールによって得られた信号を，他のモジュールの学習に利用するのに対して，提案手法では，複数のモジュールによって得られた信号の競合的な統合によって学習に利用する信号を生成している点で異なると考えられる．

しかし，提案手法は，上記の半教師有り学習や複数機能の同時学習の手法を否定するものではなく，それらの手法とは相補的に機能しうると考えられ，これらと組み合わせた取り組みを行うことは今後の課題である．

3.6.2 乳児の発達との関連性

問題設定の考察

本章では，乳児の音声模倣と語彙の発達における，マッピングとカテゴリ化の 2 つの課題のうち，特にマッピングの課題に取り組んだ．そのため，本来連続的な信号である養育者の音声を，乳児が離散的な表象として認識可能であることや養育者と乳児が共通のカテゴリを有していることなど，実際の発達途上の乳児が備えているものを超える能力を，簡単のために仮定していた．しかし本来は，そのような認識の能力の発達，すなわちカテゴリ化を同時に考慮すべきである．音声の音圧や長さが，乳児が，聴取した音声から単語を切り出すためには重要であることが指摘されている[77]．つまり，音声の連続的な情報は，音声を一つの音韻或いは単語カテゴリとして認識することに影響すると考えられ，これはカテゴリ化の課題においては重要である．また，本研究では，養育者と乳児のカテゴリの数を同一としていたが，実際の養育者は，未熟な乳児に比べて圧倒的に多い音韻やラベルのカテゴリを有していると考えられる．さらに，そのように圧倒的に多いカテゴリ集合を持つ養育者の応答は，乳児にとって

は，観測する刺激の変化として現れてくるものであり，そのような変化が，乳児のカテゴリ化の学習に影響すると考えられる．このような問題は，基本的には本研究で扱った教師信号の選択の問題と捉えられる．例えば，音声模倣のためのカテゴリ化の課題では，乳児が連続的に生成した構音運動パターンのどの部分が，その後聴取した養育者の連続音声パターンのどの部分に対応するかなどの時間的な対応を見つけることが重要であり，これによって音声を音韻系列やラベルなどのカテゴリとして認識できるようになると考えられる．そのため，本章で提案した主観的整合性のアイデアを適用し，音声模倣と語彙の共発達過程において，どのように相互促進的に各表象のカテゴリ化が可能であるかを検証することが課題である．このようなカテゴリ化の問題に対しては，主観的整合性を導入したカテゴリ化の手法について，次章で詳しく述べる．

また，本章では，音声模倣と語彙獲得に注目して共発達を議論したが，他の機能が共発達に及ぼす影響も考慮する必要がある．実際の乳児では，ある種の反射的な行動が生誕直後には備わっていることや[37; 38]，親の視線の理解[78]を発達させていくことが知られている．これらの機能によって，一見乳児が親を模倣しているように親に感じさせることで，親の模倣行動を促進したり，乳児が親の視線方向を推定できるようになることで，教示対象を特定する手がかりが得られることが考えられる．従って，これらの発達に関する学習メカニズム[79; 63]と統合することで，より精緻な共発達モデルを得ることが今後の課題となる．

実験結果からの考察

実験1では，養育者の教示率が0.2 以上の場合は，提案手法により音声模倣と語彙の発達のための対応学習が可能となるが，それ以下の教示率では，誤った学習をしてしまうことがわかった．ここで提案手法が誤った学習をしてしまうとき，対応していないものを対応しているとみなす連合が起こったと考えられた．これは発達障害の乳児や，発達途上の健常の乳児が，ときおり誤った対応音韻の転換や言葉の誤用を一貫して示すことと対応する現象であると考察される．しかしながら本章のシミュレーションは，実際の母子相互作用を単純化し，養育者が固定の教示率で応じ続けること，ま

た一定数の物体が環境に置かれていることを前提としている点に注意が必要である．一方，実際の養育者の応答は，乳児の学習の様子や相互作用の履歴によって変わりうるものであり，また，環境中の物体も乳児の発育に応じて変化し，その結果，養育者の教示率も変移するものであることが予想される．従って，これらの環境側の変遷も考慮し，より精緻な問題設定の下でのシミュレーションを実施していくことが今後の課題となる．

実験 2 の結果から，提案手法は，実場面で見られるような，養育者が乳児にほとんど模倣を示さない状況においても，語彙の学習によって獲得した対応を利用することで，音声模倣のための対応の学習を可能とすることが示された．これは，言い換えれば，外部から与えられる音声信号を，一度語彙的なものへと歪ませて認識し，対応を学習すべきかを決定することで，音声模倣のための対応の学習が可能となった，と捉えることができる．一方，乳児は，発達初期では聴取した音声の音韻的な違いを識別可能であるが，語彙の学習が始まるに連れ，次第に音韻的に異なる音声でも同じ一つの語彙として識別するようになる[80]．そして，再び音韻的な違いを識別するようになるのは，さらに月齢が過ぎてからのようである[81]．本章の実験結果から，このような乳児の聴取した音声に対する音韻的・語彙的な識別の切り替えは，音声模倣と語彙の共発達過程において，相互の学習を相互促進的に利用し合う結果として起こる可能性が示唆される．すなわち，初期の音韻的な識別が語彙の学習を促進することで，語彙的な識別を可能とし，語彙的な識別がさらに音韻の学習を促進することで，再び音韻的な識別を可能とする結果として起こると考えられる．今後は，提案手法の主観的整合性の遷移から，どのように識別能力の変化が引き起こされるのかを検証し，音声模倣や語彙の共発達過程の理解を深めることが必要である．

第4章　語意のカテゴリ化のための主観的整合性に基づくマルチモーダルカテゴリゼーション

ロボットのカテゴリ化の従来研究では，ロボットが同時に観測した視覚や聴覚，触覚などの複数のセンサ情報がすべて特定の事物を表すものであることが想定されていた．しかし，実環境においては，常にユーザとロボットが注意を共有しているとは限らないため，その想定が保証されない．そのような場合，観測の共起確率に基づいてカテゴリ化する従来の方法だけでは，誤ったカテゴリを学習してしまう可能性がある．これに対し，本章では，従来手法[6]を拡張する形で，2 章で説明した主観整合性のアイデアを導入したマルチモーダルカテゴリゼーションの手法を提案する．これによりロボットが，ユーザが必ずしも共同注意しているとは限らない状況において，語意カテゴリを形成できることを示す．

4.1　はじめに

近年のロボット技術の発展に伴い，人と同じ生活環境で活躍するロボットの実現に期待がよせられている．これに対し，ロボットは人の振舞を観察し，それに示されている人の意図を理解できる必要があり，人の発話の意味（それが示す事物）の特定，すなわち語意の理解は，その典型的な課題であると考えられる．ユーザの発話について学習する場合，対象となる語意は，対面するユーザや環境に存在する事物に応じて異なることが予想されるため，ロボットはユーザと相互作用し，自律的に語意学習できることが望ましい．また実環境においては，常にユーザとロボットが注意を共有し

ているとは限らない．すなわち，ユーザが発する言葉が必ずしもロボットが注目している事物を表すとは限らない．従って，ロボットは，ユーザとの相互作用から得られたデータすべてを単純に学習に利用するのではなく，語意の学習に必要なデータを取捨選択できることが望ましい．本章では，そのような必ずしもユーザとロボットが同じ事物に注目しているとは限らない状況におけるロボットの語意学習の実現を目指す．

ロボットがユーザとの相互作用を通じて語意を学習するための手法は，これまでいくつか提案されている．Gorin et al. は，ユーザの発話による指令を可能にすることを目的として，ユーザの発話とシステムの所望の動作の対応を示したデータベースから，ユーザの発話を分類する手法を提案している[82; 83]．しかしこれらの研究では，システムの所望の動作はあらかじめカテゴリ化されていた．ユーザの発話や指示内容は，対面するユーザや環境に存在する事物に応じて異なることが予想されるため，それを事前に仮定できない状況も考慮すべきである．

これに対し，Roy and Pentland や Yu は，ユーザの発話と，ユーザの発話が示す物体の関係を学習する課題において，発話とその発話に対応する物体についての観測情報のデータベースから，この両方ともをカテゴリ化する問題を扱った[9; 10]．また，田口らや中村らは、さらに考慮するモダリティの数を増やすことで，高精度な，あるいは人と同様のカテゴリ化ができる可能性を示している[11; 6]．しかし，これらの研究では，ユーザがロボットが注目している物体についての発話を行うなど，マルチモーダルな観測情報はモダリティ間で対応するものを捉えていることが想定されており，この想定が崩れる際に，どのようにシステムが対応すべきかには焦点はあてられていなかった．

一方，ユーザとロボットが必ずしも同じ事物に注目するとは限らない状況において，ロボットが人の発話音声とその発話に対して自身がとるべき行動の対応関係を学習する手法が提案されている．しかしながら，ロボットがとった行動に対して，予め設定した目標へ到達したかどうか[84]，あるいはユーザからの報酬[85]などの外的な評価が与えられることが想定されている．ロボットが，物体の特徴抽出器などロボットの特定の行動とは直接的な関係がみえにくいものについて学習する場合，ユーザがそれを適切に評価できるかは不明である．そのため，外的な評価に頼らない，自律的な学

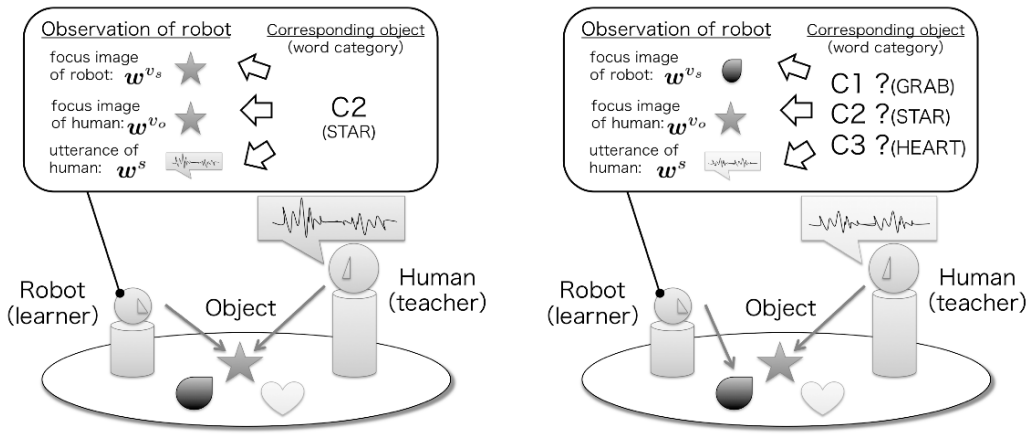
習機構の開発が望まれる．

そこで本章では，ユーザとロボットが必ずしも同じ事物に注目しているとは限らない状況においてロボットが自律的に語意を学習するための手法を提案する．ここでの問題は，ロボットの観測に，注目している事物とは対応しないデータが含まれることである．例えば，人がロボットが見ている物体の特定に失敗する場合，ロボットが見ている物体 (ロボットの注目画像特徴) と聴取した人の発話音声 (人の音声特徴) は，それぞれ別の事物を表す．しかし，その場合でも，人が見ている物体（人の注視画像特徴）と発話音声は同じ事物を表す，すなわち，部分的にはある事物に対応するデータが含まれることがある．本章では，マルチモダールな観測のモダリティ間の整合性を評価することで注目すべきモダリティを決定する主観的整合性を導入し，部分的な対応データを利用したカテゴリ化を実現できることを示す．以下では，まず，4.2 節において，本章で想定するユーザとロボットが必ずしも同じ事物に注目しているとは限らない状況における語意学習について説明し，4.3 節でこれに関する従来手法[6]を述べた上で，これを拡張する形で提案手法のアイデアを説明し，実装方法について述べる．4.4 節では，提案手法の有効性を評価するために実施した実データを用いた実験と計算機シミュレーションの結果を示し，最後に，本章で提案する手法の位置づけと今後の展望について述べる．

4.2 語意学習状況

人 (教示者) とロボット (学習者) と物体が存在する環境を考える (Fig.4.1)．ロボットは，環境中のいずれかの物体をランダムに選択し，それを注視する．ロボットの語意学習の教示者となる人は，ロボットと同様にいずれかの物体を注視し，同時に語を発声する．このような状況において，ロボットは，ロボットの注視先の画像特徴ベクトル $\boldsymbol{w}^{v_s}$ ，人の注視先の画像特徴ベクトル $\boldsymbol{w}^{v_o}$ 及び人が発する語の音声特徴ベクトル $\boldsymbol{w}^{s}$ を観測できるとし，本章では，これら 3 種の特徴ベクトルの対応関係を学習することを語意学習の課題とする．

また本章では，人とロボットが同じ事物に注目しているとは限らないことを仮定



(a) An example scene when human and robot focus the same object

(b) An example scene when human and robot donot focus the same object

Figure 4.1: Assumed environment of robot that learns words from human

し，人がロボットの見ている物体と同じ物体に注目する確率 (対応率) P_c を導入する．Fig.4.1 (a) に示すように人とロボットが常に同じものに注目する場合，すなわち $P_c = 1$ の場合，各観測 $\boldsymbol{w}^{v_s}$，$\boldsymbol{w}^{v_o}$，$\boldsymbol{w}^{s}$ は同じ物体に対応する特徴となる．これまでの研究では，このように同時に観測された特徴をカテゴリ化することで特定の物体と特徴の対応関係，すなわち，語意が学習されていた．しかし実際には，人がロボットの注目物体の特定に失敗したり，人がロボットの注目物体以外 (例えば，ロボットの顔) を見ながら物体の名称を発したり，あるいは，ロボットと同じ物体を見てはいるがその物体の名称以外を発したりする場合が起こる，すなわち $P_c < 1$ となる場合が起こると考えられる．そのような場合，$\boldsymbol{w}^{v_s}$，$\boldsymbol{w}^{v_o}$，$\boldsymbol{w}^{s}$ はそれぞれ異なる物体に対応する特徴となる (Fig.4.1 (b))．そのため，同時に観測された特徴を単純にカテゴリ化するだけでは，実際には物体とは対応しない特徴もその物体に対応する特徴として捉えてしまう可能性がある．

一方，このような場合でも，観測される複数のモダリティの一部は同じものを捉えたものである場合があると考えられる．例えば，人の音声がロボットの見ているものとは対応しないが，人自身の見ているものとは対応する場合や、人がロボットの方を

