視聴覚事象の対応付けと
それに基づく概念の獲得

西堀　研人

概要

本論文では，物体操作による視聴覚事象の対応付け，選択的注意による視聴覚事象の対応付け，および視聴覚事象の対応付けに基づく概念の獲得について提案する．物体操作による視聴覚事象の対応付けでは，物体固有の情報ではなく，一般的な法則（ゲシュタルトの群化の法則）を用いる．物体を操作しようとする脳から筋への信号の変化と，その時，視聴覚によって観測される音の変化および映像の変化の“ 同時性 ”，物体操作と視聴覚事象の間における繰り返しの“ 類似性 ”を手掛かりとして対応付ける．実験では，ロボットマニピュレータを用いてドラムをばちでたたいたり，ベルを振ったりして視聴覚事象を発生させ，マイクロフォンとカメラを用いて観測し，運動と視聴覚事象の対応付けを行った．聴覚処理では，対象音検出のため運動がなされていない時刻から雑音を推定し，混合音から雑音を除去する．その後，混合音での音オンセット検出において，周波数バンドごとに分け，適応的に検出する．検出した音オンセットのスペクトル同士の類似性から音源ごとの音オンセット時系列を得る．一方，視覚処理では物体の運動範囲を推定し，運動範囲ごとに得られる映像の重心軌跡から物体の運動方向変化の時系列を得る．統合処理により，運動と視聴覚事象時系列を同一のサンプリング周期に統一し，事象間の類似性を比較する．相関の高い事象同士をグループ化することで，運動によって生じた視覚事象と聴覚事象を対応付ける．実環境において実験を行い，結果から本手法の有効性を確認した．

選択的注意による視聴覚事象の対応付けでは，人間が目的に応じて視覚と聴覚を組み合わせ，周囲の環境から必要な情報を選択的に取得し，感覚器からの大量の情報を瞬時に処理していることに示唆を得た．視覚において動きを知覚した場合に，その動きが在る時点に聴覚の注意を集中し，雑音環境において目的音を検出することを聴覚的注意とし処理を行った．実験では，音声やホワイトノイズを雑音として付加した環境において，目的音の検出を視覚情報による時間軸上での窓関数を用いて行った．視聴覚事象の対応付けの成功率は，時間軸上での窓関数を用いないときの 78.6% か

ら用いたときの 95.2% へ上昇した．また聴覚において音が知覚されるが，視覚情報が得られない場合に，音源方向への頭部の回転や，視線方向と明るさの調整により，音に対応する運動を探すことを視覚的注意とし実験を行った．視聴覚事象の対応付けでは，成功率は 93.3% であり，本手法の有効性を示した．

視聴覚事象の対応付けに基づく概念の獲得では，人が視聴覚事象を対応付けし，事例として蓄積したものを体系化し，概念として獲得することを工学的に実現することを目標とした．本論文では，視聴覚事象を対応付けし，事例として蓄積したものを正準相関分析し，教師なし学習を行い，概念を獲得する手法を提案する．対応付けた事例を共通する特徴によって分類し，それらの分布および中心的な事例によって表されるものを概念とする．視聴覚事象についての概念を自律的に獲得するため，対応付けられた音と画像の事例の集合を用い，正準相関分析により両者の統計的関係を学習する．音と画像の特徴ベクトルを正準空間へ写像し，K-means 法による類別を行い，概念を獲得する．実験の結果，獲得した概念の中心的な事例は，カテゴリの特徴を表している．また，未知の視聴覚入力を獲得した概念のいずれかに識別する成功率は 83.2% 以上であった．さらに，未知の視覚または聴覚入力を対応する概念の聴覚または視覚の事例として想起する成功率は 81.5% 以上であり，本手法の有効性を確認した．

Abstract

This thesis proposes a human being understands the objects in the environment by integrating information obtained by the senses of sight, hearing and touch. In this integration, active manipulation of objects plays an important role. I propose a method for finding the correspondence of audio-visual events by manipulating an object. The method uses the general grouping rules in Gestalt psychology, i.e. "simultaneity" and "similarity" among motion command, sound onsets and motion of the object in images. In experiments, I used a microphone, a camera and a robot which has a hand manipulator. The robot grasps an object like a bell and shakes it or grasps an object like a stick and beat a drum in a periodic, or non-periodic motion. Then the object emits periodical/non-periodical events. To create more realistic scenario, I put other event source (a metronome) in the environment. As a result, I had a success rate of 73.8 percent in finding the correspondence between audio-visual events (afferent signal) which are relating to robot motion (efferent signal)

I propose a new method for determining the correspondence between audio and visual events based on the selective attention which is observed in a living organism. For auditory attention, our system recognizes a target sound in noisy environments using an auditory filter that focuses on a specific period corresponding to visual events. It was confirmed that the success rate of correspondence between audio and visual events increased from 78.6% without the auditory filter, to 95.2% with the auditory filter. This certifies the importance of visual information for auditory attention. I also realized a visual attention of localizing a sound source and adjusting the line of sight and the iris of a camera when the system perceived a sound, but could not obtain visual information. The success rate of experiments based on visual attention is 93.3%. These results show the

effectiveness of the proposed method.

In the real world, there are a lot of objects and it is impossible to make a system memorize all knowledge concerning the real world. Therefore, the system should autonomously learn knowledge relating to the environment. I propose the system that autonomously acquires concepts which are derived by statistical relation between audio-visual events. Firstly, the system determines correspondence between audio-visual events after extracting patterns from the external world and accumulates them as cases. Secondly, it applies a canonical correlation analysis to the cases and categorizes them by using K-means method. Finally, it identifies unknown image or sound and associates the corresponding sound or image. As the result of experiments, the identification success rate of concepts is more than 83.2%. And the association success rate of concepts is more than 81.5%. Consequently, the effectiveness of this method was confirmed.

目 次

図 目 次

表 目 次

第1章　序論

1.1　本研究の背景と目的

人間は，周囲の環境で生じる様々な事象を理解する時，受動的に観測する視聴覚情報だけでなく，対象に働きかける運動情報を通じて対象に関する様々な情報を収集し，それらを統合する．視聴覚情報のような異種感覚においては，それぞれから特徴量を抽出し，同じ次元で統一的に統合することが重要となる．E. Spelke [1] は，4ヶ月の幼児が，視聴覚パターンの間の不変な関係を認識でき，モダリティを横断する情報を常に探索することで，物体や事象のマルチモーダルな特性を見つけることを示した．

自分の意志で動かしてみることによって，どのような運動が行われたかだけではなく，どのような運動を行おうとしたかに関する遠心性のコピーと呼ばれる運動情報が得られる．また，幼児の言語の獲得には視聴覚情報だけでなく，対象物への身体運動が大きな役割を果たしているという正高ら [2, 3] の報告もある．このように，対象物に能動的に働きかけ，その運動に関連する視聴覚事象を対応付けることは，物体についての概念獲得に重要であると考える．

また，人間は目的に応じて視覚と聴覚を組合せ，受動的に観測するだけでなく，対象へ注意を向けることで周囲の環境から必要な情報を選択的に取得し，処理を行っている．複数の聴覚事象が同時に混在する環境において，特定の聴覚刺激を選択的に聞き取ることができる [4, 5]．また，視覚刺激に注意を向け，外界から必要な情報を選択するときには，対象を視力の高い視野の中心部でとらえるように，頭部運動と眼球運動を組合せて視線を移動させている [6, 7]．このように，対象へ選択的に注意を向けることで，特定の情報を効率的に取得している．

さらに，人は視聴覚事象を対応付けし，事例として蓄積したものを体系化し，概念として獲得すると考えられる．P. Kuhl ら [8] の報告では，生後 18〜20 週の幼児は発話と口の動きの対応を認識し，言語を獲得する．また，J. Saffran ら [9] の研究により，統計的な情報を用いて幼児が言語の学習を行うことが確認されている．このような事例から概念を獲得する機能を実現するため，複数モダリティの統計的な相関関係を基に，音や画像入力から概念を自律的に獲得するシステムを構築することが望まれる．実

世界の対象は非常に多様であるため，設計者が実世界に関する概念をすべてシステムやロボットに記憶させることは不可能である．そのため，システムやロボットが環境内に存在する物体についての概念を自律的に学習することが必要となる．

本研究の目的は，視聴覚事象の対応付け手法の拡張として，(1) 複数の視聴覚事象が存在する環境中において，物体操作によって生じた視聴覚事象を，操作の運動情報をもとに対応付ける手法，(2) 聴覚または視覚環境が不良である場合や視野内に対象物体がない場合に，対象物へ選択的に注意を向けることで対応付けるといったより実際的な手法，(3) 対応付けられた視聴覚事象から概念を獲得する手法，を考案することである．

1.2　本研究の応用

ロボットや知能システムの設計時に，すべての知識を組み込むことはできない．また，現実世界に柔軟に適応する能力が求められる．そこで，システムが自律的に知識を獲得することが不可欠となる．本研究の応用は，次のとおりである．

- 人間のパートナーロボット
  人間のパートナーとして人と生活空間を共にするロボットが事前知識なしで，能動的に対象物に働きかけたり，注意を向けることで物体の視聴覚事象を対応付け，自律的に対象物に対する概念を獲得する．

- 検査ロボット
  複数の視聴覚事象が存在する工場内において，システムの異常な動きや音の発生を検知する．また，配管や機械部品等をたたき，音によって，亀裂や破損がないか検査する．

- セキュリティ監視システム
  音と映像を組み合わせたシステムにより，音が発生した時刻の音源方向の映像を選択的に取得することで，効率的に広範囲を監視することができる．

- 障害者への視聴覚的保障システム
  視聴覚の一方に障害をもつ人へ，それを保障するように概念を想起する．例えば，目が見えない人の前に，犬や自転車に乗った人が現れたときに，それが犬または自転車であると聴覚的に伝える．

1.3 従来研究

視聴覚事象の対応付け手法について，早川ら [10] は，ゲシュタルト心理学の群化の要因に基づいて，物体についての固有な知識用いずに，物体の運動とそれによって生じた音を対応付けた．J. Chen ら [11, 12] は，それを複数運動，複数音源という状況や音源が移動しながら音を発するような，より複雑な環境でも対応できるような手法へ拡張した．また，K. Nakadai ら [13] は人の上半身を模倣したロボットに装着した 2 つのカメラと 3 つのマイクロフォンを用いて，複数話者の顔と発話を実時間で対応付けた．P. Aarabi [14] は 2 つのカメラと 3 つのマイクロフォンを用いて，音を発する運動物体の空間位置を画像と音において推定し，それらを合成した尤度関数を最大にする物体を求めた．しかし，これらの手法は受動的に視聴覚で観測される事象の対応付けである．

赤松ら [15] は，人が外界の対象を知覚する時，触覚，運動，視覚といった複数の感覚からの情報を統合していると考え，対象物の形状知覚に関して複数の感覚の相互関連性を心理物理的に解析した．その結果，運動情報により，視覚と触覚との対応が付きやすくなることを示した．しかし，この研究はインタフェースでの利用を想定し，呈示された複数の感覚を統合するのは人であり，計算機による自動的な統合はなされていない．橋本 [16] は，視覚情報を求心性信号とし，ロボットマニピュレータへのコマンドのコピーを遠心性のコピーとして，ビジュアルサーボ機構を実現した．

注意を工学的にモデリングする研究は，近年活発に行われている．高橋ら [17, 18] は，選択的注意として，聴覚では視覚情報を用いて教師信号を作成し，聴覚情報を検出するための線形フィルタ学習を行う研究を行った．また，視覚における選択的注意については，L. Itti ら [19, 20] の，画面上での顕著な特徴を有する領域の検出や，陳ら [21] の音源の推定位置や対象物とシステムとの相対的姿勢を保つ行動とを統合する研究が行われている．港ら [22] は, 視覚的注意を工学的にモデリングすることによって，ロボットの頭部や視線の動きを制御し，視覚情報を効率的に素早く抽出した．長井ら [23] は，幼児が養育者との間で共同注意を成立させることで，養育者から対象物に関する多くの知識や言語を学習していることに着目した．そして，ロボットに，人間の幼児にも生得的に備わっていると考えられる注視機能と自己評価学習機能を実装し，それをもとに，試行と学習を繰り返すことで，養育者からのタスク評価なしに共同注視の能力を獲得できるようにした．

大西ら [24] や川野ら [25] は，聴覚情報から音源定位を行い，音源方向にカメラワークを行うことで視覚情報を検出し，会議の自動撮影を実現した．しかし，画像上で人

物の位置を検出することにとどまり，撮影後の視聴覚事象の統合は行われていない．

中川ら [26] は，幼児がどのような情報処理によって概念形成を行っているか考察し，これに示唆を得て工学的に概念形成メカニズムを計算機上に実現した．赤穂ら [27] は，画像や音といった複数の情報源を持つマルチモーダルなシステムが，それらの情報を統合して概念を学習するための枠組みについて研究を行った．また，ロボットに言語学習をさせる研究として，D. Roy [28] は，視聴覚事象間の相互情報量が最大となる視聴覚事象を対応付け，N. Iwahashi [29] は，感覚信号から属性信号を抽出し，シンボルを作り出した．小島ら [30] は，対話を繰り返すことにより入力された連続画像と音信号から段階的に相手と共有可能な概念を獲得した．

1.4　本論文の特徴と概要

本論文では，手に持った物体を操作することで能動的に物体に働きかけ，その特徴を抽出すると共に，他に視聴覚事象が存在する環境中において事前知識なしで，自身の働きかけた物体と物体が発する音の対応付けを実現する．運動物体が何かに運動を阻害された時に，そのエネルギーの一部が音として発生する．これをゲシュタルト心理学の群化の要因である“ 同時性 ”としてマニピュレータの運動の変化，画像上の運動の変化，そして音の変化の時刻が類似するものを群化する手掛かりとする．この際，運動の制御信号である遠心性の信号と視聴覚信号のような観測された求心性信号を対応付けることが従来研究にない特徴である．

次に，選択的注意として，視覚（聴覚）情報が良好に検出できるが，聴覚（視覚）情報が検出しにくい場合において，良好な視覚（聴覚）情報を手掛かりに，検出しにくい聴覚（視覚）情報を探すことを実現する．視覚において動きを知覚した場合に，その動きがある時点に聴覚の注意を集中し，雑音環境において目的音を検出することを聴覚的注意とする．また，聴覚において音が知覚されるが，視覚情報が得られない場合に，音源方向へ頭部を回転させることと，視覚において視線方向や明るさを調整することにより，音に対応する運動を探すことを視覚的注意とする．このように選択的注意によって，完全なデータを手掛かりに不完全なデータ中の情報を検出することが，本研究の特徴である．

最後の視聴覚事象の対応付けに基づく概念の獲得では，対応付けた事例を共通する特徴によって分類し，それらの中心的な事例によって表されるものを概念として獲得する．本研究の特徴は，視聴覚事象についての概念を自律的に獲得するため，対応付けられた画像と音の事例の集合を用い，正準相関分析により画像と音の特徴との相関

関係を学習し，特徴ベクトルに対して教師なし学習を行い，概念を獲得すること，および新たに示された物体の画像（音）に対して，その物体を表す音（画像）を想起できることである．

本論文は，6 章から構成されており，本章につづく各章の概要は次のとおりである．

2 章では，本研究における対応付けの手掛かりとして用いるゲシュタルト心理学における群化の法則について説明し，また基本的な手法として，音源定位，音の特徴として用いる線形予測係数，画像特徴として用いる Hu モーメントについて述べる．また，2 群の変量間の関係を求める正準相関分析，クラスタリング手法の 1 つである K-means 法，およびクラス数未知におけるクラス数の推定法についてまとめる．

3 章では，能動的に対象に働きかけ，運動情報を利用した視聴覚事象の対応付け手法について述べる．視聴覚事象の対応付けでは，物体特有の情報ではなく，一般的な法則（ゲシュタルトの群化の法則）を用いる．物体を操作しようとする脳から筋への信号の変化する時刻と，その時，視聴覚によって観測される音の変化および映像の変化する時刻の“ 同時性 ”および物体操作と視聴覚事象の間における繰り返しの“ 類似性 ”を手掛かりとして対応付ける．実験では，マニピュレータにより物体を振り，マイクロフォンとカメラを用いて観測し，運動（遠心性信号）と視聴覚情報（求心性信号）の対応付けを行った．