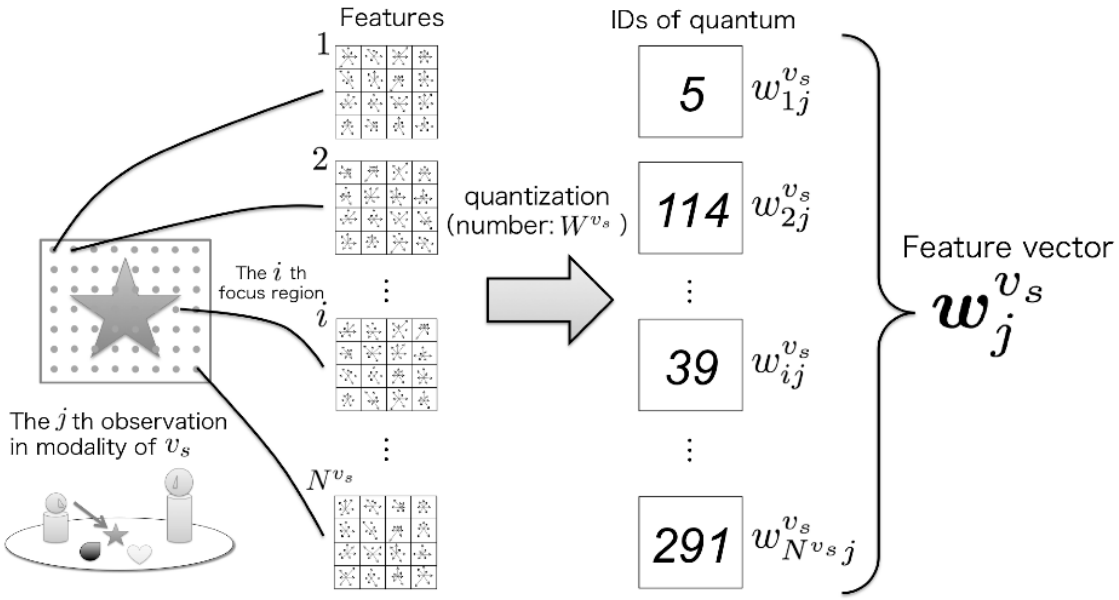


Figure 4.2: An example of caluculation process of feature vectors

見ながら，ロボットの見ているものに関する語を発する場合などである．このようなとき，観測の一部，ここでは $\boldsymbol{w}^{v_o}$ と $\boldsymbol{w}^s$ あるいは $\boldsymbol{w}^{v_s}$ と $\boldsymbol{w}^s$，は同じ物体に対応する特徴であるため，その部分的な対応関係の学習だけでも進められることが望ましい．しかし，観測された特徴のうち，どのモダリティの特徴が対応しているものであるかは状況に依存し事前には規定できないため，観測に応じて適応的にその対応をロボット自身が判断できることが必要であると考えられる．本章では，そのような部分的な対応をロボットが主観的に抽出するための主観的整合性を導入し，それに基づいた語意学習手法を提案する．

4.3 主観的整合性に基づくマルチモーダルカテゴリゼーション

本章では，マルチモーダルカテゴリゼーションの従来手法である Multimodal LDA[6] を拡張する形で，観測されるマルチモーダル情報が部分的にしか対応しないことがありうる状況に対処するためのアイデアである主観的整合性に基づいたカテゴリ化を導入し，その実装について述べる．

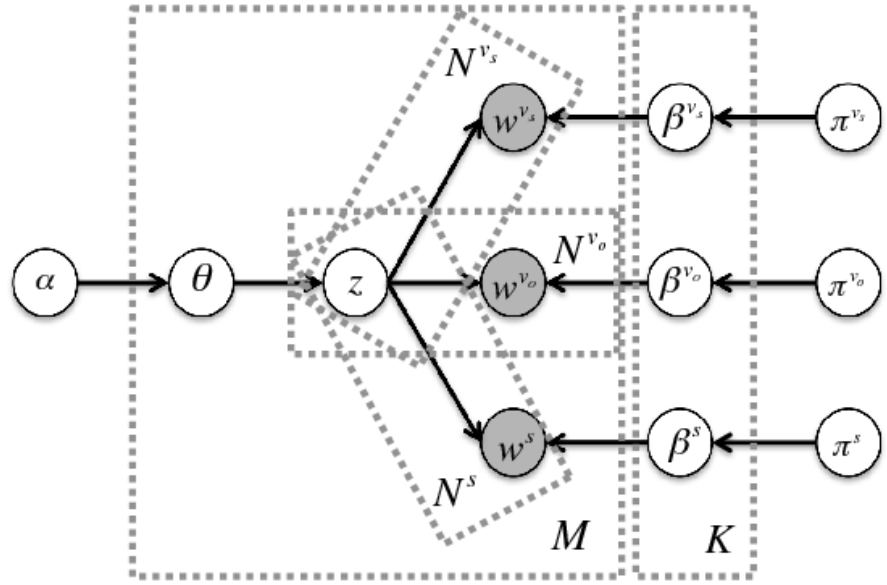
4.3.1 基本的な学習手法: Multimodal LDA

Multimodal LDA [6]は，元々は文書分類の手法として提案された LDA [86]をマルチモーダル情報を扱えるように拡張した手法であり，マルチモーダル情報をそれらの潜在変数で対応づけ，その対応に基づいて各モダリティのカテゴリを形成するものである．

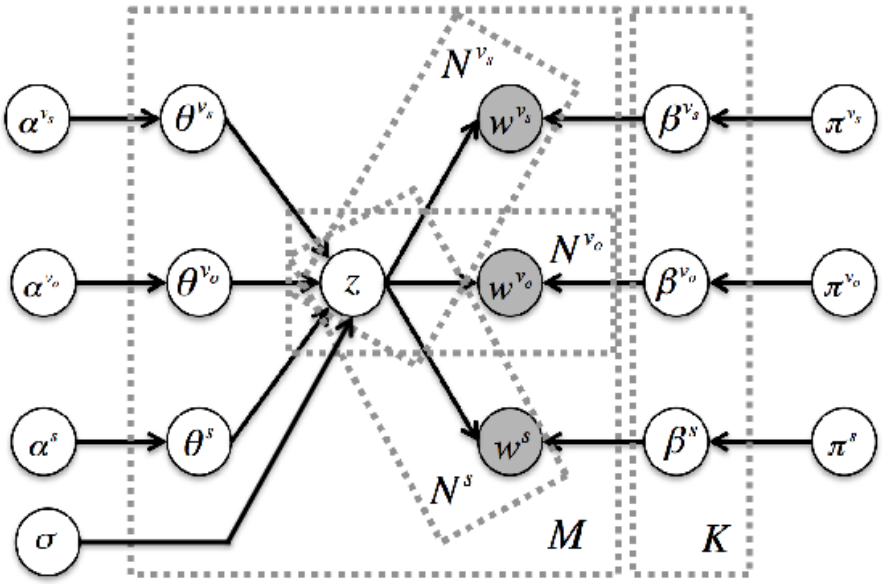
Multimodal LDA におけるパラメータの推定方法は以下の手順によって実現される．ロボットが複数のモダリティにより M 回の観測をすることを考え，一度の観測でモダリティ m において特徴ベクトル $\boldsymbol{w}^m$ が観測されるとする．これは，モダリティ m における N^m 個の各注目領域の情報を量子化した特徴を並べた N^m 次元ベクトルである (Fig.4.2)．ここでモダリティ m の各注目領域の情報が量子化される特徴の種類は W^m である．

Mulimodal LDA では，このような特徴を，Fig.4.3 (a) のグラフィカルモデルに示す生成過程によって得られるものと見なす．ここで，$\boldsymbol{\theta}$ は潜在変数の多項分布パラメータ, $\boldsymbol{\beta}^m$ はモダリティm の特徴の多項分布パラメータ，α, π^m は，それらのディリクレ事前分布パラメータ (ハイパーパラメータ) である．このグラフィカルモデルにおいて，潜在変数 z_{mij} は，$\boldsymbol{\theta}$, $\boldsymbol{\beta}^m$ を周辺化した，

$$p(z_{mij}=k|\boldsymbol{z}^{-mij},\boldsymbol{w}^m,\alpha,\pi^m) \propto \left(\frac{N_{kj}^{-mij}+\alpha}{N_{j}^{-mij}+K\alpha}\right)\frac{N_{ml(i)k}^{-mij}+\pi^m}{N_{mk}^{-mij}+W^m\pi^m} \tag{4.1}$$



(a) previous method



(b) proposed method

Figure 4.3: Graphical models of (a) previous method and (b) proposed method for the word learning of robot. Some notations in (a) are changed from original ones.

に従ったギブスサンプリングのアルゴリズムにより推定できることが示されている [6]．ここで，$i \in \{1, \ldots, N^m\}$ は，量子化した特徴ベクトルのインデックスであり，注目領域に対応する．$j \in \{1, \ldots, M\}$ は観測回，z_{mij} は j 番目の観測におけるモダリティ m の特徴ベクトルにおける i 番目のインデックスの特徴を生成した分布を示す潜在変数（語に対応するカテゴリ）である．K は潜在変数の数 (学習するカテゴリ数)，$l(i) \in \{1, \ldots, W^m\}$ は特徴ベクトルにおける i 番目のインデックスの特徴の量子 ID である．N_* は，M 回の観測を通じて，添字 $_*$ で特定される特徴が観測された回数を表す．また，上付きの添字 $^{-mij}$ は観測された全ての特徴を要素とする集合 U から j 番目の観測におけるモダリティm の特徴ベクトルにおける i 番目のインデックスの特徴を除いた集合 U^{-mij} についての量を表す指標である．従って具体的には，N_{kj}^{-mij} は U^{-mij} の集合に含まれるもののうち， j 番目の観測で得られる特徴の潜在変数が k であった回数であり，同様に N_{mk}^{-mij} はモダリティ m の特徴の潜在変数が k であった回数，$N_{ml(i)k}^{-mij}$ はモダリティm の $l(i)$ の潜在変数が k であった回数である．

式 (4.1) の右辺乗算の第一項はおよそ N_j^{-mij} と N_{kj}^{-mij} の比によって増減する項であるため全体として，同じ観測回の（共起する）別のモダリティの特徴の潜在変数も考慮して現在の特徴の潜在変数を k とするべき程度を表しているとみなすことができる．右辺乗算の第二項はおよそ N_{mk}^{-mij} と $N_{ml(i)k}^{-mij}$ の比によって増減する項であり，全体として，現在のモダリティの$l(i)$ が，別の観測回や同じ観測回の特徴ベクトルにおいて現在注目しているものとは別のインデックスの特徴も考慮したときに，潜在変数 k についてどれほど特有的に観測される特徴であるかを表しているとみなすことができる．すなわち，式 (4.1) の右辺は，現在の観測回の注目モダリティの特徴ベクトルにおいて注目しているインデクッスの特徴の潜在変数を「すべてのモダリティが示す潜在変数との共起性」(右辺乗算第一項) と,「注目している特徴の潜在変数における特有性」(右辺乗算第二項) に基づいてサンプリングすることを意味する．

式 (4.1) 中の N_* ははじめ一様乱数からサンプリングされた z_{mij} によって初期化しておき，特定の mij において，式 (4.1) に従ったサンプリング結果に基づき更新してく．具体的には，特定の mij において k とサンプリングされた場合,

$$N_{mijk} = N_{mijk} + 1 \tag{4.2}$$

と加算し，各 N_* の添字が特定する範囲に含まれている場合，これに応じて更新する．式 (4.2) により更新した N_* を用いて式 (4.1) によるサンプリングを繰り返していくことで，z_{mij} が収束し，N_* の収束値 $\bar{N}_*$ が得られる．これを用いて，最終的なパラメータの推定値が

$$\hat{\theta}_{kj} = \frac{\bar{N}_{kj} + \alpha}{N_j + K\alpha}, \quad \hat{\beta}^m_{w^m k} = \frac{\bar{N}_{mw^m k} + \pi^m}{\bar{N}_{mk} + W^m \pi^m} \tag{4.3}$$

のように算出される．

以上の手順により，式 (4.1) の右辺乗算第一項の働きにより，人とロボットが常に同じ物体に注目する場合，その共起性を手がかりに，各モダリティの特徴ベクトルにおける各インデックスの特徴が示す語のカテゴリ $z = k$ を推定することができ，正しい対応関係を示すパラメータの学習ができる．しかしながら，人とロボットが必ずしも同じ物体に注目していない場合，すなわち，いずれかのモダリティに観測される特徴が他のモダリティに観測される特徴が示す語カテゴリと一致しない分布から生成される場合，共起性が高く評価されないために，正しい語カテゴリのサンプリングへのバイアスが期待できない．

4.3.2 主観的整合性に基づく拡張

人とロボットが必ずしも同じ物体に注目していない場合の Multimodal LDA の問題に対し，本節では，マルチモーダル情報の部分的な対応関係をロボットが主観的に抽出し，そのような場合でも，正しい語カテゴリのサンプリングを可能にする語意学習手法を提案する．

これまでの手法では，観測回毎に，一つのカテゴリ分布パラメータ $\boldsymbol{\theta}$ を求めていた．しかしこれでは，マルチモーダルデータの一部が他と対応しないことを表現できない．そこで，本章では，観測したモダリティ毎に潜在変数の分布パラメータ $\boldsymbol{\theta}^{v_s}, \boldsymbol{\theta}^{v_o}, \boldsymbol{\theta}^{s}$ を考慮したグラフィカルモデルを考える (Fig.4.3 (b))．ここで，4.3.1 節と同様に，j 番目の観測のモダリティ m の特徴ベクトルにおける i 番目のインデックスの特徴を生

成した分布を示す潜在変数 z_{mij} を

$$p(z_{mij}=k|\boldsymbol{z}^{-mij},\boldsymbol{w}^m,\boldsymbol{\alpha},\pi^m) \propto \left(\sum_{r\in\{v_s,v_o,s\}} \lambda^r_{mij}\frac{N^{-mij}_{rkj}+\alpha^r}{N^{-mij}_{rj}+K\alpha^r}\right)\frac{N^{-mij}_{mw^mk}+\pi^m}{N^{-mij}_{mk}+W^m\pi^m} \tag{4.4}$$

に従ってサンプリングすることを考える．ここで，λ^r_{mij} は， z_{mij} が，別のモダリティ r のデータを観測したことで得られる潜在変数の分布からどの程度生成されうるかの度合い (以後，主観的整合性と呼ぶ) を表している．前章で説明したように，主観的整合性に基づいて各モダリティのデータを重みづけることで，他のモダリティのデータとは対応しない整合性の低いデータを排除して対応学習できる．本章では，この方法を応用し，式 (4.4) のようにモダリティ毎に示される潜在変数の分布を主観的整合性によって重みづけることを考える．主観的整合性は，以下のように定義する．

$$\lambda^r_{mij} = \frac{\kappa^r_{mij}}{\sum_{o\in\{v_s,v_o,s\}}\kappa^o_{mij}}. \tag{4.5}$$

ここで，κ^r は，モダリティ r が示す潜在変数の分布が，他のモダリティが示す分布とどの程度一致しているかを表し，以下の式で表す．

$$\kappa^r_{mij} = \exp\left(-\sum_{\substack{q\in\{v_s,v_o,s\}\\ q\neq r}} D\left(P(z_{mij}|\tilde{\boldsymbol{\theta}}^r_j), P(z_{mij}|\tilde{\boldsymbol{\theta}}^q_j)\right)/\sigma\right), \tag{4.6}$$

ただし，$P(z_{mij}|\tilde{\boldsymbol{\theta}}^{m'}_j)$ は，j 番目の観測において，モダリティ m' のデータを観測したことで得られる潜在変数の分布に相当し，パラメータは，

$$\tilde{\theta}^{m'}_{kj} = \frac{N^{-mij}_{m'kj}+\alpha^{m'}}{N^{-mij}_{m'j}+K\alpha^{m'}} \tag{4.7}$$