4 章では，目的に応じて視覚と聴覚を組み合わせ，周囲の環境から必要な情報を選択的に取得する選択的注意による，視聴覚事象の対応付けについて述べる．視覚において動きを知覚した場合に，その動きがある時点に聴覚の注意を集中し，雑音環境において目的音を検出することを聴覚的注意とした．また聴覚において音が知覚されるが，視覚情報が得られない場合に，音源方向への頭部の回転と，視線方向と明るさの調整により，音に対応する運動を探すことを視覚的注意とした．実験では，2 個のマイクロフォン，1 台のカメラを備えた左右方向に回転可能なロボットヘッドを用いる．視覚的注意の実験では，音源定位により，音が発生した方向に注意を向け，周辺部が明るく中央部が暗い穴の中に存在する対象物を検出し，視聴覚事象を対応付ける．聴覚的注意の実験では，視覚情報により得られる物体の運動方向変化の時刻を基に作成した時間軸上での窓関数を用い，雑音と混合された対象物体の音オンセットを検出し，得られた視聴覚信号の対応付けを行う．

5 章では，視聴覚事象を対応付けし，それらを事例として蓄積したものを体系化することにより，概念として獲得する手法について述べる．ここでは，対応付けた事例を共通する特徴によって分類し，それらの中心的な事例によって表されるものを概念

とする．視聴覚事象についての概念を自律的に獲得するため，対応付けられた音と画像の事例の集合を用い，正準相関分析により両者の統計的関係を学習する．音と画像の特徴ベクトルを正準空間へ写像した後，K-means 法による類別を行い，得られた各分類の中心事例を概念として獲得する．実験では，対象物体の運動によって発生した画像と音を群化の法則を基に対応付け，それぞれの特徴量を算出する．そして，正準相関分析により作成した正準空間においてクラスタリングを行う．最後に，カテゴリ未知の入力事例から対応する概念の同じモダリティ（異なるモダリティ）における中心的な事例を提示し，識別（想起）を行う．実験では，対象物の視聴覚事象を計測するため，1 箇所に固定したマイクロフォンと異なる視点で観測するために位置をかえたカメラを用い，外界から得た視聴覚データの統計的関係から視聴覚事象の概念を獲得した．

6 章では，本論文をまとめるとともに今後の課題について述べる．

第2章　本論文で用いた基本的な考え方と手法

計算機によって外界の視聴覚事象を情報として取り入れ，その特徴を用いるために，従来から様々な手法が用いられてきた [31–33]．本章では，本論文に用いる基本的な考え方や手法についてまとめる．

2.1　ゲシュタルト心理学における群化の法則

物の形は，人間が物を認識する上で，重要な手掛かりとなる．このとき，人間が知覚するものは，個々の刺激要素ではなく，要素に還元できない全体性をもつ形態（ゲシュタルト，Gestalt）である．図形や形は，それを構成する点や線などの要素がまとまったもので，そのような知覚的なまとまりを形成することを，知覚的体制化という [34]．

M. Wertheimer は，ゲシュタルトが知覚される際に働く，**表 2.1** に示すような心理法則（ゲシュタルトの法則；群化の要因）を挙げた [35, 36]．1) 近接の要因：空間的，時間的に近いものがまとまりとして知覚されやすい．2) 類同の要因：色や形などの類似性が高いものどうしがまとまりやすい．3) 閉合の要因：閉じた領域を形成するものが知覚されやすい．4) よい連続の要因：なめらかに特性が変化するものがまとまりとして知覚されやすい．5) よい形の要因：より規則的な形をしたものがまとまる．6) 共通運命の法則：同期して運動や変化するものはまとまって知覚されやすい．

これら以外にもいくつかの要因が挙げられているが，これらの法則は，形態が知覚されるときには，できるだけ全体がもっとも簡潔なよい形になる傾向があるとするプレグナンツ (Prägnanz) の法則に包括できるとされる．本論文では，視聴覚事象の対応付けにおいて，このゲシュタルトの群化の要因を対応付けの手掛かりとして用いる．

2.2　音源定位

空間的に配置された複数のマイクロフォンで音響信号を受音すると，各受音信号の間には時間差や振幅差が生じる．音源定位はこれらの信号の差を利用して行う [37]．

表 2.1: ゲシュタルトの法則
Table 2.1 Gestalt laws

1	Law of proximity	Spatial or temporal proximity of elements may induce the mind to perceive a collective or totality
2	Law of similarity	When more than one kind of element is present, those which are similar tend to form groups
3	Law of closure	Other things being equal, lines which enclose a surface tend to be seen as a unit
4	Law of continuity	Parts of a figure which have a good contour, or common destiny, tend to form units
5	Law of symmetry	Symmetrical images are perceived collectively, even in spite of distance
6	Law of common fate	Elements are grouped when they move simultaneously and in a similar manner

図 2.1 は，2 つのマイクロフォン（L と R の位置）による音源 S の定位を示す．音源位置は，左側のマイクロフォンの計測音から，後述する **3.3.1**(a) の処理で求めた音オンセット時刻における，左右マイクロフォンの計測音間の相互相関関数が最大値となる時間差より推定する [12]．2 つの入力信号 $l(n)$ と $r(n)$ 間の相互相関関数 cc_{rl} は次式で求める．

$$cc_{rl}(d) = \sum_{n=N_1}^{N_2} l(n)r(n-d) \tag{2.1}$$

サンプルは音オンセットを中心にとった 1024 点のデータとする ($N_1 = -511, N_2 = 512$)．$cc_{rl}(d)$ が大きいことは相関があることを示し，$cc_{rl}(d)$ の値が最大となる時間差 d は，両耳間の音の到達時間差 ITD(inter-aural time difference) となる．音源から 2 つのマイクロフォンまでの距離 D_L，D_R が，マイクロフォン間距離 D_M よりも充分大

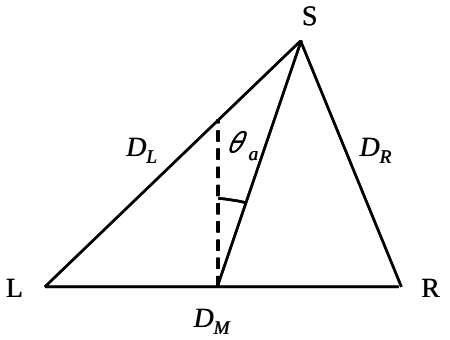


図 2.1: 2 つのマイクロフォンによる音源定位
Fig. 2.1 Sound localization by two microphones.

きいとき，1 つの音オンセットで得られる音源の方位角 θ_a は次式で近似できる．

$$\begin{aligned} \theta_a &\simeq \sin^{-1}\left\{\frac{D_L - D_R}{D_M}\right\} \\ &= \sin^{-1}\left\{\frac{V_{sound} \times ITD}{D_M}\right\} \end{aligned} \tag{2.2}$$

ここで，V_{sound} は音速を表す．

2.3 線形予測係数

人間は，コミュニケーション手段として音声を用いる．音声の生成メカニズムは，声帯振動や声道を呼気が通過するときに発生する気息雑音という音源を入力信号，声道をフィルタとする調音フィルタとしてモデル化することができる．本論文では，そのような調音フィルタを実現する一手法である，線形予測分析によって得られる線形予測係数 (LPC : linear prediction coefficient) を音特徴として用いる [38]．線形予測分析は，次式の自己回帰過程により，音の生成モデルを定義している．

$$s(n) + a_1 s(n-1) + \cdots + a_M s(n-M) = u(n) \tag{2.3}$$

ここで，$s(n)$ は音信号，$u(n)$ は音源信号，$a_i (i = 1, ..., M)$ は線形予測係数，M は線形予測分析における分析次数であり，式 (2.3) は次のように変形できる．

$$s(n) = -\sum_{k=1}^{M} a_k s(n-k) + u(n) \tag{2.4}$$

音信号 $s(n)$ と音源信号 $u(n)$ の z-変換をそれぞれ $S(z)$，$U(z)$ とすると式 (2.4) は z-変換により，次式のように変形できる．

$$(1 + \sum_{k=1}^{M} a_k z^{-k}) S(z) = U(z) \tag{2.5}$$

調音フィルタの伝達関数は $H(z) = S(z)/U(z)$ であるので調音フィルタ $H(z)$ は，次式のような全極形の伝達関数となる．

$$H(z) = \frac{1}{1 + \sum_{k=1}^{M} a_k z^{-k}} \tag{2.6}$$

これは，調音フィルタの共振のみに着目したモデルであり，人間の聴覚特性がスペクトルのピークに敏感であることから，妥当な仮定として受け入れられている．入力 $u(n)$ が未知の場合，信号 $s(n)$ は式 (2.4) において，$u(n) = 0$ として次式で近似できる．

$$\hat{s}(n) = -\sum_{k=1}^{M} a_k s(n-k) \tag{2.7}$$

予測係数は，次式の予測誤差の 2 乗和を最小化で求める．

$$J \equiv \sum_n e^2(n) = \sum_n \big(s(n) + \sum_{k=1}^{M} a_k s(n-k)\big)^2 \tag{2.8}$$

すなわち，次式を満足する係数 a_i を求める．

$$\frac{\partial J}{\partial a_i} = 0, \quad 1 \le i \le M \tag{2.9}$$

式 (2.8)，式 (2.9) より，

$$\begin{aligned} \frac{\partial J}{\partial a_i} &= 2\sum_n \Big(\big(s(n) + \sum_{k=1}^{M} a_k s(n-k)\big) s(n-i) \Big) \\ &= 2\sum_n s(n)s(n-i) + 2\sum_{k=1}^{M} a_k \big(\sum_n s(n-k)s(n-i) \big) \\ &= 0 \end{aligned} \tag{2.10}$$

であるから，正規方程式

$$\sum_{k=1}^{M} a_k \big(\sum_n s(n-k)s(n-i) \big) = -\sum_n s(n)s(n-i), \ \ 1 \le i \le M \tag{2.11}$$

が得られる．自己相関関数 $r(i) = \sum_n s(n-i)s(n)$ を用いると上式は 2 乗誤差の最小値 J_{min} は，式 (2.8) に式 (2.11) を代入することにより，次式となる．

$$J_{min} = \sum_n s(n)s(n-i) + \sum_{k=1}^{M} a_k \big(\sum_n s(n)s(n-k) \big) \tag{2.12}$$

$$\sum_{k=1}^{M} a_k r(k-i) = -r(i), \ \ 1 \le i \le M \tag{2.13}$$

$$J_{min} = r(0) + \sum_{k=1}^{M} a_k r(k) \tag{2.14}$$

式 (2.13) より，正規方程式は，

$$\boldsymbol{T}\boldsymbol{a} = \boldsymbol{c} \tag{2.15}$$

と表される．ここで，$\boldsymbol{a}^t = (a_1, a_2, \cdots, a_M)$，$\boldsymbol{c}^t = (-r(1), -r(2), \cdots, -r(M))$ であり，$\boldsymbol{T}$ は Toeplitz 行列である．

$$\boldsymbol{T} = \begin{bmatrix} r(0) & r(1) & \cdots & r(M-1) \\ r(1) & r(0) & \cdots & r(M-2) \\ \vdots & \vdots & \ddots & \vdots \\ r(M-1) & r(M-2) & \cdots & r(M-1) \end{bmatrix} \tag{2.16}$$

Toeplitz 正規方程式を解く効率的な手法として，Levison-Durbin のアルゴリズムがある．$s(n-i), s(n-i+1), \cdots, s(n-1)$ によって，$s(n)$ を予測した際の誤差を $e_i(n)$ とする．

$$e_i(n) \equiv s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{i} a_k^{(i)} s(n-k) = \sum_{k=0}^{i} a_k^{(i)} s(n-k) \tag{2.17}$$

ここで，$a_0^{(i)} \equiv 1$ とする．予測誤差の 2 乗和を最小化する予測係数 $a_k^{(i)}$ は，正規方程式

$$\begin{bmatrix} r(0) & r(1) & \cdots & r(i-1) \\ r(1) & r(0) & \cdots & r(i-2) \\ \vdots & \vdots & \ddots & \vdots \\ r(i-1) & r(i-2) & \cdots & r(0) \end{bmatrix} \begin{bmatrix} a_1^{(i)} \\ a_2^{(i)} \\ \vdots \\ a_i^{(i)} \end{bmatrix} = - \begin{bmatrix} r(1) \\ r(2) \\ \vdots \\ r(i) \end{bmatrix} \tag{2.18}$$

の解として求められる．いま，最小化された予測誤差の 2 乗和を E_i とすると，式 (2.14) より，

$$E_i(n) = r(0) + \sum_{k=1}^{i} a_k^{(i)} r(k) = \sum_{k=0}^{i} a_k^{(i)} r(k) \tag{2.19}$$

が成り立つ．また，変数 w_i を次式で定義する．

$$w_i \equiv \sum_{k=0}^{i} a_k^{(i)} r(i+1-k) = r(i+1) + \sum_{k=1}^{i} a_k^{(i)} r(i+1-k) \tag{2.20}$$

式 (2.18)〜(2.20) より，次式が成り立つ．

$$\begin{bmatrix} r(0) & r(1) & \cdots & r(i) & r(i+1) \\ r(1) & r(0) & \cdots & r(i-1) & r(i) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ r(i) & r(i-1) & \cdots & r(0) & r(1) \\ r(i+1) & r(i) & \cdots & r(1) & r(0) \end{bmatrix} \begin{bmatrix} 1 \\ a_1^{(i)} \\ \vdots \\ a_i^{(i)} \\ 0 \end{bmatrix} = - \begin{bmatrix} E_i \\ 0 \\ \vdots \\ 0 \\ w_i \end{bmatrix} \tag{2.21}$$

係数行列式の対称性より，

$$\begin{bmatrix} r(0) & r(1) & \cdots & r(i) & r(i+1) \\ r(1) & r(0) & \cdots & r(i-1) & r(i) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ r(i) & r(i-1) & \cdots & r(0) & r(1) \\ r(i+1) & r(i) & \cdots & r(1) & r(0) \end{bmatrix} \begin{bmatrix} 0 \\ a_i^{(i)} \\ \vdots \\ a_1^{(i)} \\ 1 \end{bmatrix} = - \begin{bmatrix} w_i \\ 0 \\ \vdots \\ 0 \\ E_i \end{bmatrix} \tag{2.22}$$

も成り立つ．ここで，$k_{i+1} \equiv -w_i/E_i$ とおき，式 (2.21)+(2.22)$\times$ k_{i+1} を計算すると，

$$\begin{bmatrix} r(0) & r(1) & \cdots & r(i) & r(i+1) \\ r(1) & r(0) & \cdots & r(i-1) & r(i) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ r(i) & r(i-1) & \cdots & r(0) & r(1) \\ r(i+1) & r(i) & \cdots & r(1) & r(0) \end{bmatrix} \begin{bmatrix} 1 \\ a_1^{(i)} + k_{i+1} a_i^{(i)} \\ \vdots \\ a_1^{(i)} + k_{i+1} a_1^{(i)} \\ k_{i+1} \end{bmatrix} = - \begin{bmatrix} E_i + k_{i+1} w_i \\ 0 \\ \vdots \\ 0 \\ w_i + k_{i+1} E_i \end{bmatrix} \tag{2.23}$$

が得られる．式 (2.21) より，

$$\begin{bmatrix} r(0) & r(1) & \cdots & r(i-1) \\ r(1) & r(0) & \cdots & r(i-2) \\ \vdots & \vdots & \ddots & \vdots \\ r(i-1) & r(i-2) & \cdots & r(0) \end{bmatrix} \begin{bmatrix} 1 \\ a_1^{(i)} \\ \vdots \\ a_i^{(i)} \end{bmatrix} = - \begin{bmatrix} E_i \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{2.24}$$

である．一方，式 (2.23) は，k_{i+1} の定義より

$$\begin{bmatrix} r(0) & r(1) & \cdots & r(i) & r(i+1) \\ r(1) & r(0) & \cdots & r(i-1) & r(i) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ r(i) & r(i-1) & \cdots & r(0) & r(1) \\ r(i+1) & r(i) & \cdots & r(1) & r(0) \end{bmatrix} \begin{bmatrix} 1 \\ a_1^{(i)} + k_{i+1} a_i^{(i)} \\ \vdots \\ a_1^{(i)} + k_{i+1} a_1^{(i)} \\ k_{i+1} \end{bmatrix} = - \begin{bmatrix} E_i + k_{i+1} w_i \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} \tag{2.25}$$

となるから，式 (2.25) において，

$$\begin{aligned} a_j^{(i+1)} &\equiv a_j^{(i)} + k_{i+1} a_{i-j+1}^{(i)} \qquad (j = 1, \cdots, i) \\ a_{i+1}^{(i+1)} &\equiv k_{i+1} E_{i+1} \equiv E_i + k_{i+1} w_i = E_i - (k_{i+1})^2 E_i = (1 - k_{i+1}^2) E_i \end{aligned}$$

とおくと，式 (2.25) は式 (2.24) の i を $i+1$ に拡張した式となる．

したがって，正規方程式 (2.16) は，$i = 1, 2, \cdots, M$ の i を $i+1$ に対して以下の式を求めることにより，巡回的に解くことができる．

$$E_0 = r(0) \tag{2.26}$$

$$k_i = -\frac{r(i) + \sum_{j=1}^{i-1} a_j^{(i-1)} r(i-j)}{E_{i-1}} \tag{2.27}$$

$$a_i^{(i)} = k_i \tag{2.28}$$

$$a_j^{(i)} = a_j^{(i-1)} + k_i a_{i-j}^{(i-1)}, 1 \le j \le i-1 \tag{2.29}$$

$$E_i = (1 - k_i^2) E_{i-1} \tag{2.30}$$

ここで，E_i は i 次の予測器における最小 2 乗誤差和 $J_{min}^{(i)}$ である (式 (2.19))．方程式 (2.26)〜(2.30) の最終的な解は

$$a_j = a_j^{(M)}, \qquad 1 \le j \le M \tag{2.31}$$

で与えられる．Levison は，予測器の次数の増加に応じて，E_i が非増加であることを示した．E_i は 2 乗誤差であるため非負である．したがって，次式が成り立つ．

$$0 \le E_i \le E_{i-1}, \qquad E_0 = r(0) \tag{2.32}$$

媒介量 k_i は，反射係数あるいは偏相関係数 (PARCOR:partial correlation coefficient) である．

2.4 Hu モーメント

Hu モーメントは，平行移動，拡大縮小，回転に対して不変なモーメントである [39]．画像に対して，$p+q$ 次のモーメント m_{pq} が次式で近似される．

$$m_{pq} = \sum_x \sum_y x^p y^q f(x,y) \tag{2.33}$$

ここで，x と y は任意の原点に関連するピクセル座標，$f(x,y)$ はピクセルの特徴量を表す．平行移動，拡大縮小，回転に対して不変なモーメントを求めるために，中心モーメント μ が次のように計算される．

$$\mu_{pq} = \sum_x \sum_y (x-\bar{x})^p (y-\bar{y})^q f(x,y) \tag{2.34}$$

ここで，$\bar{x} = m_{10}/m_{00}$，$\bar{y} = m_{01}/m_{00}$ である．

正規化された中心モーメント η は次式で計算される．

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\lambda}} \tag{2.35}$$

ここで，$\lambda = (p+q)/2+1$，$p+q \geq 2$ である．正規化された中心モーメント η_{pq} から，Hu モーメント集合 ϕ が以下のように計算される．

$$\begin{aligned}
\phi_1 &= \eta_{20} + \eta_{02} \\
\phi_2 &= (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \\
\phi_3 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\
\phi_4 &= (\eta_{30} - \eta_{12})^2 + (\eta_{21} - \eta_{03})^2 \\
\phi_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})((\eta_{30} + \eta_{12})^2 - 3(\eta_{21} - \eta_{03})^2) \\
&+ (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})(3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2) \\
\phi_6 &= (\eta_{20} - \eta_{02})((\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2) + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\
\phi_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})((\eta_{30} + \eta_{12})^2 - 3(\eta_{21} - \eta_{03})^2) \\
&+ (3\eta_{12} - \eta_{03})(\eta_{21} + \eta_{03})(3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2)
\end{aligned} \tag{2.36}$$

2.5 正準相関分析

正準相関分析は，p 個の変量を含む変数群 $\boldsymbol{x} = (x^{(1)}, \cdots, x^{(p)})^t$ と q 個の変量を含む変数群 $\boldsymbol{y} = (y^{(1)}, \cdots, y^{(q)})^t$ があるとき，$\boldsymbol{x}$ と $\boldsymbol{y}$ の相関が最も高くなるように写像する

手法である [40–42]．本研究では，α 番目のサンプルの i 次元目の聴覚特徴量を $x_\alpha^{(i)}$，j 次元目の視覚特徴量を $y_\alpha^{(j)}$ とする ($\alpha = 1, \cdots, n;\ i = 1, \cdots, p\ ; j = 1, \cdots, q$)．聴覚特徴量 $x_\alpha^{(i)}$ の変数群を $\boldsymbol{x}_\alpha = (x_\alpha^{(1)}, \cdots, x_\alpha^{(i)}, \cdots, x_\alpha^{(p)})^t$ とし，同様に視覚特徴量 $y_\alpha^{(j)}$ の変数群を $\boldsymbol{y}_\alpha$ とする．また，聴覚の i 次元目の特徴量 $\boldsymbol{x}^{(i)} = (x_1^{(i)}, \cdots, x_\alpha^{(i)}, \cdots, x_n^{(i)})$ と視覚の j 次元目の特徴量 $\boldsymbol{y}^{(j)} = (y_1^{(j)}, \cdots, y_\alpha^{(j)}, \cdots, y_n^{(j)})$ をあらかじめ平均 0，標準偏差 1 に標準化しておく．サンプル数 n の学習データ $\{(\boldsymbol{x}_\alpha, \boldsymbol{y}_\alpha)\}_{\alpha=1}^n$ に対して，相関行列 R_X，R_Y と相互相関行列 R_{XY} は次式で計算される．

$$R_X = \sum_{\alpha=1}^{n} \boldsymbol{x}_\alpha \boldsymbol{x}_\alpha^t / n \tag{2.37}$$

$$R_Y = \sum_{\alpha=1}^{n} \boldsymbol{y}_\alpha \boldsymbol{y}_\alpha^t / n \tag{2.38}$$

$$R_{XY} = \sum_{\alpha=1}^{n} \boldsymbol{x}_\alpha \boldsymbol{y}_\alpha^t / n = R_{YX}^t \tag{2.39}$$

視聴覚特徴量 $\boldsymbol{x}_\alpha$，$\boldsymbol{y}_\alpha$ は係数ベクトル $\boldsymbol{z}^{(k)}$，$\boldsymbol{w}^{(k)}$ により線形変換され，次式のように新変量 $a_\alpha^{(k)}$，$v_\alpha^{(k)}$ に写される．

$$a_\alpha^{(k)} = {\boldsymbol{z}^{(k)}}^t \boldsymbol{x}_\alpha,\ \ v_\alpha^{(k)} = {\boldsymbol{w}^{(k)}}^t \boldsymbol{y}_\alpha \tag{2.40}$$

新変量 $\{(a_\alpha^{(k)}, v_\alpha^{(k)})\}_{\alpha=1}^n$ が，以下の 3 条件を満たすとき，これらを正準変量といい，$a_\alpha^{(k)}$ と $v_\alpha^{(k)}$ の相関係数を第 k 正準相関と呼ぶ．

第 1 条件は，k 次元目の正準変量 $\boldsymbol{a}^{(k)} = (a_1^{(k)}, \cdots, a_\alpha^{(k)}, \cdots, a_n^{(k)})$ と $\boldsymbol{v}^{(k)} = (v_1^{(k)}, \cdots, v_\alpha^{(k)}, \cdots, v_n^{(k)})$ の平均が 0，分散は 1 であることである．すなわち，次式が成立することが条件である．

$$\sum_{\alpha=1}^{n} a_\alpha^{(k)} / n = 0,\ \ \sum_{\alpha=1}^{n} v_\alpha^{(k)} / n = 0 \tag{2.41}$$

$$\begin{aligned} \sum_{\alpha=1}^{n} {a_\alpha^{(k)}}^2 / n &= \sum_{\alpha=1}^{n} {\boldsymbol{z}^{(k)}}^t \boldsymbol{x}_\alpha \boldsymbol{x}_\alpha^t \boldsymbol{z}^{(k)} / n \\ &= \boldsymbol{z}_k^t R_X \boldsymbol{z}_k = 1 \end{aligned} \tag{2.42}$$

$$\sum_{\alpha=1}^{n} {v_\alpha^{(k)}}^2 / n = {\boldsymbol{w}^{(k)}}^t R_Y \boldsymbol{w}^{(k)} = 1 \tag{2.43}$$

第 2 条件は，$k \neq l$ であるとき，$\boldsymbol{a}^{(k)}$ と $\boldsymbol{a}^{(l)}$，$\boldsymbol{v}^{(k)}$ と $\boldsymbol{v}^{(l)}$，$\boldsymbol{a}^{(k)}$ と $\boldsymbol{v}^{(l)}$ は無相関である．

$$\left.\begin{array}{l} \sum_{\alpha=1}^{n} a_\alpha^{(k)} a_\alpha^{(l)} = 0 \\ \sum_{\alpha=1}^{n} v_\alpha^{(k)} v_\alpha^{(l)} = 0 \\ \sum_{\alpha=1}^{n} a_\alpha^{(k)} v_\alpha^{(l)} = 0 \end{array}\right\} \ (k \neq l; k, l = 1, 2, \cdots, s) \tag{2.44}$$

第 3 条件は，$\boldsymbol{a}^{(k)}$ と $\boldsymbol{v}^{(k)}$ の相関係数を r_k とし，大きさの順に並べられていることである．

$$\begin{array}{l}\sum_{\alpha=1}^{n} a_\alpha^{(k)} v_\alpha^{(k)}/n = r_k \quad (k = 1, 2, \cdots, s) \\ r_1 \geq r_2 \geq \cdots \geq r_s > 0\end{array} \tag{2.45}$$

次に，$k = 1$ である第 1 正準変量 $\boldsymbol{a}^{(1)}$ と $\boldsymbol{v}^{(1)}$ の場合において，相関係数 r_1 が最大になるように $\boldsymbol{z}(= \boldsymbol{z}^{(1)})$，$\boldsymbol{w}(= \boldsymbol{w}^{(1)})$ を求める．

$$\begin{array}{rl} r_1 &= \sum_{\alpha=1}^{n} a_\alpha^{(1)} v_\alpha^{(1)}/n \\ &= \sum_{\alpha=1}^{n} \boldsymbol{z}^t \boldsymbol{x}_\alpha \boldsymbol{w}^t \boldsymbol{y}_\alpha/n = \boldsymbol{z}^t R_{XY} \boldsymbol{w}\end{array} \tag{2.46}$$

そこで，式 (2.42)，(2.43) の条件の下で Lagrange の未定乗数 μ_1，μ_2 を用い，次式の関数を最大にする．

$$\begin{array}{rl} f(\boldsymbol{z}, \boldsymbol{w}, \mu_1, \mu_2) &= \boldsymbol{z}^t R_{XY} \boldsymbol{w} \\ & -\frac{\mu_1}{2}(\boldsymbol{z}^t R_X \boldsymbol{z} - 1) - \frac{\mu_2}{2}(\boldsymbol{w}^t R_Y \boldsymbol{w} - 1)\end{array} \tag{2.47}$$

上式を $\boldsymbol{z}$，$\boldsymbol{w}$ で偏微分してゼロとおき，次式を得る．

$$R_{XY} \boldsymbol{w} = \mu_1 R_X \boldsymbol{z},\ R_{YX} \boldsymbol{z} = \mu_2 R_Y \boldsymbol{w} \tag{2.48}$$

式 (2.42)，(2.43) の条件および式 (2.46) から，次式とする．

$$\mu_1 = \mu_2 = \lambda = \boldsymbol{z}^t R_{XY} \boldsymbol{w} = r_1 \tag{2.49}$$

式 (2.48) に R_X^{-1} および R_Y^{-1} を左からかけると，次式が得られる．

$$R_X^{-1} R_{XY} \boldsymbol{w} = \lambda \boldsymbol{z},\ R_Y^{-1} R_{YX} \boldsymbol{z} = \lambda \boldsymbol{w} \tag{2.50}$$

式 (2.50) から一般化固有値問題

$$R_{XY} R_Y^{-1} R_{YX} \boldsymbol{z} = \lambda^2 R_X \boldsymbol{z} \tag{2.51}$$

を得る．この固有値問題を解いて s 個の固有値 $\lambda_k^2 (r_k = \lambda_k\ (k = 1, 2, \cdots, s))$ と，固有ベクトル $\{\boldsymbol{z}^{(k)}, \boldsymbol{w}^{(k)}\}\ (k = 1, 2, \cdots, s)$ が得られ，これらを用いて計算される $\boldsymbol{a}^{(k)}$，$\boldsymbol{v}^{(k)}$ は正準変量となる．すなわち，次式を満足する．

$$\begin{array}{l} R_X^{-1} R_{XY} \boldsymbol{w}^{(k)} = r_k \boldsymbol{z}^{(k)}, \\ R_Y^{-1} R_{YX} \boldsymbol{z}^{(k)} = r_k \boldsymbol{w}^{(k)}\end{array} \tag{2.52}$$

固有値の大きいものから順に s 個の係数ベクトル $\boldsymbol{z}^{(k)}$，$\boldsymbol{w}^{(k)}$ を取り，係数行列 $Z = (\boldsymbol{z}^{(1)}, \cdots, \boldsymbol{z}^{(k)}, \cdots, \boldsymbol{z}^{(s)})^t$，$W = (\boldsymbol{w}^{(1)}, \cdots, \boldsymbol{w}^{(k)}, \cdots, \boldsymbol{w}^{(s)})^t$ とする．視聴覚特徴量 $\boldsymbol{x}_\alpha$，$\boldsymbol{y}_\alpha$ は係数行列 Z，W により線形変換され，次式のように新変量 $\boldsymbol{a}_\alpha$，$\boldsymbol{v}_\alpha$ に写される．

$$\boldsymbol{a}_\alpha = Z \boldsymbol{x}_\alpha,\ \boldsymbol{v}_\alpha = W \boldsymbol{y}_\alpha \tag{2.53}$$

ここで，新変量 $\boldsymbol{a}_\alpha = (a_\alpha^{(1)}, \cdots, a_\alpha^{(k)}, \cdots, a_\alpha^{(s)})^t$，$\boldsymbol{v}_\alpha = (v_\alpha^{(1)}, \cdots, v_\alpha^{(k)}, \cdots, v_\alpha^{(s)})^t$ とする．

2.6　K-means 法

K-means 法は，クラス数 K を与えてクラスタリングを行う処理である [43]．処理の手順は次のとおりである．

Step 1 入力順に K 個のパターン P_1〜P_k を選び，それらの位置を中心とする K 個のクラス C_1〜C_k を作る．変数 $i := 1$ とする．

Step 2 パターン P_i とクラス C_i〜C_k の中心との距離を調べ，パターン P_i を最小の距離にあるクラス $C_j (1 \leq j \leq K)$ に属させる．

Step 3 クラス C_j の中心を，現時点でクラス C_j に属するすべてのパターンの平均をとることによって更新する．

Step 4 $i = N$ なら **Step5** へ，それ以外のときは $i := i + 1$ として **Step2** に戻る．

Step 5 K 個のクラス中心位置が，**Step2**〜**Step4** までの処理によって更新された場合，現在のクラス中心を初期のクラス中心とみなし，$i := 1$ として **Step2** に戻る．一方，全く更新されなかった場合は処理を終了する．