と計算される．また，$D(P(z_{mij}|\tilde{\boldsymbol{\theta}}^{m'}_j), P(z_{mij}|\tilde{\boldsymbol{\theta}}^{m''}_j))$ は，$P(z_{mij}|\tilde{\boldsymbol{\theta}}^{m'}_j), P(z_{mij}|\tilde{\boldsymbol{\theta}}^{m''}_j)$ の分布間の距離を表し，σ は その距離に対する逆感度パラメータである．これらの分布は，z_{mij} に関する離散分布であるので，分布間の距離は，以下のように Intersection で評価する．

$$D\left(P(z_{mij}|\tilde{\boldsymbol{\theta}}^{m'}_j), P(z_{mij}|\tilde{\boldsymbol{\theta}}^{m''}_j)\right) = \sum_{k=1}^{K}\min\left(\tilde{\theta}^{m'}_{kj}, \tilde{\theta}^{m''}_{kj}\right). \tag{4.8}$$

式 (4.4) は，先行研究におけるサンプリングの式 (4.1) と比べて，右辺乗算第一項が異なることがわかる．式 (4.4) の右辺乗算第一項は，式 (4.1) のように，単純にすべてのモダリティが示す潜在変数との共起性を評価するのではなく，よく対応する，すなわち整合性のとれるモダリティが示す潜在変数との共起性を評価するように変更されたものと見なすことができる．これにより，あるモダリティのデータが示す潜在変数が，他のモダリティが示すものと一致していない場合でも，一致しているモダリティの潜在変数を重視し，それが示す語カテゴリへバイアスしてサンプリングできると考えられる．

N_* は，従来研究と同様に更新していくが，その更新量は，現在の mij における潜在変数の分布が，他のモダリティが示す潜在変数の分布とどの程度一致しているかに応じて変更するよう，以下のように計算する．

$$N_{mijk} = N_{mijk} + \kappa^m_{mij}. \tag{4.9}$$

最終的なパラメータの推定値は，N_* の収束値を用いて，

$$\hat{\theta}^m_{kj} = \frac{\bar{N}_{mkj} + \alpha^m}{\bar{N}_{m}j + K\alpha^m}, \quad \hat{\beta}^m_{w^m k} = \frac{\bar{N}_{mw^m k} + \pi^m}{\bar{N}_{mk} + W^m \pi^m} \tag{4.10}$$

と算出される．

以上のように，整合性の高いモダリティとの共起性を手がかりに語カテゴリを推定することで，あるモダリティで観測された特徴が示す語のカテゴリが，他のモダリティで観測された特徴が示す語のカテゴリと一致しない場合でも，共起性が高く評価されない問題を抑制できると考えられる．また，その整合性に応じて，N_* の更新量を決定することで，注目している特徴が，サンプリングされた語カテゴリが示す特徴をどの程度捉えたものとみなすかを調整できる．これにより，ある特徴が示す語と他のモダリティによってバイアスされてサンプリングされた語が異なる場合には，その特徴をその語カテゴリの特徴としないようにカテゴリ化できると考えられる．

4.4 実験

提案手法の有効性を確かめるために，ロボット及び計算機シミュレーションを用いて以下の二つの実験を実施した．

実験 1:実データを用いた検証　提案手法が実環境において有効であるかを検証する．具体的には，まず，人とロボットが常に同じ物体に注目する状況での人の教示を記録し，そのデータの組み合わせを対応率にもとづき非対応データと入れ替える．これにより，仮想的に人とロボットが同じ物体に注目しない状況を作り，ロボットが提案手法によって正しく語意を学習できるかを検証する．

実験 2:人工データを用いた詳細な検証　提案手法の効果を詳細に検証するために，人工データを用いた計算機シミュレーションを実施する．具体的には，3 つのモダリティの内，いずれか 1 つのモダリティのデータが他の 2 つのモダリティのデータと正しく対応しない場合 (実験 2-1)，特定のモダリティのみ他のモダリティと対応しない場合 (実験 2-2) に提案手法により対応しないデータがどのように抑制できるようになり，どのように対応関係が獲得されていくかを検証する．

4.4.1 実験 1：実データを用いた検証

基本設定

Fig.4.4 のように，ロボットと人の間に複数の物体がランダムに配置されており，ロボットはこのうち一つをランダムに選択して注視するとする．そして人は，一定の対応率 P_c でロボットの注視物体の語を正しく教示する．すなわち確率 P_c で，ロボットと同じ物体を注視すると同時に，この物体の名称を発し，確率 $1 - P_c$ で，そうでない非教示的な振る舞い（注視および発話）をするとする．ロボットはこのいずれの場合においても，ロボットの注目先の画像特徴ベクトル $\boldsymbol{w}^{v_s}$，人の注目先の画像特徴ベクトル $\boldsymbol{w}^{v_o}$，および人が発した音声特徴ベクトル $\boldsymbol{w}^{s}$ を観測する．これら一連の観測

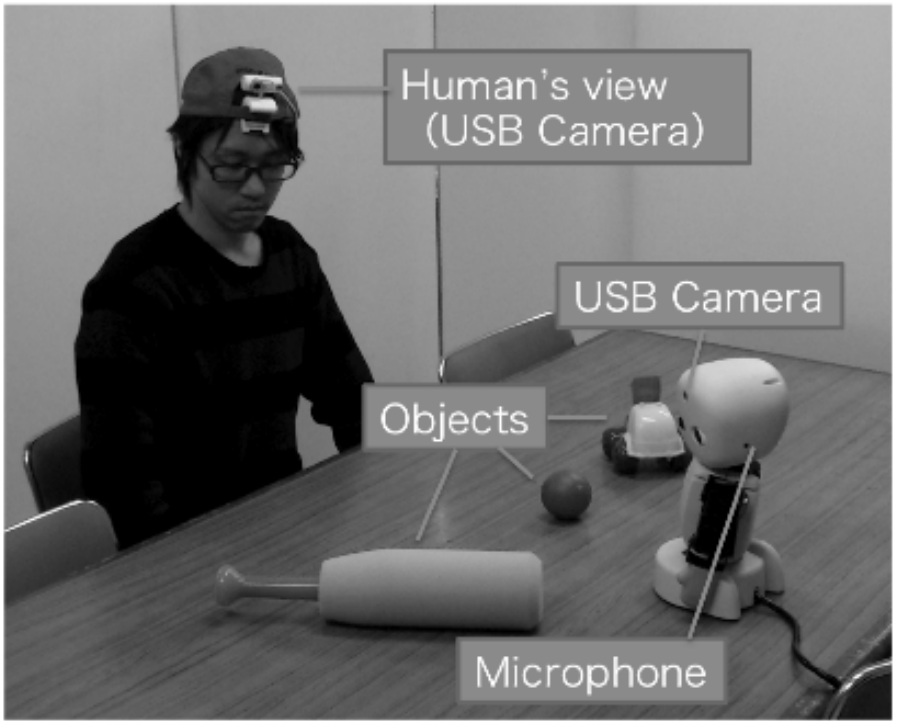


Figure 4.4: The environment in experiment 1

をまとめて 1 ステップと呼び，M ステップ繰り返して得られたデータを用いて語意学習を行う．

正しい教示がなされる場合，$\boldsymbol{w}^{v_s}, \boldsymbol{w}^{v_o}, \boldsymbol{w}^{s}$ は対応する物体に関する特徴を含むデータの組となるため，対応データと呼ぶ．これに対し，人が非教示的に振る舞う場合に得られるものを非対応データと呼ぶ．本章では，非対応データとして，いずれか二つのモダリティの特徴は対応しているが，残りの一つのみがこれとは異なるものを含むものとし，どの二つのモダリティが対応するものとなるかはランダムに決定されるとする．これは，

(1) ロボットの注目先の画像特徴が対応しない場合

人は自分の見ているものの名称を発しているが，ロボットの注目物体と異なる物体を見ている

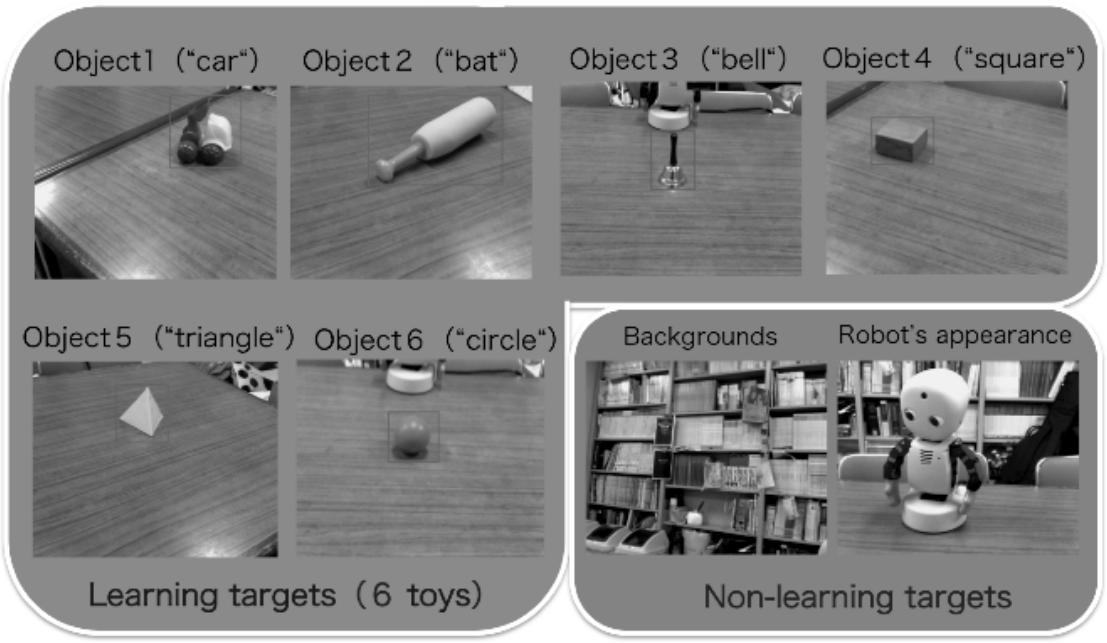


Figure 4.5: Examples of image data of learning target and non-learning target

(2) 人が発した音声特徴が対応しない場合

ロボットと人が見ているものは同じであるが，人がこの物体の名称以外の語を発する

(3) 人の注目先の画像特徴が対応しない場合

人はロボットが見ているものの名称を発するが，人がその物体を見ていない

という 3 つ場合が等確率に起こることを仮定したことに相当する．なお，これら以外に，

(4) どのモダリティの特徴も対応しない場合

人はロボットが見ているものには注目しておらず，また自分の見ているものの名称以外の語を発する．

という場合も考えられる．しかし，本章では，マルチモーダルな観測に含まれてしまう他とは対応しないモダリティのデータをいかに排除した学習が可能であるかに焦点

をあてるため，この (4) は想定しないこととした．

対応データ，非対応データの一例として，人の注目先の画像を Fig.4.5 に示す．本実験では，Fig.4.5 に示すような 6 つの子供の玩具を学習対象とし，各玩具について 20 回観測しておくことで得た各モダリティのデータを再サンプルすることを各ステップにおける観測とみなす．これらのうち，いずれかのモダリティの観測を対応しないものに入れ替えたものが非対応データとして扱われる．例えば，Fig.4.5 の物体 1 を観測とする場合，非対応データとして入れ替えられる候補となる観測データは，物体 2 や物体 3 など学習対象の内の別の物体に関するものに加え，ロボットの外見などの非学習対象を捉えた観測である．学習対象でないものの観測は，モダリティごとにそれぞれ 20 回ずつ行った．モデルのハイパーパラメータは，経験的に，$\alpha^m = 1.0$, $\pi^m = 1.0\,(m \in \{v_s, v_o, s\})$, $\sigma = 0.5$ とした．

特徴ベクトル

ロボットが観測する画像，音声特徴として，中村ら[6]と同様に，各モダリティのセンサ情報を量子化した特徴ベクトルを用いる．

ロボットの注目先の画像特徴ベクトル $\boldsymbol{w}^{v_s}$ は，ロボットの頭部の USB カメラから取得した画像の物体領域 (ROI) において 10 ピクセル間隔で観測される，合計 4 個のマルチスケール (半径 4,8,12,16 ピクセルの領域) の各領域における DSIFT 特徴量[87]を 500 個のコードブックベクトルを用いて量子化した特徴を並べたベクトルとする．一般に，カメラ画像から物体を認識する場合には，オクルージョンの問題があるが，ここでは，カメラ画像には一つの物体のみ含まれることとした．

人の注目先の特徴量ベクトル $\boldsymbol{w}^{v_o}$ の取得には，人の注目点の画像情報を得る必要がある．これに関して，人の顔画像から推定する方法等が提案されている（例えば，[88]）が，ここでは対応しないデータを排除した学習に主眼を置くため，この問題は扱わず，人の頭部に USB カメラを設置し取得した画像を人の注目画像として用いる．$\boldsymbol{w}^{v_o}$ は，このようにして得た人の注目画像に，$\boldsymbol{w}^{v_s}$ の取得と同様の処理を施して取得する．

音声特徴ベクトル $\boldsymbol{w}^{s}$ は，ロボットの頭部側面にあるマイクロフォンから取得した

音声から 20 ms 毎の各領域における 12 次元の MFCC 特徴量[89]を 50 個のコードブックベクトルを用いて量子化した特徴を並べたベクトルとする．

評価方法

主観的整合性を用いないで学習した場合 (式 (4.1)(4.2) に基づいた学習手法：従来手法) と，主観的整合性を用いて学習した場合 (式 (4.4)(4.9) に基づいた学習手法：提案手法) の学習後の分類性能を比較する．分類性能は，学習後に，物体毎に新たに 10 回づつ観測して得られたデータ，$\boldsymbol{w}^{v_s}_{test}$ ，$\boldsymbol{w}^{v_o}_{test}$，$\boldsymbol{w}^{s}_{test}$ を，学習時に推定したパラメータを用いて，

$$\begin{aligned}\hat{\boldsymbol{z}} &= \arg\max_{\boldsymbol{z}} P(\boldsymbol{z}|\boldsymbol{w}^{v_s}_{test}, \boldsymbol{w}^{v_o}_{test}, \boldsymbol{w}^{s}_{test}) \\ &= \arg\max_{\boldsymbol{z}} \int P(\boldsymbol{z}|\boldsymbol{\theta}) P(\boldsymbol{\theta}|\boldsymbol{w}^{v_s}_{test}, \boldsymbol{w}^{v_o}_{test}, \boldsymbol{w}^{s}_{test}) d\boldsymbol{\theta}\end{aligned} \tag{4.11}$$

のように分類した結果が，正しく対応する語意であるかどうかで評価する．ただし，主観的整合性を用いる場合は，$\boldsymbol{\theta} = \{\boldsymbol{\theta}^{v_s}, \boldsymbol{\theta}^{v_o}, \boldsymbol{\theta}^{s}\}$ である．また，$P(\boldsymbol{\theta}|\boldsymbol{w}^{v_s}_{test}, \boldsymbol{w}^{v_o}_{test}, \boldsymbol{w}^{s}_{test})$ は，式 (4.1)(提案手法の場合は，式 (4.4)) を，学習時の収束値 $\bar{N}_{mw^m k}$ を初期値として用いて再計算することで求める．