2.7　初期クラスとクラス数の決定

K-means 法は，初期クラス中心位置と点データの入力順序にある程度依存する．そのため，パターンの入力順序が異なるデータ列を用いてクラスタリングを複数回行い，最も分離精度の良い結果を採用する．クラス j の中心（平均ベクトル）を $\boldsymbol{c}_j$，クラス j のサンプル数を n_j，クラス j に含まれる i 番目のパターンを $\boldsymbol{p}_i^{(j)}$，クラス数を K，全パターン数を n，全パターンの平均ベクトルを $\boldsymbol{m}$ とする．クラス内分散 σ_W^2(within class variance) とクラス間分散 σ_B^2(between class variance) を式 (2.54)，(2.55) で表す．

$$\sigma_W^2 = \sum_{j=1}^{K} \sum_{i=1}^{n_j} (\boldsymbol{p}_i^{(j)} - \boldsymbol{c}_j)^t (\boldsymbol{p}_i^{(j)} - \boldsymbol{c}_j)/n \tag{2.54}$$

$$\sigma_B^2 = \sum_{j=1}^{K} n_j (\boldsymbol{c}_j - \boldsymbol{m})^t (\boldsymbol{c}_j - \boldsymbol{m})/n \tag{2.55}$$

ここで，標準化により $\boldsymbol{m} = \boldsymbol{0}$ である．クラスタリング結果の評価は，クラス内分散 σ_W^2 とクラス間分散 σ_B^2 の比である

$$J_\sigma = \frac{\sigma_B^2}{\sigma_W^2} \tag{2.56}$$

の値が大きい結果を用いる．

また，クラス数 K は未知のため，次のクラスタリングの有効性分析によって最適なクラス数 K を決める．すなわち，情報量を基準とした次式の $A(K)$ を最小にする K を採用する [44]．

$$A(K) = \log_{10} V_{mean} + \frac{K}{2} \tag{2.57}$$

ここで，V_{mean} は分散の全クラス平均であり，次式で求められる．

$$V_{mean} = \left\{ \sum_{j=1}^{K} \sum_{i=1}^{n_j} \frac{|\boldsymbol{c}_j - \boldsymbol{p}_i^{(j)}|^2}{n_j} \right\} / K \tag{2.58}$$

式 (2.57) において，第一項は学習データに対するモデルの当てはまりの良さであり，第二項はモデルの自由度である．

第3章　物体操作による視聴覚事象の対応付け

3.1　はじめに

人間は，周囲の環境で生じる様々な事象を認識・理解するとき，受動的な観測で得られる視聴覚情報だけでなく，対象に働きかける運動情報を通じて，対象に関する様々な情報を収集し，それらを統合する．例えば，赤ん坊が玩具を手に取り，遊ぶことでその玩具が発する音や，玩具の形状，色，重さ，質感等の玩具の属性についての知識を得る．

視聴覚情報の統合としては，物体についての固有な知識を用いない，物体の運動で生じた視聴覚事象の対応付け [11] や，視聴覚事象間の相互情報量が最大となる事象の統合がある [45, 46]．しかし，これらの手法は受動的に視聴覚で観測される事象の対応付けであった．人は事象を認識する場合，受動的に観測するだけでなく，能動的に対象へ働きかけ，より注意を向けていると考えられる [47]．

自分の意志で動かしてみることによって，どのような運動が行われたかだけではなく，どのような運動を行おうとしたかに関する，遠心性のコピーと呼ばれる運動情報が得られる [16, 48–50]．赤松ら [15] は，人が外界の対象を知覚するとき，触覚，運動，視覚といった複数の感覚からの情報を統合していると考え，対象物の形状知覚に関して，複数の感覚の相互関連性を心理物理学の手法で解析した．そして，運動情報（位置決め動作）により，視覚と触覚との対応が付きやすくなることを示した．しかし，この研究はインタフェースでの利用を想定し，呈示された複数の感覚を統合するのは人であり，計算機による自動的な統合はなされていない．

また，幼児の言語の獲得には視聴覚情報だけでなく，対象物への身体運動（指さしや物体操作）が大きな役割を果たしているという，正高らの報告 [2, 3] もある．このように，対象物に能動的に働きかけ，運動に関連する視聴覚事象を対応付けることは，物体についての知識や概念の獲得に重要であると考えられる．

能動的に物体に働きかけることによって，物体の特徴を抽出でき，さらに運動情報を介することで，複数の視聴覚事象が存在する環境中においても，自身が操作した対

象物体に関連する視聴覚事象を対応付けることができる [51].

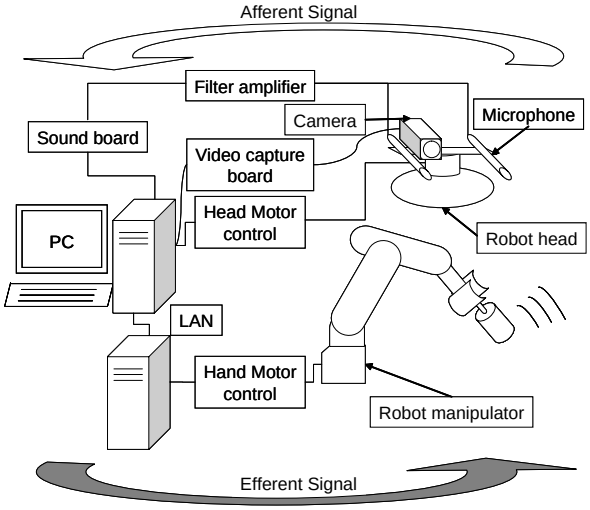
本章では，手に持った物体を操作することで能動的に物体に働きかけ，対象物体の特徴を抽出すると共に，複数の視聴覚事象が存在する環境中において事前知識なしで，自身の働きかけた物体と物体が発する音の対応付けを実現する．実験では，マニピュレータがドラム，ベルといった物体に対してたたく，振るといった操作で働きかけ，その運動によって発生した事象をマイクロフォンとカメラで観測する．ゲシュタルト心理学の群化の要因に基づいて，運動物体が何かに運動を阻害されたときに，そのエネルギーの一部が音として発生する時刻の“ 同時性 ”を尺度として，マニピュレータへ運動信号の変化，画像上の運動の変化，そして音の変化の時刻が類似するものを群化する手掛かりとする．運動の制御信号である遠心性の信号と，視聴覚信号のような観測された求心性信号を対応付けることで，他に視聴覚事象が存在する状況で，運動によって生じた視聴覚事象を対応付ける方法を提案する．

本章の構成は次のとおりである．まず **3.2** では，視聴覚事象の能動的対応付けについて説明する．**3.3** においては，運動，視聴覚における特徴抽出と対応付け処理について述べる．**3.4** で実験方法，**3.5** では，実験結果を示すとともに考察を行い **3.6** で本章をまとめる．

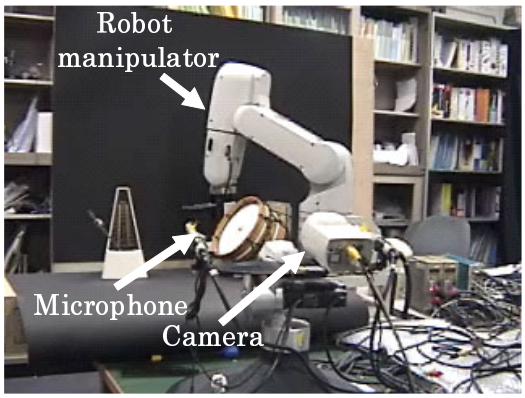
3.2　遠心性と求心性信号の対応付け

視聴覚事象の対応付けを能動的に行う本章のロボットの構成を**図 3.1**(a) に，実験シーンを (b) に示す．ロボットはマニピュレータを持ち，カメラとマイクロフォンを備えている．ロボットの脳に相当する計算機はマニピュレータに制御信号を送る．同時に，計算機はこの制御信号のコピーを得る（神経生理学での，脳から筋への制御命令である遠心性信号のコピー）．マニピュレータによる運動で発現された事象に対して，マイクロフォンとカメラにより観測した信号を，視聴覚によって知覚される求心性信号とする．脳で遠心性のコピーと求心性信号が統合されるように，本章ではこれら遠心性の信号と求心性信号とを計算機上で統合し，視聴覚事象を対応付ける．

陳ら [11] は，ゲシュタルト心理学で用いられる群化の要因を，視聴覚の対応付けの手掛かりとした．その中で述べられている“ 共通運命の要因 ”とは，運命を共にするもの，共に変化し，共に動くものは，1 つにまとまる傾向である．また“ 類同の要因 ”とは，多くの刺激がある場合，他の条件が一定であるならば同種のものがまとまる傾向である [52].



(a) Robot system



(b) Experiment scene

図 3.1: 遠心性と求心性の信号の統合システム
Fig. 3.1 System of integrating afferent and efferent signals

“ 共通運命の要因 ”の尺度を本章では，運動物体が何かに運動を阻害されたときに，そのエネルギーの一部が音として発生する“ 同時性 ”として，音の変化と画像上の運動の変化の時刻が類似するものが群化すると考え，対応付けの手掛かりとする.“ 類同の要因 ”の尺度を運動の繰り返しと音の変化，立ち上がりと画像上の変化の繰り返しの“ 類似性 ”とし，同種のものが群化するとして，対応付けの手掛かりとする.

3.3　対応付け処理

遠心性の信号と求心性の信号を対応付ける処理の概要を**図 3.2** に示す．図のように処理は，(1) 運動処理部，(2) 聴覚部，(3) 視覚部，(4) 統合部から構成される．(1) の運動処理部では，マニピュレータへの運動指令から運動方向変化時刻の時系列 $M_{in}(t)$ を得る．本研究では，この運動指令を運動情報とよぶ．(2) の聴覚部で，音を入力として同じ音源の音オンセット (音の先頭部分で，詳細は **3.3.1.(a)**) の時系列 $A_k(t)(k = 1, \cdots, m; m$ は音源数) を出力する．(3) の視覚部で，映像中の物体の運動が行われていると推定される範囲，すなわち運動範囲ごとの運動方向変化の時系列 $V_l(i)(l = 1, \cdots, n; n$ は運動範囲数) を算出する．(1) 運動処理部の入力は遠心性信号のコピーで，(2) 聴覚部と (3) 視覚部の入力は求心性の信号である．(4) の統合部で各信号時系列間の相関を求め，相関の高いものを同じ事象であるとして対応付ける．

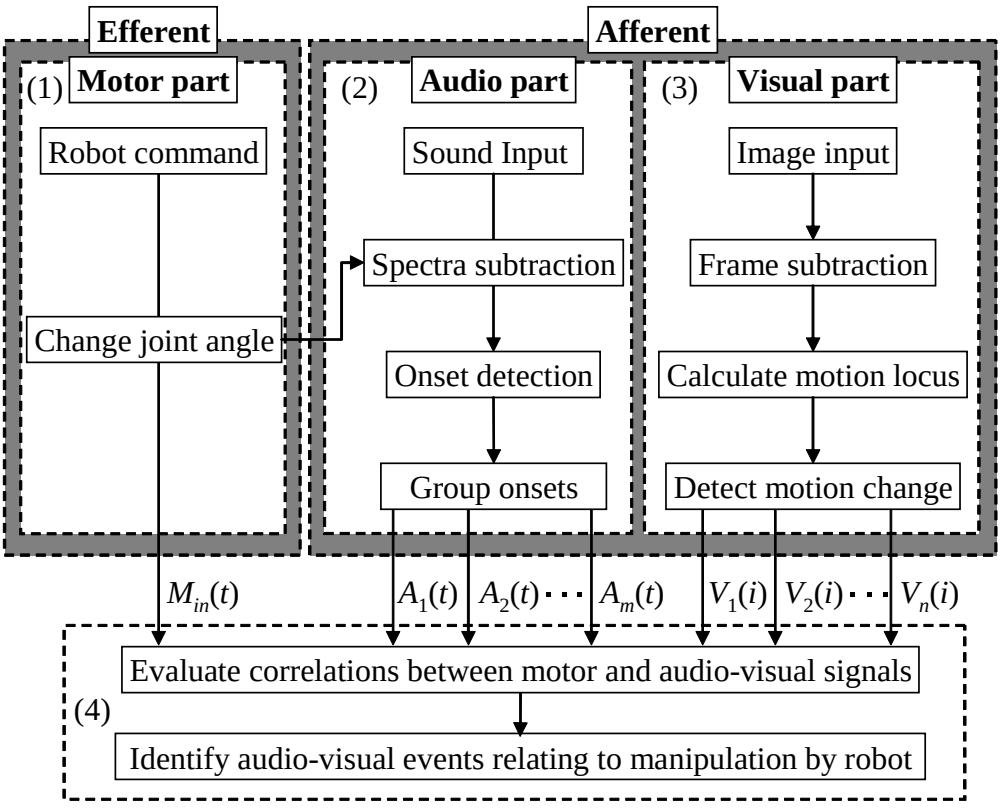


図 3.2: 遠心性と求心性の信号の統合処理
Fig. 3.2 Integration of afferent and efferent signals

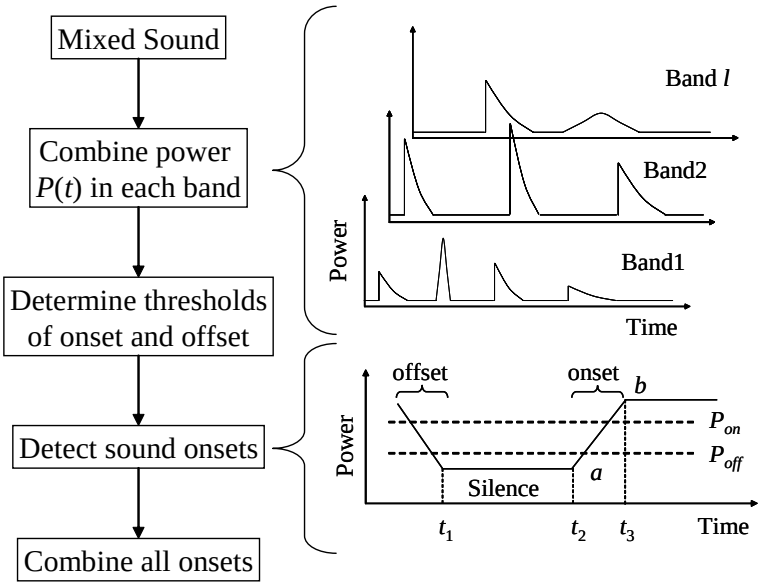


図 3.3: 周波数軸上での音オンセット検出
Fig. 3.3 Onset detection in frequency domain

3.3.1 音オンセットの検出と分離

聴覚部では複数の音が混合した 1 つの音信号から，各音源の音オンセット時系列 $A_k(t)$ を求める．その概要は次のとおりである（図 3.2(2) 参照）．まず，周波数領域で混合音から雑音を減算することで，目的音を抽出するスペクトルサブトラクション（SS : spectral subtraction）により背景雑音を音信号から除去した後，音オンセットを検出する．音オンセットのパワースペクトルを各音オンセット間で比較し，相関が高いものは同じ音としてグループ分けを行い，音オンセットの時系列 $A_k(t)$ を算出する．

(a) 運動情報を利用した雑音除去

雑音環境下での対象音検出のために，本章ではスペクトルサブトラクションを行い，混合音から推定した雑音を周波数領域において除去する．雑音は，マニピュレータが動き出す直前の 10 フレーム (1 フレーム 512 点) の平均から推定する．なお，実験では，フーリエ変換のフレーム長を 512 点，シフト量をフレーム長の半分とし，窓関数にハミング窓を用いる．

(b)　音オンセット検出

背景雑音を除去した後，入力混合音から音オンセットを検出する．音オンセットとは背景雑音以外の音がない状態で，音の大きさが急激に変化した時刻からの短時間の音である．複数の音が存在した中でも同時に発生していなければ，単一音源からの音しか含まず，また残響の影響を受けないといった特徴を持つため，音響信号の分離にこの音オンセットを利用することは有効である．また音オフセットは音信号が有音状態から背景雑音以外の音がない無音状態へ変化する音の終了部分であり，音オフセットから次の音オンセットまで十分な無音区間を設定することで音オンセットを正確に検出する．

A. Klapuri は，バンドパスフィルタを用いて周波数帯域別に音オンセット要素を検出した後，全帯域の情報を時刻と振幅に基づき統合して，全体の音オンセットを検出している [53]．しかし，音オンセット検出の閾値および音オンセットを統合するときの時間間隔の閾値を，経験的に決める必要がある．

音オンセットを検出するため，まず**図 3.3** のように，入力混合音に対してフーリエ変換 (サンプリング 512 点，シフト量 256 点) を行い，0[Hz] から重なりなく 1 [kHz] ずつ，8 つの周波数帯域に分ける．次に，音オンセットと音オフセットを検出する閾値を決定するため，周波数帯域ごとの計測時間内の音パワーの平均値を音オフセット検出閾値 P_{off} とし，音オンセット検出閾値 P_{on} を P_{off} の 2 倍と決定する．音信号のパワーが P_{on} 以上の状態から P_{off} 未満に立ち下がる部分を音オフセット，パワーが P_{off} 未満の区間 $[t_1, t_2]$ を背景雑音以外の音がない無音状態，P_{off} 未満であるパワー a から P_{on} 以上であるパワー b に音が立ち上がる部分を音オンセットとし，区間 $[t_2, t_3]$ で最初にパワーが音オンセット閾値 P_{on} 以上となる時刻を検出する．

サンプリング周波数 16 [kHz] により計測した音信号を，サンプリング 512 点，シフト量 256 点 としてフーリエ変換を行い，周波数領域で音オンセットを検出する場合，16 [kHz]/256 = 62.5[Hz] の周期で音オンセットを検出する．各周波数帯域中のパワーから音オンセットを検出した後，全帯域の音オンセットすべてを 1 つの集合とする．

(c)　音のグルーピング

複数の音がある状況では，**3.3.1.(b)** で検出された音オンセットは必ずしも同一音源ではなく，また 1 つの事象の音で残響の影響により複数の音オンセットとして，検出されていることも考えられる．そこで，音オンセットを音源ごとに分けてグルーピン

グし，音源ごとの音オンセットを検出する．まず，フーリエ変換（512 点）により，各音オンセットでのパワースペクトル $\boldsymbol{PS} = \{ps_1, \cdots, ps_i, \cdots, ps_j, \cdots\}$ を求める．得られたパワースペクトルの類似度を式 (3.1) の相関係数として算出し，相関の高い音オンセットを同一音源にし，相関の低い音オンセットはそれぞれ別音源とする．

$$Cor(ps_i, ps_j) = \frac{Cov(ps_i, ps_j)}{\sqrt{Var(ps_i)Var(ps_j)}} \tag{3.1}$$

音オンセットの分類アルゴリズムを以下に示す．

[音オンセットの分類アルゴリズム]

Step 1 すべての音オンセットでのスペクトル $\boldsymbol{PS}$ からなるグループを S_1 とし，これに含まれるスペクトル同士の相関係数を求める．$i := 1, k := 1$ とする．

Step 2 S_1 の中でスペクトル ps_i と**条件** 1，**条件** 3 を満たす音オンセットを音源グループ G_k とする．

Step 3 S_1 の中で ps_i との相関係数が最小のスペクトル ps_{min} を求める．

Step 4 S_1 の中で ps_i との**条件** 2 を満たす音オンセットは，音源グループ S_0 とする．

Step 5 $S_1 = G_k$ または，$S_0 = \phi$ なら，処理を終了する．そうでなければ，次の **Step** へ．

Step 6 $S_1 := S_1 - S_0$，$k := k + 1$，$i := min$ とし，**Step2** へ戻る．

ただし，**条件** 1〜**条件** 3 は以下のとおりである．
条件 1：相関係数が閾値 0.5 以上．
条件 2：相関係数が閾値 0.7 以上．
条件 3：同じグループの音オンセット間の時間間隔 $T_o \geq \frac{t_m}{2}$
ただし，t_m は運動方向変化の時間間隔の平均である．

条件 1 は，同一音源であっても，他の複数の音が混合した場合に，スペクトルの間の相関が低くなるために 0.5 とした．**条件** 2 は，**Step2** の処理で音源グループ G_k に含まれてなくても，相関値が高ければ G_k と同一音源の残響であると考え，**Step6** の処理で G_k を含むそれらの音源グループ S_0 を音源グループ S_1 から除くために 0.7 とした．

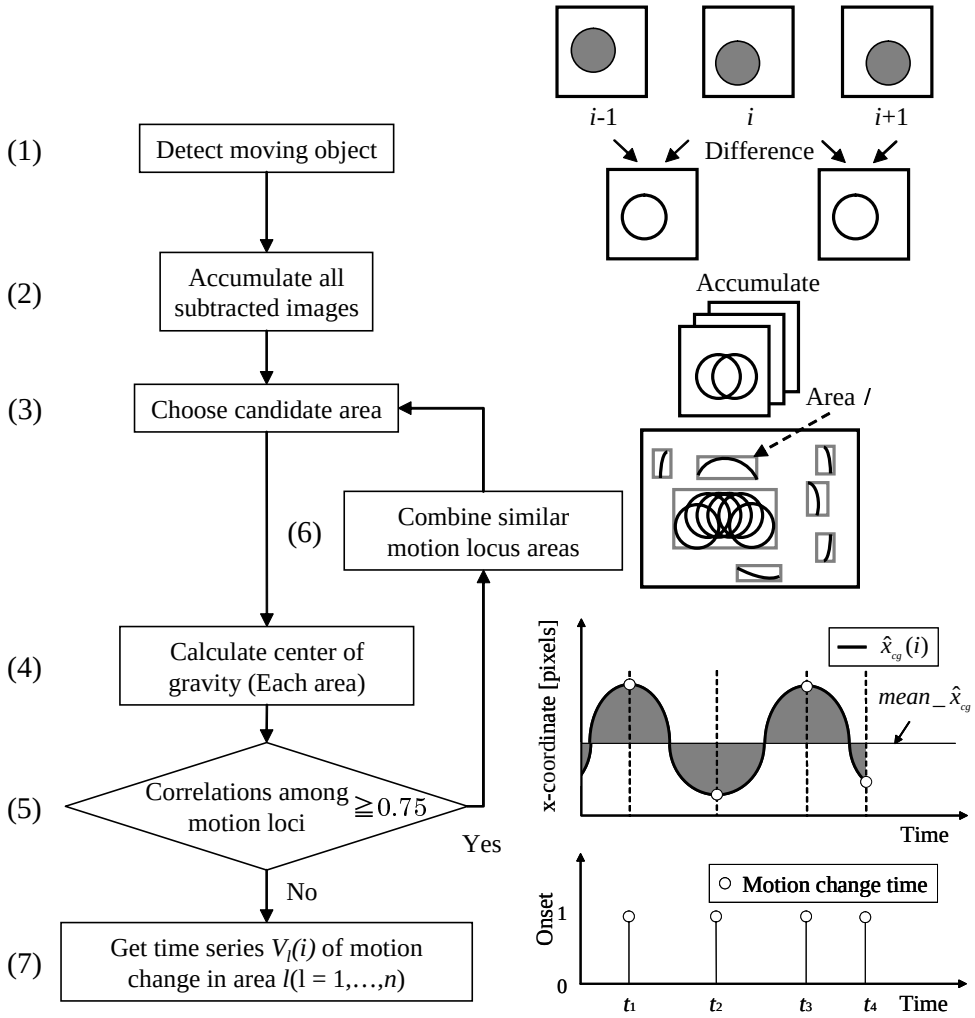


図 3.4: 視覚処理による運動方向変化の検出
Fig. 3.4 Detection of motion direction change by visual process

(d)　聴覚信号時系列 $A_k(t)$ の算出

音源グループごとに聴覚信号時系列 $A_k(t)$ を算出し，図 3.2(4) の統合部へ出力する．

3.3.2　映像上の運動方向変化の検出

視覚処理では，運動物体の検出，運動範囲の決定，運動軌跡の計算の後，運動方向変化の時系列を出力する（図 3.2 の (3) 参照）．

(a) 物体の運動範囲の決定

図 3.4 に運動方向の変化の検出方法を示す．まず，画像中の運動物体の領域を検出する．物体の運動の検出処理を簡単にするため，入力である RGB カラー画像を輝度画像に変換する．座標 (x, y) における RGB 各画素の値を $f_R(x, y)$，$f_G(x, y)$，$f_B(x, y)$ とすると，変換後の輝度画像 $g(x, y)$ は次式で求められる．

$$g(x, y) = 0.2989 f_R(x, y) + 0.5866 f_G(x, y) + 0.1145 f_B(x, y) \tag{3.2}$$

図の (1) に示すように，連続した 3 フレームの輝度画像 $g_{i-1}(x, y)$, $g_i(x, y)$, $g_{i+1}(x, y)$ の間の差で，運動物体の抽出を以下のように行う．まず，式 (3.3) により 2 画像間の差を計算する．

$$\begin{aligned} \Delta g_{i-1,i}(x, y) &\equiv |g_{i-1}(x, y) - g_i(x, y)| \\ \Delta g_{i,i+1}(x, y) &\equiv |g_i(x, y) - g_{i+1}(x, y)| \end{aligned} \tag{3.3}$$

それらの論理積を，次式により算出する．

$$g(x, y) = 1(\Delta g_{i-1,i}\,(x, y)) \cap 1(\Delta g_{i,i+1}\,(x, y)) \tag{3.4}$$

ここで，$1(u)$ は次式に示す閾値関数で，その閾値は予備実験で決めた値 $th = 15$ とする．

$$1(u) = \begin{cases} 1 & \text{if} \quad u \geq th \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

この処理によって時刻 i における物体のフレーム間差分領域が抽出される．その 4 近傍の連結領域を運動物体の領域とする．

図 3.4(2) の処理において，すべての時刻での運動物体の領域を重ね合わせ，画素点の値の和を求める．図 (3) の処理で，1 以上の値の画素点の 4 近傍の連結領域を物体の運動範囲とする (処理には影響しないが，運動範囲を明確にするため領域の外接矩形を表示する)．

(b) 対象物の移動軌跡の算出

すべての時刻 i での物体のフレーム間差分領域を運動範囲ごとに算出し，動領域内の物体重心 (center of gravity 以下，c.g.) の x 座標を次式で計算する（図 3.4(4)）．

$$x_{cg}(i) = \frac{\sum_{x=1}^{w} N_p(x, i)x}{\sum_{x=1}^{w} N_p(x, i)} \tag{3.6}$$

$x_{cg}(i)(i = 1, \cdots, m)$ は時点 i における動領域の重心の x 座標位置である．式 (3.6) において w は画像幅で，$N_p(x, i)$ は x 座標の値が x の動領域内の画素数である．運動範囲内の運動が止まった時点 i では，重心座標 $x_{cg}(i)$ 値が算出されないため，$x_{cg}(i) = 0$ のとき，$x_{cg}(i) := x_{cg}(i-1)$ として 1 時点前の重心座標によって補間する．

得られた動領域の重心の x 座標の時系列からノイズの影響を除くため，次式の移動平均により $-l$〜l 区間の平滑化を行う．

$$\tilde{x}_{cg}(i) = \frac{1}{\sum_{k=-l}^{l} w_k} \sum_{k=i-l}^{i+l} w_{k-i} x_{cg}(k) \tag{3.7}$$

上式で，$l = 4$ とし，重み w_k は次式のガウス分布で $\sigma = 5$ とする．

$$w_k = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{k^2}{2\sigma^2}\right\} \tag{3.8}$$

平滑化のため欠落した最初と最後の l 区間は，次式のように，1 つ後と 1 つ前の $\tilde{x}_{cg}(i)$ の座標値により補間する．

$$\hat{x}_{cg}(i) := \begin{cases} \tilde{x}_{cg}(i+1) & 1 \le i \le l \\ \tilde{x}_{cg}(i-1) & m-l \le i \le m \end{cases} \tag{3.9}$$

(c)　物体の運動範囲の結合

1 つの物体でも複数の運動範囲に分かれてた場合，それぞれの運動範囲での軌跡が類似すると考えられる．そこで，運動範囲ごとの軌跡 $\boldsymbol{Locus}$= $\{locus_1, \cdots, locus_k, \cdots, locus_l, \cdots\}$ 同士の相関を式 (3.10) で求め値を比較する（図 3.4 (5)）．相関係数が閾値 0.75 よりも高い値の運動範囲があれば，処理 (6) において領域を統合し処理 (3) に戻る．ここで $locus_k$ は，マニピュレータが運動を開始してから静止するまでの時間の運動範囲 k の $\hat{x}_{cg}(i)$ の集合である．

$$Cor(locus_k, locus_l) = \frac{Cov(locus_k, locus_l)}{\sqrt{Var(locus_k) Var(locus_l)}} \tag{3.10}$$

(d)　視覚信号時系列 $V_l(i)$ の算出

運動範囲ごとの，平滑化された物体の運動軌跡 $\hat{x}_{cg}(i)$ から，図 3.4(7) に示すように，運動方向が変化する時刻での値が 1 で，その他は 0 の時系列を視覚信号時系列 $V_l(i)$ と

する．運動方向が変化する時刻を求めるために，まず全体のフレーム数を N_f としたとき，運動物体の位置の平均 $mean\_\hat{x}_{cg}$ を次式で求める．

$$mean\_\hat{x}_{cg} = \frac{1}{N_f}\sum_{i=1}^{N_f}\hat{x}_{cg}(i) \tag{3.11}$$

運動方向変化の時刻 $tv_l(l = 1, \cdots, n)$ の検出方法を以下に示す．ここで，$t(i)$ は第 i フレームの時刻とする．

[運動方向変化の時系列の検出]

Step 1 運動領域の重心軌跡 $\hat{x}_{cg}(i)$ の 1〜N_f 区間の平均値 $mean\_\hat{x}_{cg}$ と運動方向変化の時間間隔の平均 t_m を求める．$i := 1$，$k := 1$，$l := 1$ とする．

Step 2 速度 $v(i+1) := \hat{x}_{cg}(i+1) - \hat{x}_{cg}(i)$ を求め，$v(i+1)v(i) \leq 0$ であれば時刻 $t_k := t(i+1)$ とする．そうでなければ $i := i+1$ として，**Step2** をもう一度行う．

Step 3 $k = 1$ であれば，**Step5** へ，そうでなければ次へ．

Step 4 時刻 t_k において $(\hat{x}_{cg}(tv_l) - mean\_\hat{x}_{cg})(\hat{x}_{cg}(t_k) - mean\_\hat{x}_{cg}) \leq 0$ かつ，$t_k - tv_l \geq t_m$ であるならば，次の **Step** へ．そうでなければ，$i := i+1$ として，**Step2** へ．

Step 5 運動方向変化の時刻 $tv_l := t_k$ とし，運動方向変化の回数を $l := l+1$ とする．

Step 6 $k := k+1$ とし，$N_f = i+1$ であれば処理を終了する．そうでなけでば，**Step3** へ戻る．

対象物をたたいて事象を発生させる場合では，対象がある側の運動方向変化の時系列を抽出し，対象物を振って事象を発生させる場合では，運動方向の変化ごとに音が鳴るとし，運動方向の変化ごとに時系列を検出しなければならない．そこで，視覚部では，運動の変化する時系列を両側，右側，左側の 3 通り求め，統合部において，運動－聴覚事象と対応付けられる最適な運動方向変化の時系列を求める．以上の処理を図 3.4(3) で求めたすべての運動範囲について行う．