分類した結果が正しいかどうかは，F 値で評価した．F 値は，適合率 P と再現率 R の調和平均として，以下のように計算される．

$$F = \frac{2 * P * R}{P + R}, \tag{4.12}$$

ただし，モデルが獲得した語のカテゴリが，実際のどの語に対応するかは，モデル自体からは判断できないので，ここでは，獲得したカテゴリ k' を実際の語 k とおき，式 (4.12) を計算する処理を，すべての $k, k' \in \{1, \ldots, K\}$ の組み合わせについて行い，最も F 値の高かった組み合わせを正しい分類とした．適合率 P と再現率 R は，以下のように計算される．

$$P = \sum_{k=1}^{K} \frac{a_k}{a_k + c_k}, \quad R = \sum_{k=1}^{K} \frac{a_k}{a_k + b_k}.$$

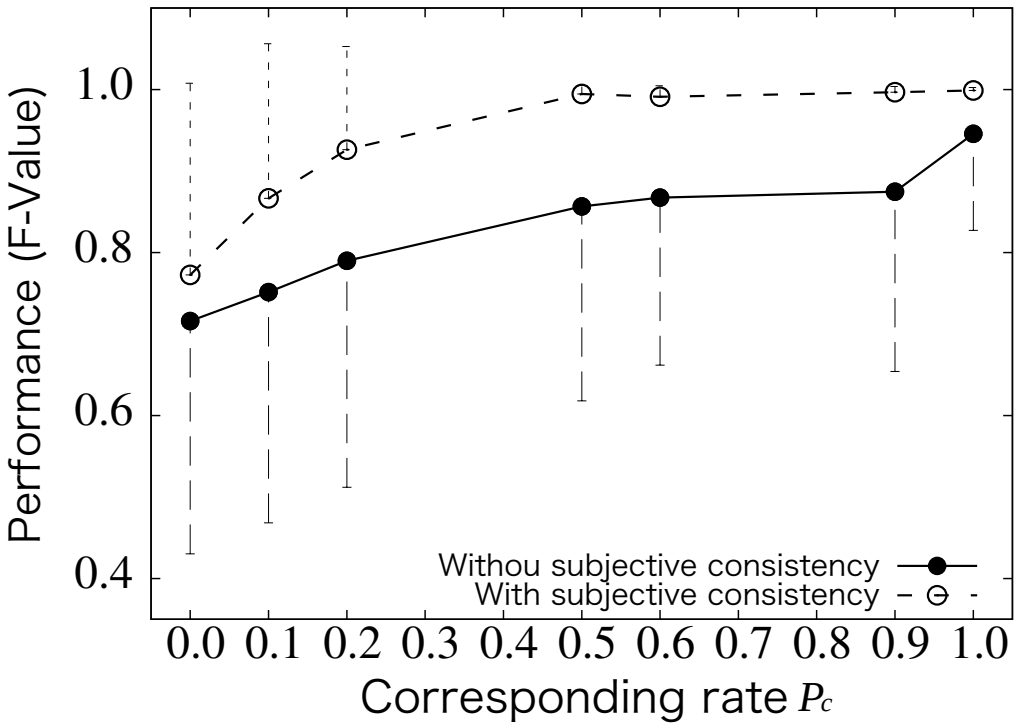


Figure 4.6: Average F-value in 10 steps with respect to corresponding rate P_c

ここで，a_k は，モデルがテストデータをカテゴリ k と分類した回数を表し，c_k は，カテゴリ k と分類されたデータの内，実際には対応する語 k 以外のデータである場合の数であり，b_k は，実際の語は k のデータであるが，それをカテゴリ k 以外に分類した場合の数である．

実験結果

対応率 P_c を 0.0，0.1，0.2，0.5, 0.6，0.9, 1.0 としたそれぞれの場合で，一つの物体につき 60 回観測したデータを用いて，総ステップ数 $M = 360$ とした語意学習のシミュレーションを実施した．Fig.4.6 は，P_c に基づいて対応データ・非対応データの組み合わせを変えて 10 回学習させた際の，学習後の分類性能の平均値と標準偏差を示したものである．

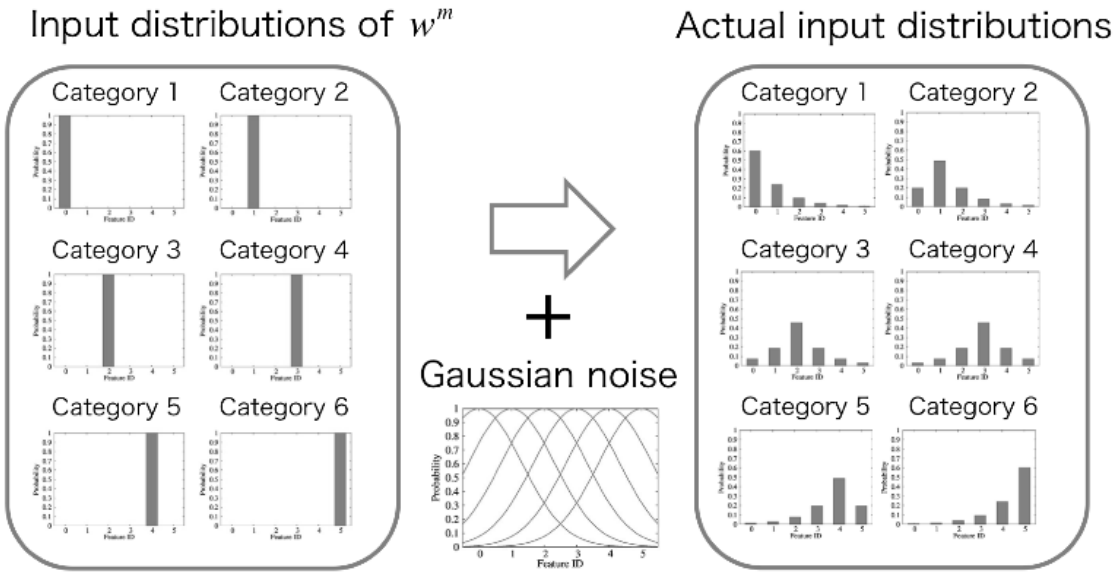


Figure 4.7: Input distributions of artificial data in experiment 2

Fig.4.6 より，全ての対応率の条件において，提案手法 (白丸) の方が主観的整合性を用いない従来手法 (黒丸) よりも高い分類性能を示していることが分かる．

全ての観測が対応するものである場合 ($P_c = 1.0$)，従来手法 (黒丸) でも，高い分類性能が得られているが，対応率が下がると ($P_c < 1$)，分類性能が低下している．これに対して，提案手法を用いた場合，中程度の対応率 ($P_c = 0.6, 0.5$) まで，対応率が 1.0 であるときと同程度の分類性能が維持されていることが分かる．また，対応率がかなり低くなった場合 ($P_c = 0.2, 0.1, 0.0$) でも，対応率が中程度である場合に比べて分類性能は低下するものの，従来手法に比べて高い分類性能が得られている．すなわち，提案手法を用いることで，対応率に対して頑健に学習が可能であると言える．

本実験では，対応率が 0.5 まで低下しても，完全に教示する状況 ($P_c = 1.0$) と同程度の分類性能が得られている．この状況は，観測データのうち半分は非対応データである場合，すなわち，人が 2 回に 1 回は完全には教示しない状況に相当しており，提案手法により，人がロボットに語意を教示する際の注意に関する制限を緩和可能であることが示唆される．

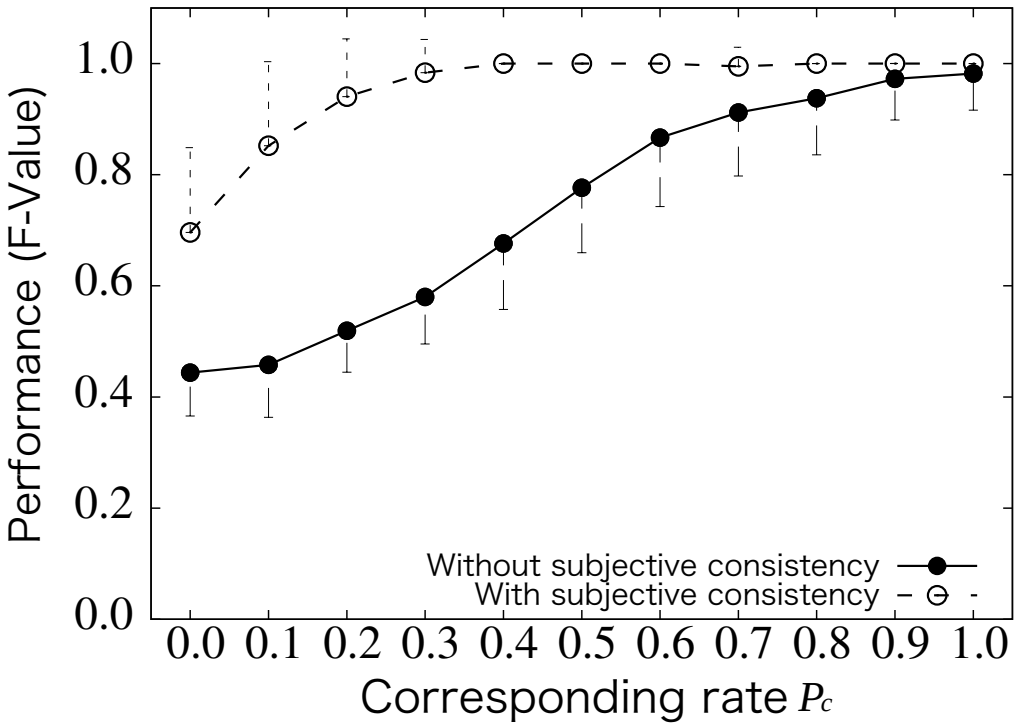


Figure 4.8: Average performance of clacssification in 100 steps with respect to correponding rare P_c

4.4.2 実験 2: 人工データを用いた詳細な検証

次に，提案手法による学習の効果を詳細に分析するために，人工データセットを用いた検証を行った．実験 2-1 として，あるモダリティのデータが他のモダリティのデータと正しく対応しない場合に，提案手法により対応しないデータを用いた学習が抑制される過程をより詳細に分析する．そのために，すべてのモダリティの特徴の種類を $W^m = 6$ と仮定し，実験 1 と同じ設定の語意学習シュミレーションを実施する．さらに，実験 2-2 として，特定のモダリティのデータのみ他のモダリティのデータと対応しない場合での検証を行う．そのために，注目率 P_a を導入し，それに応じてモダリティ v_s のみに非対応データが観測されるとした語意学習シミュレーションを実施する．

実験 2-1(モダリティ間で正しく対応するデータが得られない場合): 基本設定，特徴ベクトルおよび評価方法

実験1 と同様に，3 つのモダリティの内，いずれか 1 つのモダリティのデータが他の 2 つのモダリティのデータと正しく対応しない状況を想定する．ただし，対応率 P_c の確率で，3 つすべてのモダリティで同じカテゴリのデータ (以後，全対応データと呼ぶ) が観測され，$1 - P_c$ で，いずれか 2 つのモダリティのみ同じカテゴリのデータが観測され，残り 1 つのモダリティには別のカテゴリのデータが観測される (以後，部分対応データと呼ぶ)．

簡単のため，カテゴリ数 K を 6，特徴の種類数 W^m をすべてのモダリティで同じ 6 とする．特徴は，各カテゴリに対応する K 個の分布から生成されるとし，1 ステップにおいてその分布から生成された 100 点の特徴を特徴ベクトルとする．理想的な観測状況においては，1 ステップで観測されうる特徴ベクトルは各カテゴリにおいて排他的であるとした．すなわち，特徴は，Fig.4.7 左 のような，一次元の排他的な分布に従うと仮定する．ただし，観測ノイズを考慮し，それぞれ中心位置の異なる正規ノイズを加えた分布から生成されるものとする (Fig.4.7 右)．

1 ステップ毎に，ロボットが K 個のうちのいずれかのカテゴリに注目するか決定される．そして対応率に応じて，人がいずれのカテゴリの物体を注視するのか，またどのカテゴリの物体の名称を発するのか，がランダムに選択される．各モダリティでは，これらのカテゴリに対応する観測特徴量の分布から生成されたデータが観測される．一つの物体について 100 ステップ観測するものとし，語意学習の総ステップ数 M は 600 とする．モデルのハイパーパラメータは，実験 1 と同じ，$\alpha^m = 1.0$, $\pi^m = 1.0\,(m \in \{v_s, v_o, s\})$, $\sigma = 0.5$ とした．

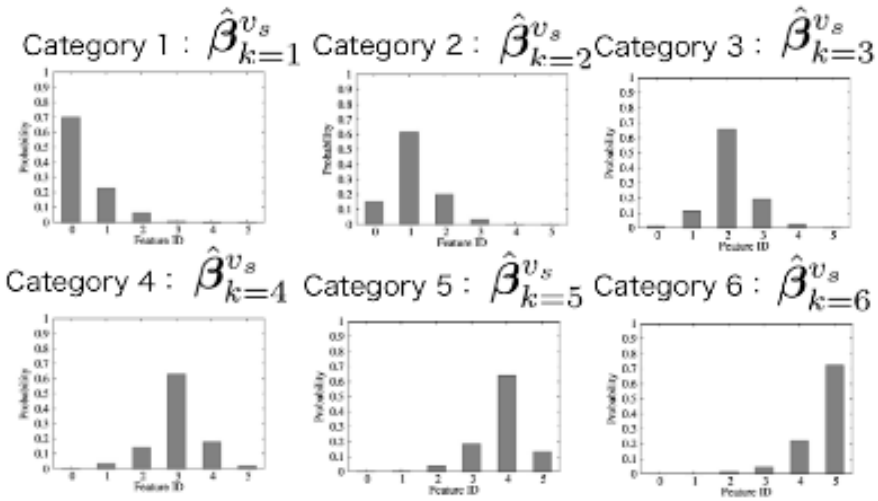
提案手法の有効性は，前節と同様の方法で評価する．ただしテストデータとして，カテゴリ毎に，各モダリティの特徴ベクトルが従う分布から新たにデータを 100 個ずつ生成したものを用いる．

実験結果 (2-1)