3.3.3 遠心性と求心性信号の統合

これまでに述べた処理により得られた時系列，すなわち，マニピュレータへの運動信号の変化時刻の時系列 $M_{in}(t)$，聴覚信号時系列 $A_k(t)$，視覚信号時系列 $V_l(i)$ を，“ 同時性 ”と繰り返しの“ 類似性 ”に着目し対応付ける．

聴覚の 16 [kHz]（周波数領域での音オンセット検出時は 62.5[Hz]）と視覚の 30[Hz] ではサンプリング周期が異なるので，統一的に評価するため，最小のサンプリング周期である視覚と同じにする．時系列 $M_{in}(t)$，$A_k(t)$，$V_l(i)$ に対し，各信号時点を中心に最大値が 1，底辺の長さが運動方向の変化の平均時間間隔の 1/2 の時間になるように，三角波の重みを付ける．そして，0 秒から 1/30 秒ごとの三角波の値により，$\hat{M}_{in}(i)$，$\hat{A}_k(i)$，$\hat{V}_l(i)$ を得る．

これら運動と視聴覚の信号を“ 同時性 ”と“ 類似性 ”を手掛かりとして対応付けるため，相関関数を評価尺度として用いる．計測誤差，および信号抽出処理での誤差の影響を低減するため，運動，聴覚，視覚情報の相関関数をシフト量を考慮して求める．そのため，次式のように聴覚-視覚の相互相関関数の最大値を算出する．

$$MC(A_k, V_l) = \max_{-8 \leq s \leq 8} \frac{Cov(\hat{A}_k(i), \hat{V}_l(i+s))}{\sqrt{Var(\hat{A}_k) Var(\hat{V}_l)}} \tag{3.12}$$

ここで，i は運動を行っている区間，s はシフト量であり，$-8 \leq s \leq 8$ (時間に換算すると ± 0.264 秒) とする．同様に，運動-聴覚，運動-視覚についても，相互相関関数の最大値を求める．

さらに，聴覚-視覚の相互相関関数の最大値に対して，運動情報を対応付ける．式 (3.12) によって求めた運動-聴覚および運動-視覚の相関を次式のように用いて，運動に関連する視聴覚事象の尺度とする．

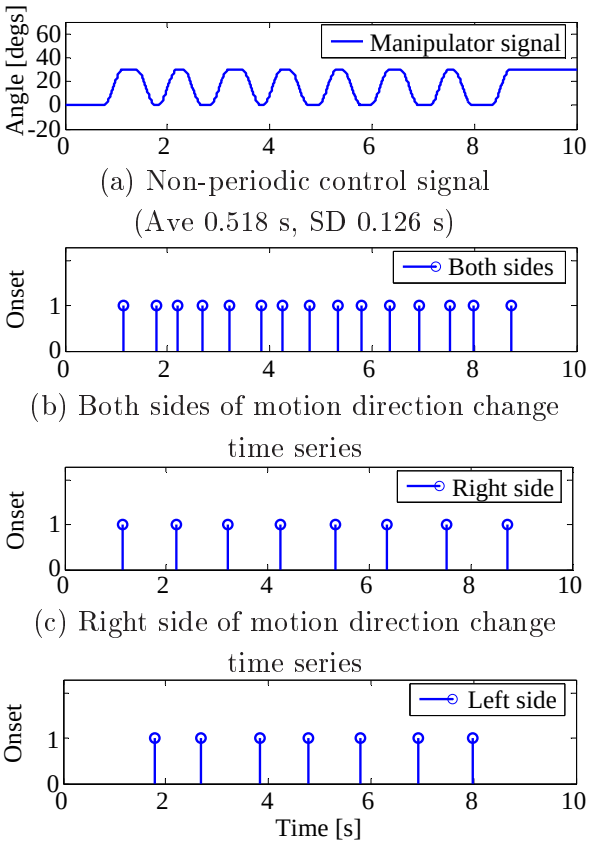
$$Motor(A_k, V_l) = MC(A_k, V_l) \frac{MC(M_{in}, A_k)}{\max_k MC(M_{in}, A_k)} \frac{MC(M_{in}, V_l)}{\max_l MC(M_{in}, V_l)} \tag{3.13}$$

上式において，最大値が 0.5 以上であるとき，その視聴覚事象は運動によって生じたものと対応付ける．

3.4　実験

3.4.1　システム構成

実験には，図 3.1(a) のシステムを使用した．マニピュレータ（Mitsubishi Electric, RV-1A）を用いて対象物を把持し，肘関節角度を変化させて，音や動きの変化といった事象を発生させる．1 個のマイクロフォン（RION，UC-30）と，1 台のカメラ（SONY, EVI-G20）を備えた擬似ヘッドと，マイクロフォンとカメラからの信号を計測する計算機（Dell, PowerEdge SC420, Linux OS），マニピュレータ制御とデータ処理のための計算機（VAIO, VGN-S70B, Windows OS）を使用する．Linux OS において，録音



(a) Non-periodic control signal (Ave 0.518 s, SD 0.126 s)

(b) Both sides of motion direction change time series

(c) Right side of motion direction change time series

(d) Left side of motion direction change time series

図 3.5: マニピュレータへの運動信号
Fig. 3.5 Motor signal to manipulator

は J. Tranter [54]，画像キャプチャは飯尾 [55] のプログラムを参考にした．マニピュレータ制御，通信に専用ソフト（3A-01C-WINJ）を用い，通信サーバを介して動作指令を送る．

3.4.2 実験条件

マニピュレータの運動による視聴覚事象の発生に加え，他の視聴覚事象が存在するという一般的な設定の一例として，対象事象の他に音や動きの事象がある状況下で視聴覚事象の対応付け実験を行った．背景は黒色の画用紙（マーメイド）とし，他事象としてメトロノームを用いた．

マニピュレータへの指令は，**図 3.5(a)** のように肘関節の角度を 30[deg] の範囲で周期的，または非周期的に変化させるように送る．(i) マニピュレータがばちを把持して，

肘関節の角度を変化させ，予め決られた位置に置いてある玩具のドラムをたたく．あるいは (ii) ベルを把持して，肘関節の角度を変化させ，振るという試行を行う．

対象物をたたいて事象を発生させる場合は，対象物を振って事象を発生させる場合と，視聴覚事象を発生させる運動が異なる．そのため，運動方向が変化する時系列を図 3.5(b) 両側，(c) 右側，(d) 左側のように 3 通り求め，統合部において，視聴覚事象と対応付けられる最適な運動方向変化の時系列を求める．非周期的な運動では，肘関節を 30[deg] 変化させるための，時間間隔を標準偏差（以下，SD）0.126 秒の正規分布に従うランダムなデータとする．

5 つの異なる平均時間間隔 (以下，Ave) で，周期的，非周期的の 2 つの異なる動きで対象物に働きかけ，各条件について 2 回，計 20 回の実験を行う．これをメトロノームが動いていない，またはメトロノームが動いている状態で，ドラム，ベルに働きかける 4 つの場合について，つまり全体で 80 回の視聴覚事象の対応付け実験を行う．

音はサンプリング周波数 16 [kHz]，量子化 8[bits]，映像は画像サイズ 320×240[pixels]，30[Hz] で，音，映像共に 10 秒間計測した．

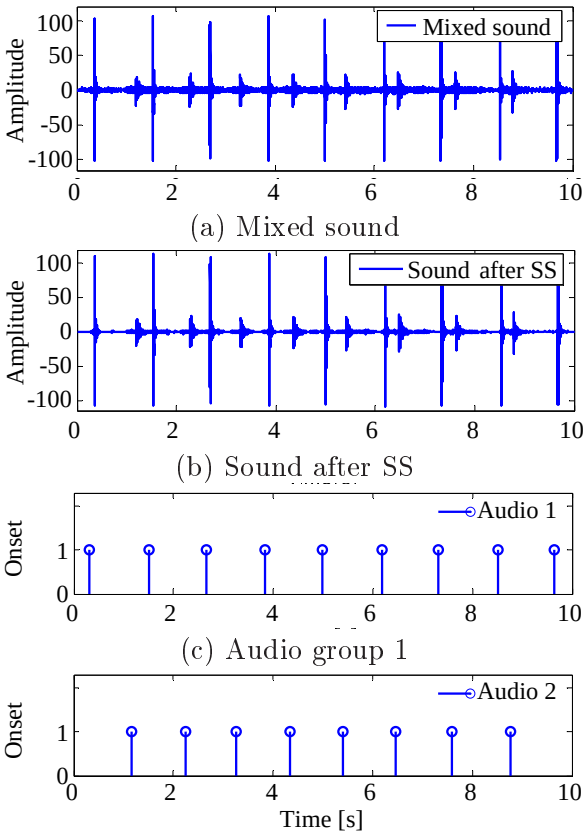
3.5　結果と考察

メトロノームが動いている状態で，マニピュレータが把持したばちでドラムをたたいた場合について説明する．マニピュレータへの運動指令は，図 3.5(a) のように，肘関節を 30[deg] の範囲で平均 0.518 秒，標準偏差 0.126 秒の正規分布に従うランダムな時間間隔となる，非周期的な運動である．

3.5.1　聴覚処理

図 3.6 に音処理の結果を示す．図 3.6(a) に示す音波形を観測し，SS により図 (b) のように雑音部分を除いた．その結果から音オンセットを検出し，グループ化した結果，図 3.6(c)，(d) の 2 グループの音オンセット時系列が得られた．音源グループ内のスペクトル同士の相関係数の平均は，グループ 1 が 0.959，グループ 2 が 0.910 であり本手法での分類が正しいことを確認した．

図 3.6(a) のような異なる振幅の混合入力音に対して，音オンセットのための閾値を周波数領域で適応的に求め，図 3.6(c)，(d) のように良好に音オンセットを検出することができた．



(a) Mixed sound

(b) Sound after SS

(c) Audio group 1

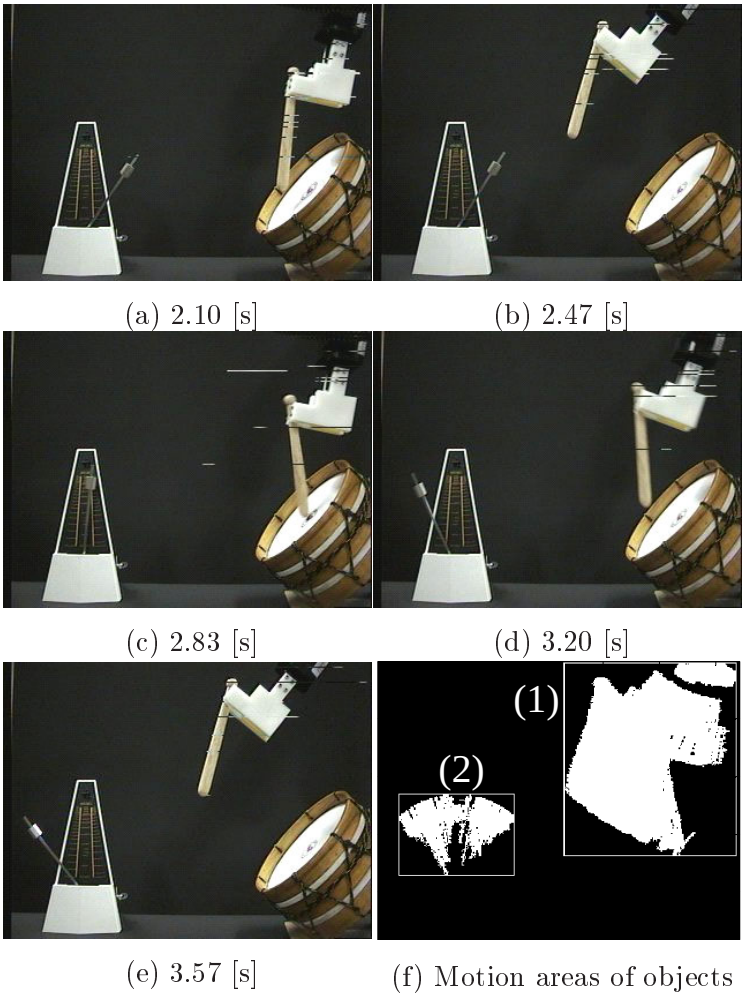
(d) Audio group 2

図 3.6: 聴覚処理の結果

Fig. 3.6 Results of audio process

音源ごとの音の振幅差が小さい混合音であれば，振幅から直接に音オンセット検出のための閾値を適応的に決定することは容易であるが，本実験のように，どのような振幅の音が混合するか事前知識を持たず，振幅差の大きな混合音を観測した場合，時間方向で音オンセットを適応的に決定することは容易ではない．本システムは周波数帯域で音オンセット閾値を適応的に決定するため，図 3.6(a) のように大きく振幅の異なる複数音に関して音オンセットの検出が可能である．本章では，類似する音が別々の音源から発生している場合はまれであると考え，想定していない．そのような場面の音源分離は難しくなるが，音源定位や映像情報の利用により分離可能であると考えられる．

図 3.6 において環境雑音は，メトロノームやドラムの音に比べ小さいため，SS の効果は小さい．しかし，周波数帯域ごとに音オンセット検出閾値を決定しており，環境雑音のパワーが強い周波数帯域において雑音の影響は無視できず SS を用いた．



(a) 2.10 [s]　(b) 2.47 [s]

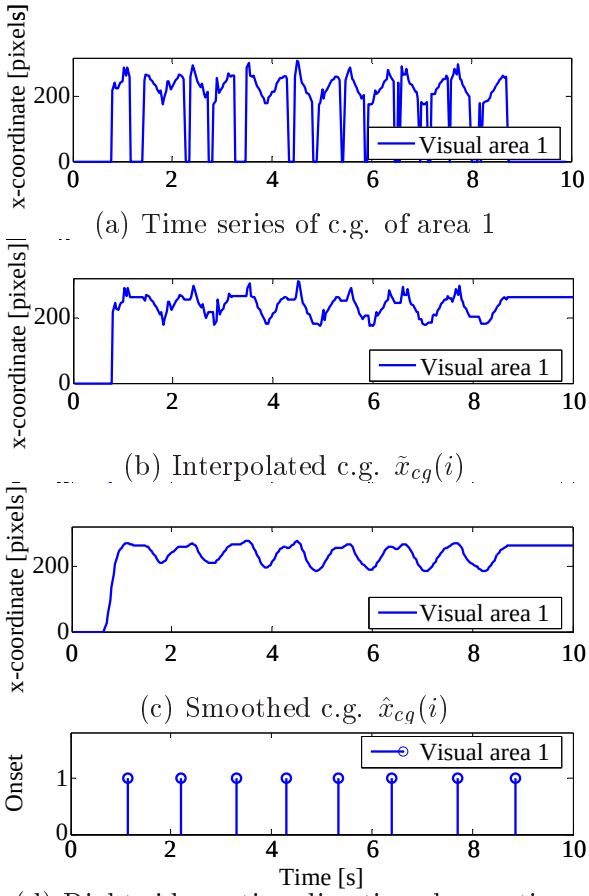
(c) 2.83 [s]　(d) 3.20 [s]

(e) 3.57 [s]　(f) Motion areas of objects

図 3.7: 抽出された運動範囲
Fig. 3.7 Extracted motion areas

3.5.2　視覚処理

図 3.7(a)〜(e) のように，2 物体が別々に運動している入力画像列に対し，運動範囲を抽出した結果，図 (f) に示すように 2 つの運動範囲が得られた．運動範囲 1 についての運動方向変化の時系列の抽出を**図 3.8** に示す．図 3.8(a) は，画像のフレーム間差分による処理で得られた物体の重心の観測値の時系列である．図中のデータの欠落は，実験で用いたドラム，ベルが運動中鳴り続けないよう，運動方向変化の後に運動を短時間停止したためである． 物体の運動軌跡を画像のフレーム間差分により求めたので，運動停止時の重心軌跡は抽出できず，データが欠落する．得られた時系列に対して欠落した時点のデータを一時点前の重心座標によって補間し，その後全体に対して平滑化処理を行った結果が，それぞれ図 3.8 (b),(c) の軌跡である．この軌跡に対して，図



(a) Time series of c.g. of area 1

(b) Interpolated c.g. $\tilde{x}_{cg}(i)$

(c) Smoothed c.g. $\hat{x}_{cg}(i)$

(d) Right side motion direction change time series

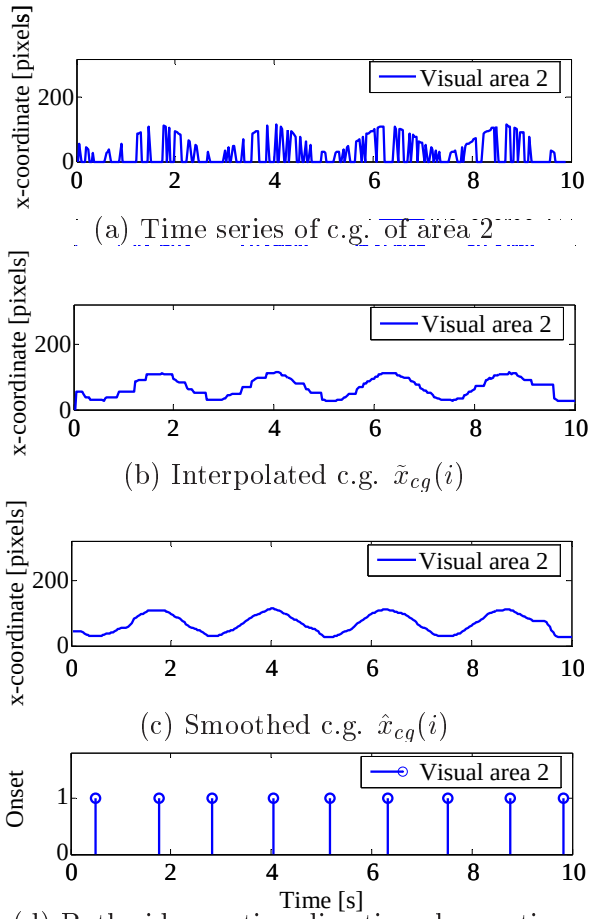
図 3.8: 運動範囲 1 における運動方向の変化
Fig. 3.8 Motion direction change of area 1

3.4(7) の処理によって得られた右，左，両側における 3 つの時系列の内，本実験例での正しい位置である，右側の運動方向変化の時系列を図 3.8(d) に示す．

同様に，運動範囲 2 における物体の重心軌跡を順に**図 3.9** (a)〜(c) に表す．図 3.9(d) は，右，左，両側の内，両側の運動方向変化の時系列である．

運動物体の位置と速度変化を組み合わせて運動方向変化の時刻を検出したため，図 3.8 に示すように対象が運動を止める場合，図 3.9 のように常に対象が動いている場合に関わらず，運動方向変化の時刻がずれることなく，良好に抽出できた．しかし，複数の運動物体の運動範囲が重なったときには，各物体の個々の動きの抽出が困難になる．

物体の運動範囲の決定において，全 80 回の実験中 1 つの運動物体が複数の運動範囲に分断されたのは 43 回あり，運動範囲の結合処理における閾値 0.75 により正しく結合されたのは 37 回，結合されなかったのは 6 回あった．また複数の物体がそれぞれ



(a) Time series of c.g. of area 2

(b) Interpolated c.g. $\tilde{x}_{cg}(i)$

(c) Smoothed c.g. $\hat{x}_{cg}(i)$

(d) Both sides motion direction change time series

図 3.9: 運動範囲 2 における運動方向の変化
Fig. 3.9 Motion direction change of area 2

の運動範囲の検出に成功した後，結合処理によって誤って結合されたのが 1 回あった．したがって，この閾値により，異なる運動範囲中の物体が，同じ物体で分断されたのか，別々の物体であるのかを判別できることを確認した．

3.5.3　統合処理

表 3.1 は，右側の運動方向変化を検出した運動－視聴覚事象間の相互相関関数の最大値である．これらは式 (3.12) で求めた運動－聴覚（表右下 Audio-Motor），運動－視覚（表左上 Motor-Visual），視覚－聴覚（表中央 Audio-Visual）の間の相互相関関数の最大値を示す．運動と視覚事象 1 において 0.743，運動と聴覚事象 2 において 0.889，聴覚事象 2 と視覚事象 1 において 0.785 となり，一意に運動と視聴覚事象の対応付け

表 3.1: 運動-視聴覚間の相互相関関数の最大値（右側運動方向変化検出時）
Table 3.1 Maximum cross-correlation coefficient among motor, audio, and visual series (Motion direction change at the right side)

		Motor	Audio group	
			1 (Met.)	**2 (Drum)**
Visual group	**1 (Drum)**	**0.743**	0.279	**0.785**
	2 (Met.)	0.290	0.640	0.235
Motor			0.419	**0.889**

表 3.2: 運動-視聴覚間の対応付け（右側運動方向変化検出時）
Table 3.2 Correspondence among motor, audio and visual series (Motion direction change at the right side)

		Audio group	
		1 (Met.)	**2 (Drum)**
Visual group	**1 (Drum)**	0.131	**0.785**
	2 (Met.)	0.118	0.092

ができた．これから，遠心性と求心性の信号に相関関係が見られ，対応付けを行えたことを確認した．さらに，式 (3.13) により，運動－聴覚，運動－視覚による相関を考慮した視聴覚事象の相関関係を**表 3.2** に示す．聴覚事象 2 と視覚事象 1 の値が 0.785 と最大となり，運動と視聴覚事象の対応付けが行えた．したがって，運動情報を用いることで，物体操作によって，発生した視聴覚事象が対応付けられた．**表 3.3** は，両側の運動方向変化を検出したときの運動－視聴覚事象間の相互相関関数の最大値である．この結果から聴覚事象 1 と視覚事象 2 において 0.895 と最大となり，メトロノームの視聴覚事象の対応付けができた．運動－視覚，運動－聴覚による相関を考慮した視聴覚事象の相関を**表 3.4** に示す．ここでは，0.327 と低い値となり，物体操作とは関係のない視聴覚事象であることを確認した．

表 3.5 は，運動方向変化の検出を，両側，右側，左側で行った場合の運動，聴覚事

表 3.3: 運動-視聴覚間の相互相関関数の最大値（両側運動方向変化検出時）
Table 3.3 Maximum cross-correlation coefficient among motor, audio, and visual series (Motion direction change at both sides)

		Motor	Audio group	
			1 (Met.)	2 (Drum)
Visual group	1 (Drum)	0.648	0.337	0.549
	2 (Met.)	0.391	**0.895**	0.351
Motor			0.385	0.637

表 3.4: 運動-視聴覚間の対応付け（両側運動方向変化検出時）
Table 3.4 Correspondence among motor, audio and visual series (Motion direction change at the both sides)

		Audio group	
		1 (Met.)	2 (Drum)
Visual group	1 (Drum)	0.204	0.549
	2 (Met.)	**0.327**	0.212

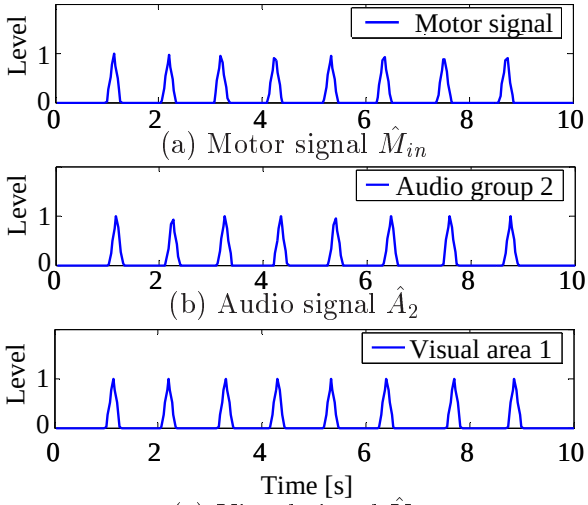
表 3.5: 異なる運動方向変化検出時の運動-視聴覚間の対応付け結果
Table 3.5 Result of correspondence among motor, audio and visual series in case where detection methods of motion direction change are different

	Correlation		Correspondence	
	(M, A)	(M, V)	(A, V)	Judgment
Both sides	0.637	0.648	0.549	false
Right side	**0.889**	**0.743**	**0.785**	**true**
Left side	0.349	0.773	0.415	false

象 2，視覚事象 1 の対応付け結果である．右側の場合の値が 0.785 で最大となり，これは右側でドラムをたたいた事象を正確に表している．一方，聴覚事象 1 と視覚事象 2 の間の値は，両側で検出した場合が 0.895 で最大となり，両側にメトロノームの針が振れたときに音が発生する事象を表している．この結果から，運動方向の変化が右側であったときのみ，視聴覚事象を発生させていることを事前知識なしで認識できた．

遠心性と求心性信号の統合を行った結果，2 つのグループが得られ，第一グループを**図 3.10**，第二グループを**図 3.11** に示す．図 3.10 は，(a) マニピュレータへの運動信号, (b) 音源グループ 2 と (c) 運動範囲 1 の映像の時系列である．図 3.11 は，(a) 音源グループ 1 と (b) 運動範囲 2 の映像の時系列である．**図 3.12** では，音オンセットを表し，物体操作と対応付けられた音オンセット検出時刻を + で表した．**図 3.13** は，入力混合音信号と音オンセットであり，**図 3.14** は，音オンセットに対応する音スペクトルと映像である．図 3.13 は，(a) 入力混合音，(b) 音オンセットであり，図 3.14 は，(a) 物体操作に対応付けられたグループ 1 に属する音スペクトルと対応付けられた映像，(b) グループ 2 に属する音スペクトルと対応付けられた映像である．

表 3.6 は，対象がドラムまたはベルで，メトロノームの視聴覚事象の有無といった条件における，周期的，非周期的な運動に関してまとめた結果である．表中の数値は，5 つの平均時間間隔について，2 回ずつの 10 回の実験の平均値である．表中の Correlation は式 (3.12) によって求められた運動－聴覚，運動－視覚の相関を，Correspondence は



(a) Motor signal $\hat{M}_{in}$

(b) Audio signal $\hat{A}_2$

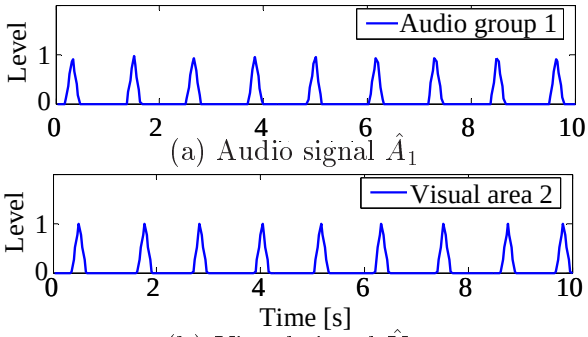
(c) Visual signal $\hat{V}_1$

図 3.10: 運動，視聴覚事象時系列（第一グループ）

Fig. 3.10 Time series of motor and audio-visual signals (group 1)

式 (3.13) によって求められた視聴覚事象間の相関である．SR（success rate）は，実験の実施回数中の成功数である．視聴覚事象の対応付けの成功率は，周期的な運動指令の場合で 75.0%，非周期的な運動指令の場合で 72.5% となり，双方の間であまり差が見られなかった．視聴覚事象の相関の平均は 0.751 となり，対応付けが 80 試行中 59 回成功し，視聴覚事象の対応付けの成功率は 73.8% であった．

成功率の目標値としては，今後の課題である概念の獲得における入力として必要な精度を目指しており，現段階では 70% 以上とした．よって，目標値を達成しており，メトロノームのような音や動きがある事象が存在する環境であっても，ドラムをたた



(a) Audio signal $\hat{A}_1$

(b) Visual signal $\hat{V}_2$.

図 3.11: 視聴覚事象時系列（第二グループ）

Fig. 3.11 Time series of audio-visual signals (group 2)

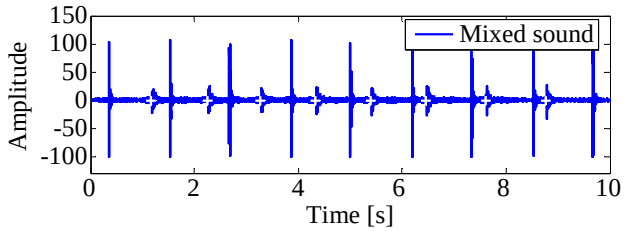
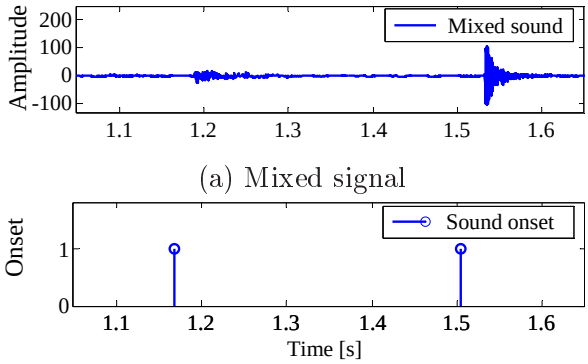


図 3.12: 物体操作に対応付けられた音オンセット
Fig. 3.12 Sound onsets corresponding to object manipulation



(a) Mixed signal

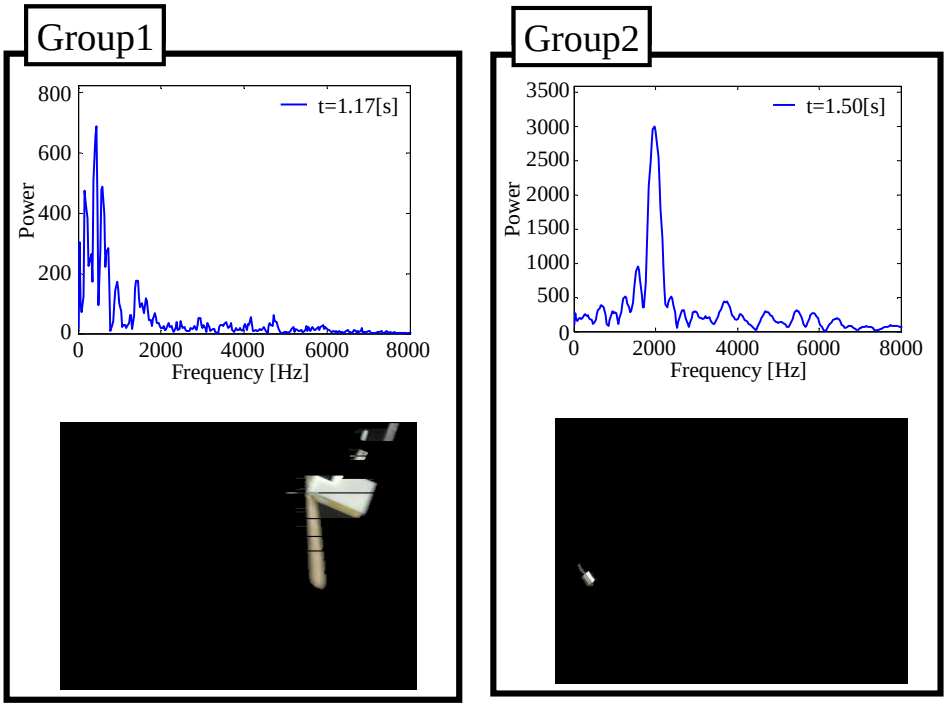
(b) Onset time series

図 3.13: 混合音と音オンセット（ドラムとメトロノーム）
Fig. 3.13 Mixed sound and sound onsets (a drum and a metronome)

く，ベルを振るといった物体操作を行うことで，発生する視聴覚事象の対応付けができることを確認した．目標値 70% の根拠は，以下に示すように対応付け条件が厳しいからである．本手法では，自らの運動でドラム，ベルといった対象物に能動的に働きかけ，それらを区別して，事前知識なしで対応付ける．対応付けの成功は，次の 3 条件をすべて満たすときである．(1) 視聴覚事象の対応が正確である，(2) 運動に関する視聴覚事象の相関が最大，(3) 運動方向変化の位置と視聴覚事象の発生位置の対応が正確である．