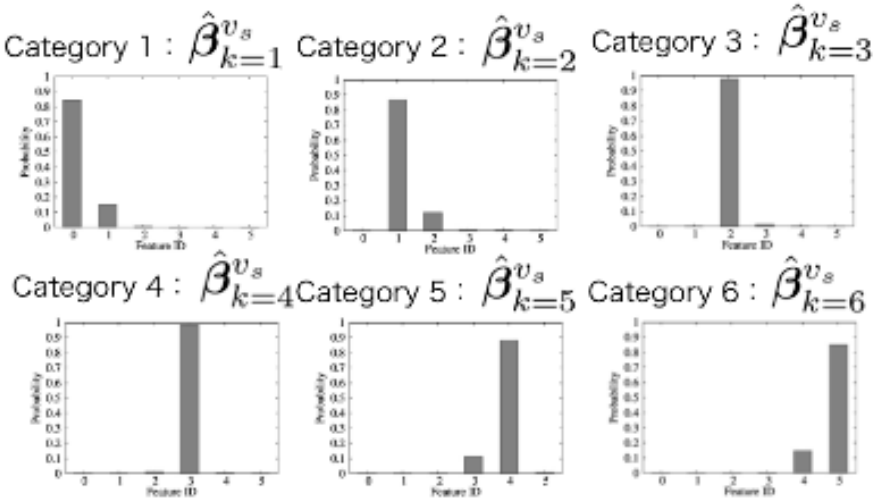
対応率 P_c を 0.0 から 1.0 まで 0.1 ずつ変化させ，600 ステップの語意学習シミュレーションを 100 回実施した．Fig.4.8 は，P_c 毎の，学習後の分類性能の 100 回の平均値と標準偏差を示している．Fig.4.8 より，$P_c = 1.0$ のとき，すなわち，全対応データのみが観測される場合では，提案手法と従来手法の両方で，高い分類性能が得られていることがわかる．しかしながら，対応率が低くなるにつれ，従来手法では $P_c = 0.8$ 付近から分類性能が低くなっているのに対して，提案手法では $P_c = 0.3$ 付近まで高い分類性能が維持されており、提案手法の有効性が確認できる．なおこの実験で，$P_c = 0.0$ であっても比較的高い分類性能が得られているのは，$P_c = 0.0$ であっても，毎回，いずれか二つのモダリティでは対応するデータが観測されるためであり，このような場合でも，対応しないモダリティのデータを排除することにある程度成功していることを示している．

Fig.4.9 は，$P_c = 1.0$ の場合のモダリティ v_s について推定された特徴量の分布 $\hat{\boldsymbol{\beta}}_k^{v_s}$ の一例を示している．従来手法では，カテゴリの形成において，裾野の広い分布が推定されている (Fig.4.9 (a)) のに対し，提案手法では，裾野の狭い分布が推定されている (Fig.4.9 (b)) ことが分かる．これは，従来手法では，観測されるノイズをそのまま含めた形で，特徴量分布を推定しているのに対し，提案手法では，そのようなノイズを排除した真の分布 (Fig.4.7 左) に近い分布を推定できていることを示している．これは提案手法により，観測データに一様に含まれるノイズが，特定のカテゴリとは対応しないデータ，すなわち非対応データとして扱われ，抑制された結果であると考えられる．

$P_c = 0.5$ のときの，ギブスサンプリング時にサンプリングされた潜在変数（カテゴリ）の遷移を Fig.4.10 に示す．Fig.4.10 は，各ステップにおいて得られた観測特徴の下での，100 回のギブスサンプリングを経て得られた潜在変数の適切さの遷移を表す．図中の白抜き逆三角形は，各ステップでギブスサンプリングの収束結果として，正しいカテゴリがサンプリングできた割合を，白抜き四角形は，対応しないカテゴリをサンプリングした割合について，100 回実施したシミュレーションにおける平均と分散を表す．サンプリングされた潜在変数が，正しいカテゴリを表すものであるかどうか

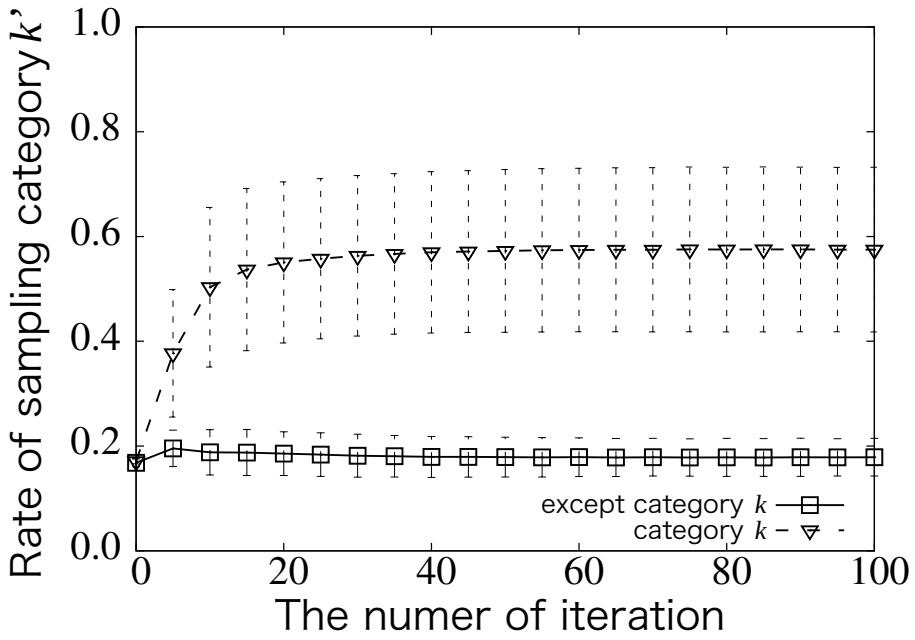


(a) With out subjective consistency

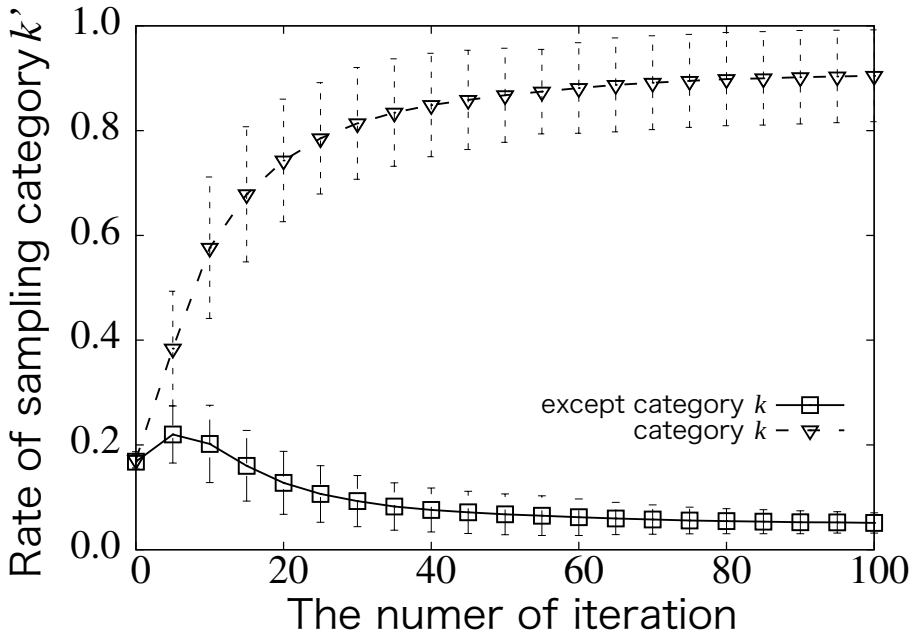


(b) With subjective consistency

Figure 4.9: An example of feature distribution of each category formed under the $P_c = 1.0$



(a) Without subjective consistency



(b) With subjective consistency

Figure 4.10: Transitions of sampled categories in learning process (a) without proposed subjective consistency and (b) with it under $P_c = 0.5$

は，学習終了時に，4.4.1 節の分類性能の評価方法に基づいて，潜在変数 k' と真のカテゴリ k の対応を特定し，それを遡って適用することで得た．従来手法（主観的整合性を用いないで学習）の場合 (Fig.4.10 (a))，正しいカテゴリのサンプリングが学習途中から，ある程度できるようになっていくが，対応しないカテゴリのサンプリングは学習終了時まで，排除できていないことが分かる．一方，提案手法（主観的整合性を用いた学習）の場合 (Fig.4.10 (b)) では，より高い頻度で，正しいカテゴリをサンプリングできるようになっていくことが分かる．語意学習は基本的に，観測特徴をサンプリングされた潜在変数に割り付ける形で進むため，提案手法の狙い通り，他のモダリティとは対応しない，すなわち整合性の低いデータが，語意学習に反映されないようになっていくことで，高い分類性能の維持が実現されていると言える．

異なる対応率の条件においても，提案手法により Fig.4.10 に示したようなサンプリングのバイアスが可能であるかを評価するため，対応率が 1 以外の条件において，学習終了時の部分対応データに対する主観的整合性 λ の（式 (4.5)）の平均値と標準偏差を Fig.4.11 に示す．図中の白抜き四角形は，部分対応データの中で対応するモダリティのデータ (以後，対応情報と呼ぶ) に対する主観的整合性を表し，白抜きの逆三角形は，対応しないモダリティのデータ (以後，非対応情報と呼ぶ) に対するものを表す．Fig.4.11 より，対応率が低くなると ($P_c < 0.3$)，その差は小さくなっていくものの，非対応情報に対する主観的整合性は，対応情報に対する主観的整合性に比べ低い値となっていることが分かる．すなわち，提案手法により整合性の高い情報が学習により貢献し，主観的整合性で適切に非対応情報を抑制できていることを示している．

実験 2-2(特定のモダリティのみ他のモダリティと対応しない場合)：基本設定，特徴ベクトルおよび評価方法

3 つのモダリティの内，特定の 1 つのモダリティのみにおいて，他の 2 つのモダリティと異なるカテゴリの観測特徴量が生成することがありうるとする．ここで想定しているのは，人がロボットに対して教示は行う，あるいはロボットの語意学習にとって正しい教示データとなる振舞をする (人の注視先と発話した名称は対応する) が，ロボットが何に注目しているかにはあまり関心がない状況である．すなわち，人が示す

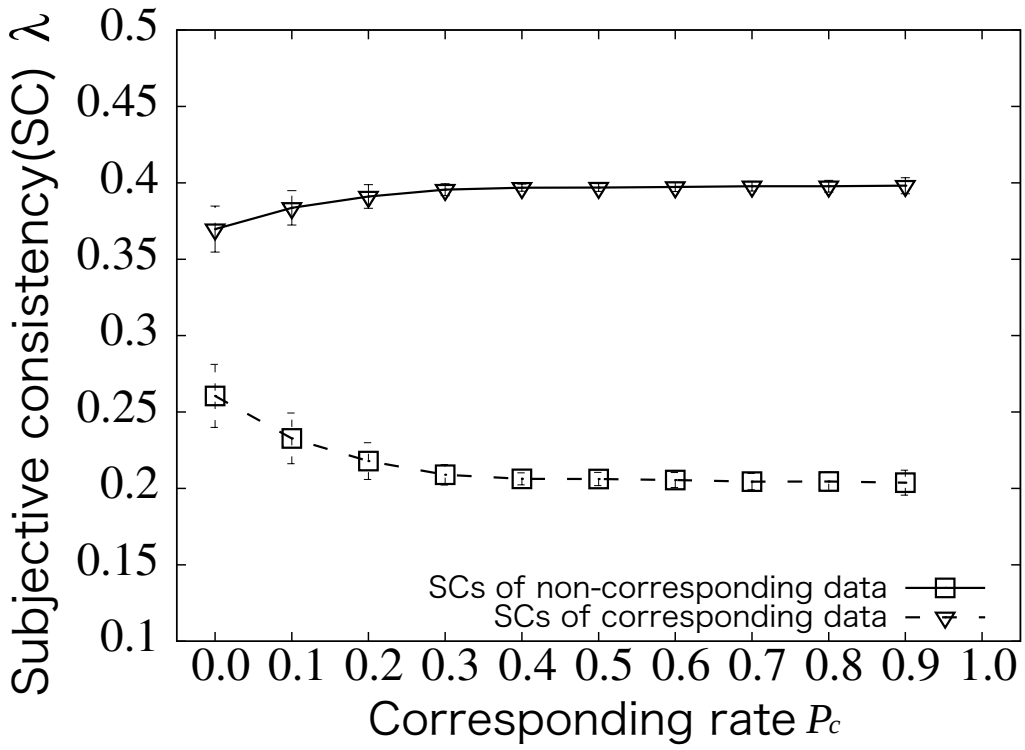


Figure 4.11: The final subjective consistency with respect to corresponding rate

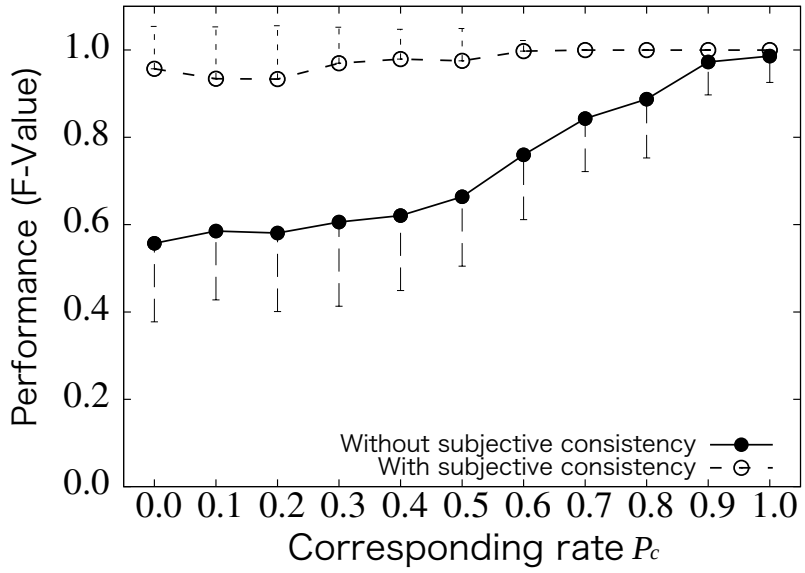
特徴が，ロボットが注目する特徴と必ずしも一致しないような状況であり，4.4.1 節で説明した非対応データの想定のうち (1) が特に起こり易いと想定できる場合である．そこで，注目率と呼ぶ確率 P_a を導入し，確率 P_a で全てのモダリティの観測が同一のカテゴリの特徴量分布から生成され，確率 $1 - P_a$ で，モダリティ v_o と s の観測は同じカテゴリの特徴量分布から生成されるが，v_s の観測は異なるカテゴリから生成されるとする．提案手法は，このように部分対応データが偏る場合においても有効に作用し，非対応データの影響を低減した語意学習が可能であると考えられる．

特徴ベクトル，その他各種パラメータは実験 2-1 と同じものを用い，提案手法の有効性も同様の方法で評価する．

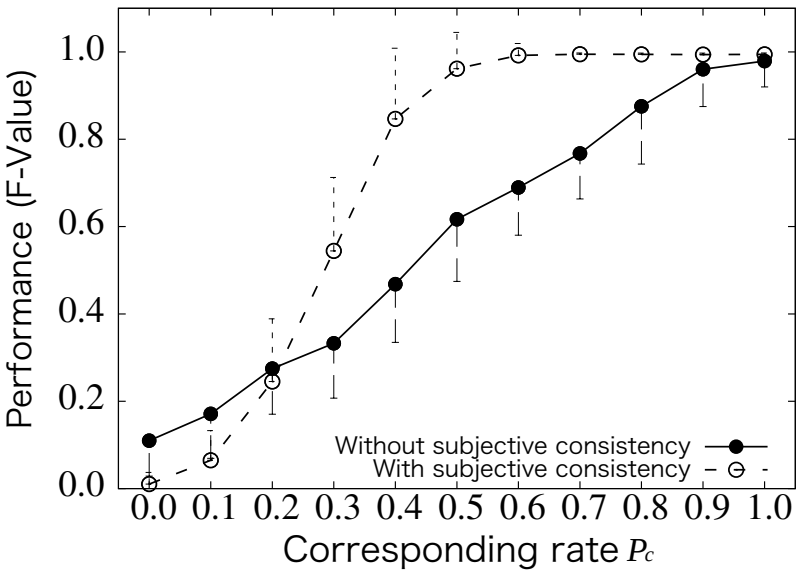
実験結果 (2-2)

注目率 P_a を 0.0 から 1.0 まで 0.1 ずつ変化させ，600 ステップの語意学習シミュレーションを 10 回実施した．非対応データが観測されるモダリティとそれ以外のモダリティの学習を比較をするために，テストデータに対して，学習後にモダリティ v_o と s を用いて分類した場合 (Fig.4.12 (a))，と v_s を用いて分類した場合 (Fig.4.12 (b)) の分類性能を見た．モダリティ v_o と s を用いた分類においては，従来手法 (Fig.4.12 (a) 黒丸) では，注目率が低くなるにつれ分類性能が低下しているのに対し，提案手法 (Fig.4.12 (a) 白丸) では，高い分類性能が維持できていることが分かる．また，Fig.4.12 (b) から，モダリティ v_s を用いた分類においても，従来手法（黒丸）に比べ，提案手法（白丸）が，注目率が 0.5 付近以上のある程度高い状況では，高い分類性能を維持できていることが分かる．しかし，注目率 P_a が 0.5 付近以下に低くなると，提案手法を用いても，モダリティ v_s の分類性能は低下していく．これは本実験設定では，モダリティ v_s に他のモダリティ v_o と s と対応するデータが観測される確率は $1 - P_a$ であるため，提案手法により，モダリティ v_s のデータが，学習からほとんど排除されてしまうことによるものと考えられる．

$P_a = 0.5$ のときの，ギブスサンプリング時に提案手法よって計算された各モダリティの主観的整合性の遷移を Fig.4.13 に示す．図中の白抜き逆三角形と白抜き菱形はそれぞれモダリティ v_o と s についての主観的整合性であり，ほぼ重なったまま学習初期に上昇したのち，その値を維持し，0.4 弱に収束していることが分かる．これに対し，図中の白抜き四角形はモダリティv_s についての主観的整合性であり，学習初期から下がり始め，0.2 強程度に収束していることが分かる．このように提案手法では，ある特定のモダリティのみが他のモダリティと対応する確率が低い場合においても，対応しないモダリティの観測を学習に反映させる程度を主観的整合性に基づいて低減させることで，部分的に対応する観測についての学習を阻害せず，データ中の対応するモダリティの観測を効率的に利用した学習が可能になっていると考えられる．



(a) With data in modalities of v_o and s



(b) With data in modality of v_s

Figure 4.12: Average performance of clacssification in 100 steps with respect to correponding rare P_c (a) with data in modalities of v_o and s, and (b) with ones in v_s

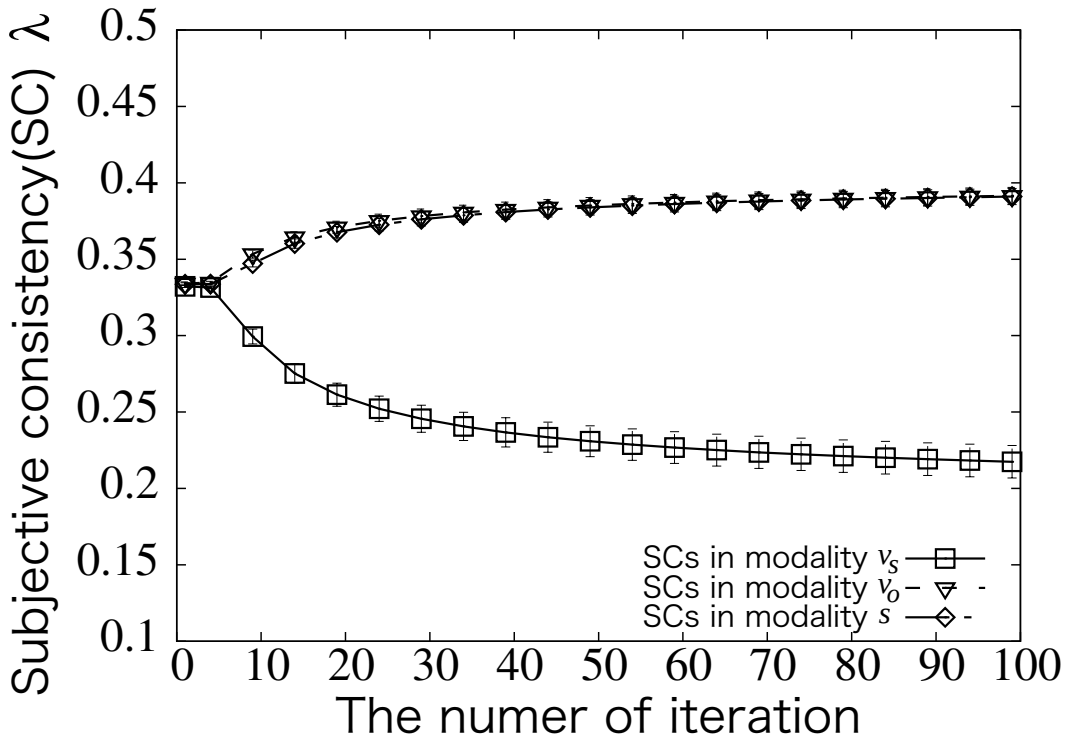


Figure 4.13: Transitions of subjective consistencies of data in each modality under $P_c = 0.5$

4.5 考察

4.5.1 ロボットの語意学習手法としての位置づけ

ロボットの言語学習における課題として，岩橋は，記号接地，パースペクティブ，信念共有，同時協調的行為，発話交代の 5 つをあげている[90]．本章で扱った，人とロボットが必ずしも同じ物体に注目するとは限らない状況におけるロボットの語意学習は，どのように言葉と事物を対応付けるかの記号接地と，どのように他者と共通の知識・信念を持つかの信念共有の課題に関連すると考えられる．

記号接地に関して，これまでいくつか研究されており，4.1 節で述べたマルチモー

ダル情報の共起確率に基づいた手法や[9; 10; 11; 6]，事前に想定されたバイアスを利用する手法[91; 62; 63]，が提案されている．これに対して本章は，マルチモーダル情報を一つの語意に接地している点や主観的整合性という一種のバイアスを学習に利用している点で類似している．しかし，本章では，問題として必ずしも正しく対応するマルチモーダル情報が共起して観測されるとは限らない状況を取り扱う点，また，形状類似性や顕著性などの単一のモダリティで規定されるバイアスではなく，整合性という複数のモダリティで規定されるバイアスを導入している点，で異なると考えられる．

信念共有に関して，Steels は，複数のエージェントが自分自身の語意知識に基づいて物体への名前付けをし合うことで他者の知識との違いを知り，それを共通化する手法を提案している[92]．また，岩橋は，ユーザとロボットがどの程度信念を共有しているかを表現した確信度関数を導入し，ロボットがそれに基づいて発話することで，ユーザとロボット間で相互適応的に信念を一致させる手法を提案している[93]．本章で提案した手法でも，主観的整合性を導入し，他者の知識との一種の違いを推定している点では同じであるが，整合性の高い情報を他者と共通化すべき知識として積極的に学習に利用している点が特徴である．しかし，提案手法は，上記の学習手法を否定するものではなく，それらの手法とは相補的に機能しうると考えられ，これらと組み合わせた取り組みを行うことは今後の課題である．

4.5.2 提案手法のスケーラビリティ

語意カテゴリ数に関して

本章で実施した実験では，語意カテゴリの数は予め既知としてた．しかし，ロボットが学習すべき語意の数は，環境やユーザの要求に応じて異なることが予想されるため，自動で決定されることが望まれる．これに関して，階層ディリクレ過程[94]を用いたノンパラメトリック化の手法が既に提案されている[95]．本章で提案した主観的整合性とこのようなノンパラメトリック化の手法を組み合わせた取り組みを行うことが今後の課題である．

モダリティ数に関して

本章では，3 つのマルチモーダル情報から語意を学習することを想定していたが，ロボットが学習すべき語意数の増加に伴って，必要なモダリティ数も増加すると考えられる．例えば，「赤い」という語意の学習には，形状のみではなく色の情報が必要であり，「置く」などの動作を表す語意の学習には，運動の系列情報が必要であると考えられる．学習に必要なモダリティが増えると，ある語意を表す正しい対応関係はより部分的にしか観測されなくなると予想され，提案手法はそのような状況において，より有効に働くと考えられる．今後は，多くのモダリティを必要とする問題設定において，提案手法の有効性を詳細に検証していくことが必要である．

学習対象の複雑さに関して

本章で実施した実験において，提案手法により，観測データに一様に含まれるノイズのような特定のカテゴリとは対応しないデータを抑制してカテゴリを形成できることを確認した．これは提案手法により，特定の学習対象 (カテゴリ) に対して排他的に対応する特徴の抽出が可能であることを示しており，カテゴリ弁別における有効性が期待できる．しかし学習対象が増える，あるいは個々の学習対象が様々な特徴を持つようになると，あるカテゴリの特徴が多くのカテゴリに含まれない場合でも，別の特定のカテゴリに含まれてしまうような場合が生じやすくなり，排他性を重視しすぎることで，そのような特徴が抽出されにくくなる恐れがある．そのような場合には，提案モデルにおける整合性に対する逆感度パラメータ σ を調整することを検討すべきである．すなわち σ を少し高い値に設定することで，観測データに一様に含まれるノイズのように，整合性がかなり低い情報の抑制効果は保持したまま，一部のカテゴリ間のみで共有される特徴のように，ある程度の整合性を示す情報の利用が可能になると考えられる．この学習対象の複雑さに応じて逆感度パラメータを動的に調整する機能については今後の課題である．

対応率に関して

本章では，あるモダリティのデータが他のモダリティのデータと正しく対応しない状況を，相互作用を通して一定の対応率でマルチモーダル情報の正しい対応が観測される状況と想定して実験した．しかし，実環境においてロボットに応じるユーザは，ロボットとの相互作用の履歴や自身の状態に応じて教示の仕方を変えうるものであり，その場合対応率は変遷していくと考えられる．そのため，今後は，実世界で人とロボットの相互作用を観察する実験を実施し，実際のユーザがどのように (どのような対応率で) ロボットに応答しうるかを考慮して，提案手法の有効性を検証する必要がある．しかし，極端に対応率が低くなってしまった場合などは，提案手法のみでは学習が困難になる可能性があり，そのような場合には，逐次的な学習や相互作用の履歴から環境側へ働きかける仕組みが必要になると考えられる．これに対して，追加的なギブスサンプリング[96] を用いたオンライン化の手法[97]や教示者への能動的な発話手法[98; 99]が利用できると考えられ，それら手法と提案手法を組み合わせた手法を開発することが今後の課題の一つである．本実験では，人とロボットの注意が共有されないことの影響およびそれに対する提案手法の有効性の検証に主眼を置くため，音声情報は簡易的に人の発話音声から抽出した MFCC を特徴量として取り扱っていた．これはロボットが発話に変換可能な形式ではないため，このままでは上述のロボットの発話手法とそのまま組み合わせるのが困難である．しかしこれは，音声認識器の認識結果を特徴量とすることで対応可能であると考えられる．文書分類の手法と同様に，複数の単語からなる人の発話の音声認識の結果は単語のカウントベクトルとして表現し，多項分布によってモデル化できる．提案手法では，入力変数は多項分布に従うと仮定しているため，この拡張に対応可能である．また，音声情報が単語で表現されるため，単語ごとに予めロボットの音声を登録しておくことで，学習結果をそのままロボットの発話に利用することができると考えられる．

第5章　結論

本論文では，共同注意が不完全なマルチモーダル環境におけるロボットの学習課題を扱い，そのような環境でもロボットが自律的にカテゴリ化と対応学習が可能な手法の構築に取り組んだ．

1 章では，ロボットのカテゴリ化と対応学習の従来研究を概観し，それら研究における想定の問題を指摘した．従来研究では，ユーザとロボットが共同注意し，ロボットが観測する複数の情報が特定の事物を表すマルチモーダル環境が扱われており，実環境へ応用した際などその想定が崩れる場合にロボットがどのように学習を進めるべきかには焦点は当てられていなかった．そこで，本論文ではそのような想定がおけない，ユーザとロボットの共同注意が不完全なマルチモーダル環境におけるロボットのカテゴリ化と対応学習の実現を目的とした．

2 章では，1 章で述べた目的を達成するための方法として提案した主観的整合性の概念について，その着想となった乳児のカテゴリ化や対応学習に関する発達心理学の知見及び人の認知や脳処理についての社会心理学や認知科学，脳科学の知見を紹介し，主観的整合性の基本アイデアについて述べた．

3 章では，教示者が必ずしも学習者を模倣するとは限らない状況における音声模倣学習の課題に対して主観的整合性のアイデアを導入した対応学習手法を提案した．具体的には，認知発達ロボティクスのアプローチにより[61]，乳児が音声模倣と語彙獲得を共発達させている現象に注目し，そこに内在するメカニズムとして，主観的整合性に基づく音声模倣と語彙獲得との相補的対応学習の手法を提案した．簡易的なロボット実験およびロボットと養育者とのインタラクション場面を想定した計算機シミュレーションを実施し，養育者がどの程度学習者に模倣や提示などの教示的振舞を示すかを表した教示率に対する提案手法の学習の頑健性を確認した．また，実場面の

乳児と養育者のインタラクションで見られる，養育者が乳児に対してほとんど模倣を示さない状況においても，提案手法により，語彙のための対応の学習を利用することで模倣のための対応の学習が可能となることを確認し，音声模倣と語彙獲得の共発達のための相互促進的な対応学習が実現された．

4 章では，ユーザとロボットが必ずしも同じ事物に注目しているとは限らない状況においてロボットの自律的な語意学習を可能とする主観的整合性に基づいたマルチモーダルカテゴリゼーション手法を提案した．実ロボットを用いた実験結果から，提案手法を用いることで対応率に対して頑健に語意を学習できることを確認した．この結果は，提案手法により，人がロボットに語意を教示する際の制限 (常に教示的に振る舞わなければならない等) が緩和可能であることを示唆する．また，人工データを用いた実験結果から，提案手法を用いることで，対応するデータの主観的整合性を，他のモダリティとは対応しないデータに比べて高く計算することで，すなわち，部分的な対応関係を重視することで，より高い頻度で正しいカテゴリのサンプリングが可能になることによるものであると分析された．

最後に，今後の展望として，3 つの研究の方向性を提案する．

1 つ目は，実世界での検証を進める方向である．本論文では，実世界で起こりうる人とロボットの共同注意が不完全な状況におけるロボットの学習実験を実施したが，人がロボットに対応を与える，すわなち共同注意する割合が一定であったり，学習対象である環境中の事物の数が固定であるなど，状況が単純化されていた．実世界における人とロボットのやり取りについては，クラウドソーシングやシミュレータを用いた方法によりいくつか研究されており[100; 101; 102]，それらの研究により実世界の人とロボットのやり取りをモデル化できると期待される．しかし実際には，ロボットの振舞や学習の履歴，対面する人に応じてやり取りは様々であると考えられ，事前に全てのやり取りを想定することは困難である．そのため，今後は，提案手法を実装したロボットを実世界に投入し，実世界での人とのやり取りを通じてどのように対応やカテゴリの学習が可能であるかを検証していくことが必要である．また，ロボットが，今後より多くの生活場面で活躍していくには，人からの様々な要求に応えられる必要があり，それに応じてロボットの学習対象は多様化していくと考えられる．ロボット

がセンシングできるデータが増えるほど，様々な対象を扱えうると期待されるが，学習対象とは関係のないデータも増加すると考えられる．すなわち，共同注意が不完全なマルチモーダル環境の想定を無視できなくなっていくと考えられる．提案手法は，そのような場合でも整合性を指標にしてより関係のあるデータのみに基づいた効率的な学習を可能にすると考えられる．今後は，3，4 章で述べた個別課題に取り組み手法を精緻化していくと共に，実世界での検証を進めていくことで，実世界で活躍するロボットの実現に近づくと期待される．

2 つ目は，本論文で提案した主観的整合性の学習手法としての有効性を検証していく方向である．近年，機械学習分野で Deep Learning と呼ばれる手法が注目されている．Hinton et al. によって提案された Deep Belief Nets[103] が発端となり，Deep Boltzmann Machines[104] や Autoencoder[105] などが提案されている．Deep Learning は，画素などの低次な特徴から人や猫の顔などの高次な特徴を教師無しで学習できる特徴抽出器であり，ロボットの行動学習などにも応用されている[106]．マルチモーダル情報を扱える Multimodal Deep Boltzmann Machines[107] も提案されており，これにより Web などにある大量のラベルと画像データから，風景や動物などの高次なカテゴリの分類や検索 (ラベルと画像のどちらか一方から他方を想起) が可能になることが示されている．しかしながら，この研究では，ラベルと画像が対応したセットが訓練データとして与えられる，すなわち，異なるモダリティ間で対応するデータが学習時に与えられることが想定されている．これに対して，主観的整合性を導入することで，訓練データに対応しないラベルと画像のセットが含まれる場合でも，対応するデータ (特徴)，あるいは対応するモダリティを積極的に利用し，対応しないデータを排除した効率的な学習が実現できると考えられる．今後は，上記機械学習手法と組み合わせた取り組みを実施し，主観的整合性を学習手法として精練させていくことが課題である．

3 つ目は，人の発達や認知，脳処理の仕組みとしての主観的整合性の妥当性を検証していく方向である．2 章で述べたように，現象として人は乳児期から様々なレベルで情報の整合性に敏感であり，整合性が崩れる際にはそれを解消しようとしていることが伺えるが，そこに内在するメカニズムは明らかではない．これに対し，例えば，

本論文で実施した実験のような同時に複数のモダリティにおいて未知の刺激を観測したときのカテゴリ化及び対応学習プロセスに焦点を当て，人を対象とした学習実験を行い，そのスコアから未知刺激（人工データ）に対する人のカテゴリ化及び連関の関連性を記述する．そして，主観的整合性のパラメータ（主観的整合性に対する感度など）を用いて，これをモデル化することで，人の学習戦略としての主観的整合性の妥当性検証を行う．これによって，整合性に基づいた人の発達や認知，脳処理のメカニズムの理解に繋がると期待される．実験パラダイムの精緻化と実験結果のモデル化手法の検討が今後の課題である．

関連図書

[1] http://www.honda.co.jp/ASIMO/.

[2] http://www.sony.jp/products/Consumer/aibo/.

[3] http://jpn.nec.com/robot/.

[4] http://www.softbank.jp/robot/special/pepper/.

[5] 中村友昭, 長井隆行, 岩橋直人. ロボットによる物体のマルチモーダルカテゴリーゼーション. 電子情報通信学会論文誌, Vol. J91-D, No. 10, pp. 2507–2518, 2008.

[6] 中村友昭, 長井隆行, 岩橋直人. Gibbs sampling による物体のマルチモーダルカテゴリゼーション. 第 28 回日本ロボット学会学術講演会, pp. 2I1–3, 2010.

[7] Yuichiro Yoshikawa, Minoru Asada, Koh Hosoda, and Junpei Koga. A constructivest approach to infants' vowel acquisition through mother-infant interaction. *Connection Science*, Vol. 15, No. 4, pp. 245–258, 2003.

[8] Katsushi Miura, Yuichiro Yoshikawa, and Minoru Asada. Unconscious anchoring in maternal imitation that helps finding the correspondence of caregiver's vowel categories. *Advanced Robotics*, Vol. 21, pp. 1583–1600, 2007.

[9] Deb K. Roy and Alex P. Pentland. Learning words from sights and sounds: a computational model. *Cognitive Science*, Vol. 26, No. 1, pp. 113–146, 2002.

[10] Chen Yu. The emergence of links between lexical acquisition and object categorization: a computational study. *Connection Science*, Vol. 17, No. 3-4, pp. 381–397, 2005.

[11] 田口亮, 岩橋直人, 船越孝太郎, 中野幹夫, 能勢隆, 新田恒雄. 統計的モデル選択に基づいた連続音声からの語意学習. 人工知能学会論文誌, Vol. 25, No. 4, pp. 549–559, 2010.

[12] Tomoaki Nakamura, Takaya Araki, Takayuki Nagai, and Naoto Iwahashi. Grounding of word mearnings in latent dirichlet allocation-based multimodal concepts. *Advanced Robotics*, Vol. 25, No. 17, pp. 2189–2206, 2011.

[13] Aris Alissandrakis, Chrystopher L.Nehaniv, and Kerstin Dautenhahn. Correspondence mapping induced state and action metrics for robotic imitation. *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, PART B: CYBERNETICS, SPECIAL*, Vol. 37, No. 2, pp. 299–307, 2007.

[14] Verena V. Hafner and Frederic Kaplan. Interpersonal maps and the body correspondence problem. In *Proceedings of the Third International Symposium on Imitation in animals and artifacts*, pp. 48–53, 2005.

[15] Frank H. Guenther, Michelle Hampson, and Dave Johnson. A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review*, Vol. 105, pp. 611–633, 1998.

[16] Hisashi Kanda, Tetsuya Ogata, Kazunori Komatani, and Hiroshi G. Okuno. Vocal imitation using physical vocal tract model. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1846–1851, 2007.

[17] Jennifer A.Schwade and Michael H.Goldstein. Social feedback to infats' babbling facilitates rapid phonological learning. *Psychological Science*, Vol. 19, pp. 515–523, 2008.

[18] Susan S.Jones. Infants learn to imitate by being imitated. In *Proceedings of IEEE/RSJ International Conference on Development and Learning*, 2006.

[19] Julie Gros-Luis, Meredith J.West, Michael H.Goldstein, and Andrew P.King. Mothers provide differential feedback to infants' prelinguistic sounds. *International Journal of Behavioral Development*, Vol. 30, No. 6, pp. 509–516, 2006.

[20] Catherine S.Tamis-LeMonda, Marc H.Bornstein, and Lisa Baumwell. Maternal responsiveness and children's achievement of language milestones. *Child Development*, Vol. 72, No. 3, pp. 748–767, 2001.

[21] Sandra R.Waxman and Irena Braun. Consistent (but not variable) names as invitations to form object categories: new evidence from 12-month-old infants. *Cognition*, Vol. 95, pp. B59–B68, 2004.

[22] H.Henny Yeung and Janet F.Werker. Learning words' sounds before learning how words sound: 9-month-old infants use distinct objects as cues to categorize speech information. *Cognition*, Vol. 113, No. 2, pp. 234–243, 2009.

[23] Elizabeth Ray and Cecilia Heyes. Imitation in infancy: the wealth of the stimulus. *Developmental Science*, Vol. 14, No. 1, pp. 92–105, 2011.

[24] Paul C.Quinn. Perceptual categorization of cat and dog silhouettes by 3- to 4-month-old infants. *Journal of Ecperimental Child Psycology*, Vol. 79, pp. 78–94, 2001.

[25] G.Cameron Marean, Lynne A.Werner, and Patricia K.Kuhl. Vowel categorization by very young infants. *Developmental Psychology*, Vol. 28, No. 3, pp. 396–405, 1994.

[26] Jessica Maye, Janet F.Werker, and LouAnn Gerken. Infant sensitivity to distributional information can affect phonetic discrimination. *Cogtnition*, Vol. 82, pp. B101–B111, 2002.

[27] Jenny R.Saffran, Richard N.Aslin, and Elissa L.Newport. Statistical learning by 8-month-old infants. *Science*, Vol. 274, No. 5294, pp. 1926–1928, 1996.

[28] Natasha Z.Kirkham, Jonathan A.Slemmer, and Scott P.Johnson. Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition*, Vol. 83, pp. B35–B42, 2002.

[29] Jill Lany and Jenny R.Saffran. *Comprehensive Developmental Neuroscience: Neural Circuit Development and Function in the Brain*, Vol. 3, chapter Statistical Learning Mechanisms in Infancy, pp. 231–248. 2013.

[30] Alissa L.Ferry, Susan J.Hespos, and Sandra R.Waxman. Categorization in 3- and 4-month-old infants: An advantage of words over tones. *Child Development*, Vol. 81, No. 2, pp. 472–479, 2010.

[31] Alison J.Greuel and Janet F.Werker. Audio-visual categorization of feature correlations in four-month-old infants. In *Proceedings of the 17th Biennial International Coference on Infant Studies*, 2010.

[32] 大藪泰. 赤ちゃんの模倣行動の発達: 形態から意図の模倣へ. バイオメカニズム学会誌, Vol. 29, No. 1, pp. 3–8, 2005.

[33] Susan S. Jones. Imitation in infancy the development of mimicry. *Psychological Science*, Vol. 18, pp. 593–599, 2007.

[34] Theano Kokkinaki and Giannis Kugiumutzakis. Basic aspects of vocal imitation in infant-parent interaction during the first 6 months. *Journal of reproductive and infant psychology*, Vol. 18, No. 3, pp. 173–187, 2000.

[35] Martha Pelaez-Nogueras, Jacob L. Gewirtz, and Michael M. Markham. Infant vocalizations are conditioned both by maternal imitation and motherese speech. *Infant Behavior and Development*, Vol. 19, No. 1, p. 670, 1996.

[36] Cecilia Heyes. *Perspectives on imitation: From neuroscience to social science*, chapter Imitation by association, pp. 157–176. MIT Press, 2005.

[37] Andrew N. Meltzoff and M. Keith Moore. Newborn infants imitate adult facial gestures. *Child Development*, Vol. 54, No. 3, pp. 702–709, 1983.

[38] Xin Chen, Tricia Striano, and Hannes Rakoczy. Auditory-oral matching behavior in newborns. *Developmental Science*, Vol. 7, No. 1, pp. 42–47, 2004.

[39] Emese Nagy, Hajnalka Compagne, Hajnalka Orvos, Attila Pal, Peter Molnar, Imre Janszky, Katherine A.Loveland, and Gyorgy Bardos. Index finger movement imitation by human neonates: Motivation, learning and left-hand preference. *Pediatric Research*, Vol. 58, pp. 749–753, 2005.

[40] Eugene Abravanel and Ann D.Sigafoos. Exploring the presence of imitation during early infancy. *Child Development*, Vol. 55, pp. 381–392, 1984.

[41] Roger Fontaine. Imitative skills between birth and six months. *Infant Behavior and Development*, Vol. 7, pp. 323–333, 1984.

[42] Sandra W.Jacobson. Matching behavior in the young infant. *Child Development*, Vol. 50, pp. 425–430, 1979.

[43] Susan S.Jones. Imitation or exploration? young infants' matching of adults' oral gestures. *Child Development*, Vol. 67, pp. 1952–1969, 1996.

[44] Emmanuel Devouche. Imitation across changes in object affordances and social context in 9-month-old infants. *Developmental Science*, Vol. 1, No. 1, pp. 65–70, 1998.

[45] Leon Festinger. *A Theory of cognitive dissonance*. Stanford University Press, 1957.

[46] Fritz Heider. *The psychology of interpersonal relations*. John Wiley and Sons, 1958.

[47] Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, Vol. 264, pp. 746–748, 1976.

[48] Shoko Kanaya and Kazuhiko Yokosawa. Perceptual congruency of audio-visual speech affects ventriloquism with bilateral visual stimuli. *Psychonomic Bulletin and Review*, Vol. 18, pp. 123–128, 2011.

[49] Matthew Botvinick and Jonathan Cohen. Rubber hands 'feel' touch that eyes see. *Nature*, Vol. 391, p. 756, 1998.

[50] Alvin M.Liberman and Ignatius G.Mattingly. The motor theory of speech perception revised. *Cognition*, Vol. 21, No. 1, pp. 1–36, 1985.

[51] 柏野牧夫. 音声知覚の運動理論をめぐって. 日本音響学会誌, Vol. 62-5, pp. 391–396, 2006.

[52] Sophie K. Scott and Ingrid S. Johnsrude. The neuroanatomical and functional organization of speech perception. *Trends in Neurosciences*, Vol. 26, No. 2, pp. 100–107, 2003.

[53] Jane E. Warren, Richard J.S. Wise, and Jason D. Warren. Sounds do-able: auditory-motor transformations and the posterior temporal plane. *Trends in Coginitive Sciences*, Vol. 28, pp. 636–643, 2005.

[54] Stephen M.Wilson, Ayse Pinar Saygin, Martin I.Sereno, and Marco Iacoboni. Listening to speech activates motor areas involved in speech production. *Nature Neuroscience*, Vol. 7, No. 7, pp. 701– 702, 2004.

[55] Friedemann Pulvermuller, Martina Huss, Ferath Kherif, Fermin Moscoso del Prado Martin, Olaf Hauk, and Yury Shtyrov. Motor cortex maps articulatory features of speech sounds. *Proceedings of the National Academy of Sciences*, Vol. 103, No. 20, pp. 7865–7870, 2006.

[56] Toshiaki Imada, Yang Zhang, Marie Cheour, Samu Taulu, Antti Ahonen, and Patricia K. Kuhl. Infant speech perception activates broca's area: a developmental magnetoencephalography study. *Neuroreport*, Vol. 17, pp. 956–962, 2006.

[57] Dorothee Saur, Bjorn W.Kreher, Susanne Schnell, Dorothee Kummerer, Philipp Kellmeyer, Magnus-Sebastian Vry, Roza Umarova, Mariacristina Musso, Volkmar Glauche, Stefanie Abel, Walter Huber, Michel Rijntjes, Jurgen Hennig, and Cornelius Weiller. Ventral and dorsal pathways for language. *Proceedings of the National Academy of Sciences*, Vol. 105, No. 46, pp. 18035–18040, 2008.

[58] Gregory Hickok and David Poeppel. The cortical organization of speech processing. *Nature Reviews*, Vol. 8, pp. 393–402, 2007.

[59] 中市健志, 坂本真一, 力丸裕. 先天性難聴者における劣化雑音音声の知覚. 電子情報通信学会技術研究報告. TL, 思考と言語, Vol. 104, No. 316, pp. 37–42, 2004.

[60] 橘亮輔, 力丸裕. 劣化雑音音声知覚に関連する脳内活動: functional mri による研究. 電子情報通信学会技術研究報告. SP, 音声, Vol. 104, No. 695, pp. 37–42, 2005.

[61] Minoru Asada, Koh Hosoda, Yasuo Kuniyoshi, Hiroshi Ishiguro, Toshio Inui, Yuichiro Yoshikawa, Masaki Ogino, and Chisato Yoshida. Cognitive developmental robotics: a survey. *IEEE Transactions on Autonomous Mental Development*, Vol. 1, No. 1, pp. 12–34, 2009.

[62] 菊池匡晃, 荻野正樹, 浅田稔. 顕著性に基づくロボットの能動的語彙獲得. 日本ロボット学会誌, Vol. 26, No. 3, pp. 261–270, 2008.

[63] 中野吏, 吉川雄一郎, 浅田稔, 石黒浩. 相互排他性原理に基づくマルチモーダル共同注意. 日本ロボット学会誌, Vol. 27, No. 7, pp. 814–822, 2009.

[64] Elizabeth Bates, Philip S. Dale, and Donna Thal. *The Handbook of Child Language*, chapter 4: Individual Differences and their Implications for Theories of Language Development, pp. 96–151. Blackwell Publishing, 1995.

[65] Elise Frank Masur and Doreen L.Eichorst. Infants' spontaneous imitation of novel versus familiar words: Relations to observational and maternal report measures of their lexicons. *Merrill-Palmer Quarterly*, Vol. 48, No. 4, pp. 405–426, 2002.

[66] Janet F.Werker and Suzanne Curtin. Primir: A developmental framework of infant speech processing. *Language Learning and Development*, Vol. 1, No. 2, pp. 197–234, 2005.

[67] Lakshmi J. Gogate, Laura H. Bolzani, and Eugene A. Betancourt. Attention to maternal multimodal naming by 6- to 8-month-old infants and learning of word-object relations. *Infancy*, Vol. 9, No. 3, pp. 259–288, 2006.

[68] David H.Ackley, Geoffrey E. Hinton, and Terrence J.Sejnowski. A learning algorithm for boltzmann machines. *Cognitive Science*, Vol. 9, pp. 147–169, 1985.

[69] Patricia K.Kuhl. Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience*, Vol. 5, No. 11, pp. 831–843, 2004.

[70] Barbara A.Younger and Leslie B.Cohen. Infant perception of correlations among attributes. *Child Development*, Vol. 54, No. 4, pp. 858–867, 1983.

[71] http://baby.goo.ne.jp.

[72] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors. *Semi-supervised Learning*, chapter 1: Introduction to Semi-Supervised Learning, pp. 2–12. MIT Press, 2006.

[73] Kamal Nigam, Andrew McCallum, Sebastian Thrun, and Tom Mitchell. Learning to classify text from labeled and unlabeled documents. In *Proceedings of the 15th National Conference on Artificial Intelligence*, pp. 792–799, 1998.

[74] Rie K.Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, Vol. 6, pp. 1817–1853, 2005.

[75] Daniel M. Wolpert and Mitsuo Kawato. Multiple paired forward and inverse models for motor control. *Neural Networks*, Vol. 11, No. 7-8, pp. 1317–1329, 1998.

[76] Eiji Uchibe and Kenji Doya. Competitive-cooperative-concurrent reinforcement learning with importance sampling. In *Proceedings of International Conference on Simulation of Adaptive Behavior: From Animals and Animates*, pp. 287–296, 2004.

[77] 佐藤久美子, 梶川祥世, 坂本清恵, 松本博文. 日本語母語乳児の文中からの単語切り出しにおけるアクセントと音素配列の役割. 音声研究, Vol. 2007, No. 3, pp. 38–47, 11.

[78] Dare A. Baldwin. Infants' contribution to the achievement of joint reference. *Child Development*, Vol. 62, No. 5, pp. 875–890, 1991.

[79] Andrew N. Meltzoff and M. Keith Moore. Explaining facial imitation: A theoretical model. *Early Development and Parenting*, Vol. 6, pp. 179–192, 1997.

[80] Christine L.Stager and Janet F.Werker. Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, Vol. 388, pp. 381–382, 1997.

[81] Janet F.Werker and Christopher E.Fennell. *Waving a Lexicon*, chapter From listening to sounds to listening to words: Early steps in word learning, pp. 79–109. MIT Press, 2004.

[82] A.L. Gorin, S.E. Levinson, A.N. Gertner, and E. Goldman. Adaptive acquisition of language. *Computer Speech and Language*, Vol. 5, No. 2, pp. 101–132, 1991.

[83] A.L. Gorin, S.E. Levinson, and A Sankar. An experiment in spoken language acquisition. *IEEE Transactions on Speech and Audio Processing*, Vol. 2, No. 1, pp. 224–240, 1994.

[84] 小野広司, 左祥, 伊丹英樹, 尾関基行, 岡夏樹. 最終行動ヒューリスティクスを用いた状況推定による自由発話音声データからの語句意味学習. HAI シンポジウム 2009, pp. 2B–2, 2009.

[85] 中谷仁, 植村竜也, 荒木修, 西垣貴央, 尾関基行, 岡夏樹. ロボットの語意獲得のためのユーザの発話分類. 第 25 回人工知能学会全国大会予稿集, pp. 3B1–OS22b–4, 2011.

[86] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, Vol. 101, No. 1, pp. 5228–5235, 2004.

[87] Andrea Vedaldi and Brian Fulkerson. Vlfeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/`, 2008.

[88] 小野泰弘, 岡部孝弘, 佐藤洋一. 低解像度画像からの視線方向推定: カーネル多重線形モデルによる個人差への対応. 電子情報通信学会論文誌, Vol. J90-D, No. 8, pp. 2212–2222, 2007.

[89] Paul Boersma and David Weenink. Praat: doing phonetics by computer [computer program]. http://www.praat.org/, 2014.

[90] 岩橋直人. ロボットと言語ー言語コミュニケーション能力の機械学習ー. 人工知能学会誌, Vol. 27, No. 6, pp. 563–568, 2012.

[91] 田口亮, 木村優志, 小玉智志, 篠原修二, 入部百合絵, 桂田浩一, 新田恒雄. 幼児の学習バイアスを利用したエージェントによる語意学習の効率化. 人工知能学会論文誌, Vol. 22, No. 4, pp. 444–453, 2007.

[92] Luc Steels. Grounding language through evolutionary languages games. *Language Grounding in Robots*, pp. 1–22, 2012.

[93] 岩橋直人. 人とロボットの言語コミュニケーションにおける間主観性. 人工知能学会誌, Vol. 26, No. 4, pp. 352–359, 2011.

[94] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, Vol. 101, No. 476, pp. 1566–1581, 2006.

[95] 中村友昭, 荒木孝弥, 長井隆行, 岩橋直人. 階層ディリクレ過程に基づくロボットによる物体のマルチモーダルカテゴリゼーション. 計測自動制御学会論文集, Vol. 49, No. 4, pp. 469–478, 2013.

[96] Arindam Banerjee and Sugato Basu. Topic models over text streams: A study of batch and online unsupervised learning. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, 2007.

[97] Takaya Araki, Tomoaki Nakamura, Takayuki Nagai, Kotaro Funakoshi, Mikio Nakano, and Naoto Iwahashi. Online object categorization using multimodal information autonomously acquired by a mobile robot. *Advanced Robotics*, Vol. 26, No. 17, pp. 1995–2020, 2012.

[98] 中村慎也, 岩橋直人, 長井隆行. 実世界における人とロボットの共有信念の推定に基づいた適応的な発話生成. 知能と情報 (日本知能情報ファジィ学会誌), Vol. 21, No. 5, pp. 663–682, 2009.

[99] Komei Sugiura, Naoto Iwahashi, Hisashi Kawai, and Satoshi Nakamura. Situated spoken dialogue with robots using active learning. *Advanced Robotics*, Vol. 25, No. 17, pp. 2207–2232, 2011.

[100] Cynthia Breazeal, Nick DePalma, Jeff Orkin, Sonia Chernova, and Malte Jung. Crowdsourcing human-robot interaction: New methods and system evaluation in a public environment. *Journal of Human-Robot Interaction*, Vol. 2, No. 1, pp. 82–111, 2013.

[101] Severin Lemaignan, Marc Hanheide, Michael Karg, Harmish Khambhaita, Lars Kunze, Florian Lier, Ingo Lutkebohle, and Gregoire Milliez. *Simulation, Modeling, and Programming for Autonomous Robots Lecture Notes in Computer Science*, Vol. 8810, chapter Simulation and HRI Recent Perspectives with the MORSE Simulator, pp. 13–24. 2014.

[102] Norbert Schmitz, Jochen Hirth, and Karsten Berns. A simulation framework for human-robot interaction. In *Proceedings of the Third International Conference on Advances in Computer-Human Interactions*, pp. 79–84, 2010.

[103] Geoffrey F.Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, Vol. 18, pp. 1527–1554, 2006.

[104] Ruslan Salakhutdinov and Geoffrey Hinton. Deep boltzmann machines. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, Vol. 5, pp. 448–455, 2009.

[105] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. *Advances in Neural Information Processing Systems 19*, pp. 153–160, 2007.

[106] Kuniaki Noda, Hiroaki Arie, Yuki Suga, and Tetsuya Ogata. Multimodal integration learning of object manipulation behaviors using deep neural networks. In *Proceedings of IEEE-RSJ International Conference on Intelligent Robots and Systems*, pp. 1728–1733, 2013.

[107] Nitish Srivastava and Ruslan Salakhutdinov. Multimodal learning with deep boltzmann machines. *Journal of Machine Learning Research*, Vol. 15, pp. 1–32, 2012.

研究業績リスト

学術雑誌

1. 笹本勇輝, 吉川雄一郎, 浅田稔. ロボットの語意学習のための主観的整合性に基づくマルチモーダルカテゴリゼーション. 人工知能学会論文誌, Vol.29, No.5, pp.436–448, 2014

2. 笹本勇輝, 吉川雄一郎, 浅田稔. 音声模倣と語彙獲得の共発達のための主観的整合機構に基づく対応学習. 日本ロボット学会誌, Vol.31, No.1, pp.71–82, 2013.

国際会議における発表（査読あり）

1. Yuki Sasamoto, Naoto Nishijima, and Minoru Asada. Towards understanding the origin of infant directed speech: A vocal robot with infant-like articulation. In Proceedings of the Third Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics, CD-ROM, 2013.

2. Yuki Sasamoto, Yuichiro Yoshikawa, and Minoru Asada. Towards simultaneous categorization and mapping among multimodalities based on subjective consistency. In Proceedings of the First Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics, CD-ROM, 2011.

3. Yuki Sasamoto, Yuichiro Yoshikawa, and Minoru Asada. Mutually constrained multimodal mapping for simultaneous development: modeling vocal imitation and lexicon acquisition. Proceedings of the 9th International Conference on Development and Learning, CD-ROM, 2010.

4. Yuki Sasamoto, Yuichiro Yoshikawa, and Minoru Asada. Selective integration based on subjective consistency facilitates simultaneous development of vocal imitation and lexicon acquisition. Proceedings of Ninth International Conference on Epigenetic Robotics, pp.239–240, 2009.

国内学会・シンポジウム等における発表（査読なし）

1. 笹本勇輝, 主観的整合機構に基づいたロボットのマルチモーダル語意学習. 身体性認知科学と実世界応用に関する研究専門委員会 (ECSRA) 第１１回研究会, 2013.

2. 笹本勇輝, 西嶋直人, 浅田 稔. 乳児型調音ロボットの試作 対乳児発話の発生原理解明を目指して . 日本赤ちゃん学会第 13 回学術集会プログラム抄録集, pp.84, 2013.

3. Yuki Sasamoto. A vocal robot with infant-like articulation as a platform for understanding the origin of infant directed speech. 5th Symposium on Cognitive Neuroscience Robotics, 2013.

4. Yuki Sasamoto. Vocal robot with infant-like articulatory system for understanding the origin of infant directed speech. International Symposium on Cognitive NeuroScience Robotics, 2013

5. 笹本 勇輝, 西嶋 直人, 浅田 稔. 対乳児発話の発生原理解明を目指して 乳児様発声ロボットの試作 . 大阪大学・玉川大学 GCOE 合同ワークショップ, 2013

6. 笹本 勇輝, 西嶋 直人, 浅田 稔. 親の対乳児発話を誘発する乳児様発声機構の試作. 身体性認知科学と実世界応用に関する研究専門委員会 (ECSRA) 第１０回研究会, 2012.

7. 笹本勇輝, 吉川雄一郎，浅田稔. 主観的整合機構に基づく音声模倣と語彙の共発達. 身体性認知科学と実世界応用に関する研究専門委員会 (ECSRA) 第 8 回研

究会, 2010.

8. Yuki Sasamoto, Yuichiro Yoshikawa, and Minoru Asada. Selective integration based on subjective consistency facilitates simultaneous development of vocal imitation and lexicon acquisition. Development of the Social Brain Workshop, 2010

9. 笹本勇輝, 吉川雄一郎, 浅田稔. 主観的コンシステンシーに基づく模倣と語彙の共発達. 身体性認知科学と実世界応用に関する研究専門委員会 (ECSRA) 第 7 回研究会, 2009.

10. 笹本 勇輝, 吉川 雄一郎, 浅田 稔. 主観的コンシステンシーに基づく模倣と語彙の共発達. 第 27 回日本ロボット学会学術講演会, CD-ROM, 3S2-01, 2009.

11. 笹本 勇輝, 吉川 雄一郎, 浅田 稔. 語彙獲得を通じた模倣発達メカニズムの構成. 日本赤ちゃん学会第 9 回学術集会プログラム抄録集, pp.71, 2009.

その他（受賞）

1. Best Poster Award. The Third Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics.

2. ロボカップジャパンオープン 2010 大阪，サッカーヒューマノイド アダルトサイズ, ドリブル and キック及びテクニカルチャレンジ 優勝．Team JoiTech，熱田洋史，田中剛， 池田昌弘，志原開，笹本勇輝，荻野正樹，横山智彰.

3. AAAI-08 AI Video Competition, Best Video Award Nominee(as 2 Finalists), Tomoyuki Noda, Shuhei Ikemoto, Daniel Quevedo, Toshihiko Shimizu, Hidenobu Sumioka, Hisashi Ishihara, Yuki Sasamoto, Yuichiro Yoshikawa, Takashi Minato, Hiroshi Ishiguro, Minoru Asada. CB2: Child Robot with Biomimetic Body. AAAI-08 AI Video Competition, Chicago, 2008. 7. 12.

4. 2008 ECSIS Symposium on Learning and Adaptive Behaviors in Robotic Systems(LABRS2008), LABRS BEST Video Award and $1000 prize, Cognitive Developmental Robotics with a Biomimetic Child-robot. Tomoyuki Noda, Shuhei Ikemoto, Daniel Quevedo, Toshihiko Shimizu, Hidenobu Sumioka. Hisashi Ishihara, Yuki Sasamoto, Yuichiro Yoshikawa, Takashi Minato, Hiroshi Ishiguro, Minoru Asada, University of Edinburgh, UK, August 2008.

5. RoboCup JapanOpen 2008 Numazu. RoboCup Soccer Humanoid League, Technical Challenge in Kid size. 2nd PRIZE. Team JEAP: Michael Norbert Mayer, Fuke Sawa, Takanori Nagura, Yuki Sasamoto, Nguyen Mai.