また，運動－聴覚間の相関は 0.802，運動－視覚間の相関は 0.745 であった．これらから，視覚情報が得られにくい暗い場合や，聴覚情報が得られにくい雑音環境の場合においても，視聴覚どちらかの情報が得られれば運動との対応付けができると考えられる．そのような場合での実験による検証は今後の課題である．

表 3.7 は，ドラムをたたいた場合，ベルを振った場合について，運動の平均時間間隔に関して，運動-聴覚，運動-視覚，視覚-聴覚間の相関と運動に関する視聴覚事象間



(a) Spectrum of the sound onset and image belonging to Group 1 which relate to object manipulation

(b) Spectrum of the sound onset and image belonging to Group 2

図 3.14: 対応付けられた視聴覚事象（ドラムとメトロノーム）
Fig. 3.14 Correspondence of two movements and two sounds (a drum and a metronome)

の対応付け成功率をまとめた結果である．メトロノームの視聴覚事象の有無といった条件での周期的，非周期的な運動の 4 つの場合についての 2 回ずつ，計 8 回の実験の平均値である．ドラム，ベル共に運動時間間隔に関して視聴覚事象の対応付けの成功回数に差が見られないことから，運動の時間間隔は，対応付け結果に及ぼす影響が少ないことを確認した．視聴覚事象の対応付けの成功率は対象がドラムの場合に 75.0%，ベルの場合に 72.5% であった．これから，対象をたたく片側での運動方向の変化と対象を振る両側での運動方向の変化とでは，発生した視聴覚事象の対応付けの成功率に差が少ないと考えられる．

視聴覚事象の対応付けが成功率 73.8% と難しくしている要因として，本手法は「右，左，両側における運動方向変化」によって音が発生するという事前知識を用いないことがある．この事前知識を用いた場合，視聴覚事象の対応付けは全体で 67 回成功し，

表 3.6: 対象物，周期性，他事象に関する運動-視聴覚事象間の対応付け
Table 3.6 Correspondence between Audio-Visual events relating to objects, periodicity and another event

Objects	Periodicity	Correlation		Correlation	
		(M, A)	(M, V)	(A, V)	SR
Drum	p.	0.825	0.602	0.659	7/10
	np.	0.842	0.751	0.839	8/10
Drum + Met.	p.	0.805	0.588	0.617	6/10
	np.	0.800	0.669	0.713	9/10
Bell	p.	0.793	0.879	0.785	9/10
	np.	0.823	0.822	0.774	7/10
Bell + Met.	p.	0.712	0.800	0.784	8/10
	np.	0.839	0.820	0.823	5/10
Periodical motion		0.781	0.735	0.722	75.0%
Non-periodical motion		0.824	0.755	0.781	72.5%
Total		0.802	0.745	0.751	73.8%

成功率は 83.8% である．また「右，左，両側における運動方向変化」の区別をしない場合，成功条件は対応付け条件 (1)，(2) であり，(3) を考慮しない．このとき視聴覚事象の対応付けは全体で 71 回成功し，成功率は 88.8% である．陳らの手法 [11] は，視覚事象をフレーム間差分画像の閾値を超えた画素数の変化のみによって抽出しているため,「右，左，両側における運動方向変化」の区別をしておらず，本研究と同じデータを適用した結果，視聴覚事象の対応付けは 80 試行中 48 回成功し，成功率は 60.0% である．以上の結果から，成功率 73.8% の本手法の有効性を示すことができた．なお今回の実験において SS を用いない場合の成功率は 71.3% であり，わずかに減少した．雑音の程度を変えての詳細な検証は今後の課題である．

以上のように,“ 同時性 ”,“ 類似性 ”を手掛かりとして，運動と視聴覚事象の対応付けを行い，その有効性を確認した．しかし，同時性と類似性は，信号の発生時刻における誤差の影響を受けやすい．また，統合処理で相互相関関数を用いて誤差の影響を低減させても対応付けができない場合があったが，この中のいくつかは，ロボットの手先位置，視覚での位置，音源方位といった空間情報を用いることで解決できると考えられる．

3.6　おわりに

本章では，対象物に働きかけ，物体の視聴覚事象を能動的に対応付ける方法について提案した．本手法は，視聴覚事象の対応付けに物体固有の情報ではなく，一般的な

表 3.7: 運動時間間隔に関する運動-視聴覚間の対応付け
Table 3.7 Correspondence between Audio-Visual relating to motion interval

Motion		Correlation		Correspondence	
Objects	Interval Ave [s]	(M,A)	(M,V)	(A,V)	SR
Drum/ Drum + Met.	0.518	0.847	0.657	0.715	6/8
	0.586	0.802	0.704	0.768	7/8
	0.653	0.829	0.666	0.704	8/8
	0.788	0.845	0.665	0.773	4/8
	0.923	0.767	0.584	0.610	5/8
Bell/ Bell + Met.	0.923	0.764	0.817	0.735	5/8
	1.058	0.775	0.867	0.720	5/8
	1.193	0.761	0.814	0.830	7/8
	1.328	0.756	0.822	0.788	6/8
	1.463	0.871	0.853	0.843	6/8
Drum/Drum+Met.		0.818	0.659	0.715	75.0%
Bell/Bell+Met.		0.786	0.833	0.789	72.5%

法則（ゲシュタルトの群化の法則）を用いる．物体を操作しようとする脳から筋への信号の変化と，そのとき，視聴覚によって観測される音の変化および映像の変化の“ 同時性 ”，物体操作と視聴覚事象の間における繰り返しの“ 類似性 ”，を手掛かりとして対応付ける．実験では，マニピュレータにより物体を振り，マイクロフォンとカメラを用いて観測し，運動と視聴覚情報の対応付けを行った．

視聴覚情報の対応付けにおける成功率の目標値は，対応付け条件が厳しいため，70% 以上とした．対応付けの成功は，次の 3 条件をすべて満たすときである．(1) 視聴覚事象の対応が正確である，(2) 運動に関する視聴覚事象の相関が最大，(3) 運動方向変化の位置と視聴覚事象の発生位置の対応が正確である．実験の結果，運動情報を用いた視聴覚事象の対応付けの成功率は 73.8% と目標値を達成しており，本手法の有効性を示す．よって，メトロノームのような音や動きがある事象が存在する環境であっても，ドラムをたたく，ベルを振るといった物体操作を行うことで，運動に関する視聴覚事象の対応付けができることを確認した．

今後の課題は，空間情報を利用し，視聴覚事象の対応付け精度と頑健さの向上，および視聴覚事象の対応付け結果からの概念の獲得である．

第4章　選択的注意による視聴覚事象の対応付け

4.1　はじめに

視聴覚情報の統合として，物体についての固有な知識を用いない，物体の運動で生じた視聴覚事象の対応付けが行われている [11]．また，音と画像を統合する研究の応用として，人とコミュニケーションを行うロボットが考えられており，K. Nakadai ら [13] は，三人の話者の音声を分離し，顔写真と統合する実験について報告した．

3 章では，複数の視聴覚事象が存在する環境において，ゲシュタルト心理学の群化の法則をもとに，物体操作による視聴覚事象の対応付けを行った [56–59]．しかし，これらの実験は，画像では周囲の輝度に変化が少なく，音でも雑音がわずかな環境で行われており，照明条件や雑音についての考察はなされていない．また，この手法は物体操作により能動的に視聴覚事象を発生させているが，視聴覚に関しては受動的に観測される事象の対応付けであった．

人間は，目的に応じて視覚と聴覚を組合せ，受動的に観測するだけでなく，対象へ注意を向けることで周囲の環境から必要な情報を選択的に取得し，処理を行っている．G. Meyer ら [60] は，複数の聴覚事象が同時に混在する空間において，唇の動きの視覚情報を用いて，特定の聴覚事象を選択的に聞き取るシステムを作成した．また視覚事象に注意を向け，外界から必要な情報を選択するときには，対象を視力の高い視野の中心部で見えるように，頭部運動と眼球運動を組合せて視線を移動させている [4,6,7]．このように，対象へ選択的に注意を向けることで，特定の情報を効率的に取得している．

注意を工学的にモデリングする研究は，近年活発に行われている．M. Jagersand ら [61] は，画像上での顕著な特徴を有する領域を検出した．ロボットに注意機構を構築することにより，注意を向けたり，視覚において追跡を行う研究も行われている [62–64]．しかし，画像上で人物の位置を検出することにとどまり，撮影後の視聴覚事象の統合は行われていない．

本章の注意のアルゴリズムは，生物の注意の現象あるいは機序と直接的な対応関係はないが，生物の注意の中の選択的注意（多くの情報の中から情報を取捨選択する機

能）に示唆を得た選択的注意を提案する．本研究では，視覚において動きを知覚した場合に，その動きがある時点に聴覚の注意を集中し，雑音環境において目的音を検出することを聴覚的注意とする．また，聴覚において音が知覚された場合に，音源方向へ頭部を回転させることや，視覚の感度を調整することにより，音に対応する運動を探すことを視覚的注意とする．

従来の異種情報間の制御や統合は，感覚ごとに得られる完全なデータを統合しているのに対して，本章で提案する選択的注意は，視覚（聴覚）情報が良好に検出できるが，聴覚（視覚）情報が検出しにくい場合において，良好な視覚（聴覚）情報を手掛かりに，検出しにくい聴覚（視覚）情報を探すことである．すなわち，一方の感覚での完全なデータを手掛かりに，他方の感覚での不完全なデータ中の情報を検出することである．

本章の構成は次のとおりである．まず **4.2** では，視聴覚の選択的注意について説明する．**4.3** においては，視聴覚における注意と対応付け処理について述べる．**4.4** では，実験方法を示す．**4.5** では，実験結果を示すとともに考察を行い，**4.6** で本章をまとめる．

4.2　聴覚的注意と視覚的注意

人は，関心のある事象について注意を向け，空間の特定の領域や特定の時間に意識を集中させることで情報処理を効率的に行っている．また人間の処理能力には限界があり，目的に応じた行動を行うためには，感覚器からの大量の情報を瞬時に処理しなければならない [65]．

注意とは，心的活動における選択的集中の現象である [81]．視覚的注意は，特徴の場所に対する空間的注意と外界からの情報を得るための処理に対する注意であり，聴覚的注意は，空間，時間，周波数帯域に対する注意である．そのような選択的注意の一例として本章では，まず，聴覚的注意として，次のものを取り上げる．視覚において動きを良好に知覚した場合に，その動きがある時点に聴覚の注意を集中し，雑音環境において物体の運動によって発生した目的音を検出することである．

次に，視覚的注意として，次のような機能を取り上げる．聴覚において音が知覚されるが，視覚情報が得られない場合に，視線方向や感度の調整により，音に対応する運動を探す．それでも視覚情報が得られない場合は，視野外に対象物があると判断し，音源定位により推定される方向へ頭部を回転させ，先程と同様に音に対応する運動を探す．また，周辺が明るく，中央部が暗い場所に視聴覚事象が発生しているような局

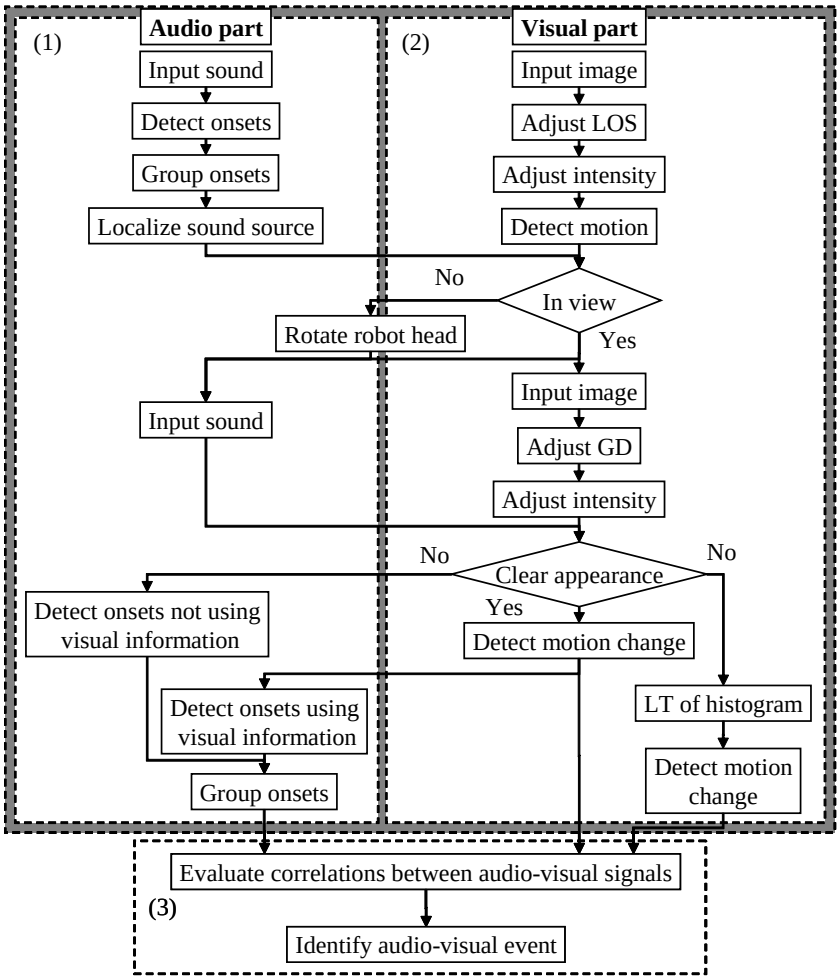


図 4.1: 視聴覚事象の注意処理
Fig. 4.1 Attentional process of audio-visual events

所的に輝度の異なる場面において，画像全体の輝度に応じて絞りを調節すると暗い部分にある物体を検出できない．そのため，暗い部分にあわせて輝度の調節を行う必要がある．そのため暗い部分に注視するように視線方向を調節し，画像の輝度を注視領域に合わせて調節することを視覚的注意とする．

4.3 注意による対応付け処理

本章のシステムは，2 つのマイクロフォンと 1 つの可動カメラ，およびそれらを備えた左右方向に回転可能なロボットヘッドから構成される．全体の処理の流れを**図 4.1** に示す．(1) 聴覚部，(2) 視覚部，(3) 統合部からなり，音入力に対して，音オンセット

を検出し，音源位置を推定する．画像入力に対しては，カメラの視線方向と絞り調節を行う．次に，運動物体が視野内に存在するか判断し，存在する場合はロボットヘッドの回転を行わず，存在しない場合は，音源の推定方向へヘッドを回転する．目的物に向いた後，音と映像を再び計測し視覚部では，絞り調節後の注視領域の輝度値の平均により，画像の輝度が良好であるか判断をする．明るさが良好であれば，運動方向変化の時刻を検出し，聴覚部における音オンセット検出時に利用する．そうでなければ，輝度ヒストグラムの線形変換を行う．画像の輝度が良好な場合，聴覚部では視覚情報を用い，周波数領域において時間軸上での窓関数を作成し対象音の検出を行う．画像の輝度が良好でない場合，視覚情報による時間軸上での窓関数を利用しない．視聴覚部からの運動の変化，音の発生時刻を統合部にて評価し対応付ける．

4.3.1　聴覚的注意による目的音検出

聴覚部では，2 つのマイクロフォンを用いて音信号を計測し，音源定位することで視野外の音を発生させている対象物の位置を推定する．その概要は次のとおりである（図 4.1 の (1) 参照）．まず，音源の音オンセット時系列を求める．検出した音オンセット時刻を手掛かりとして，計測した音信号のマイクロフォン間到達時間差から音源の定位を行う．また視野内に音を発生させている対象物があり，かつ視覚情報が良好に抽出できる場合，視覚情報を用いて **4.3.1(b)** に述べる手法により音オンセットを検出する．本研究では，視野内に対象物体がないときの音源定位や，画像が暗いときといった視覚情報を用いない場合に，**3.3.1(b)** で述べた音オンセット検出法を用いる．

(a)　音源定位

周波数領域で検出した音オンセットを，音オンセットのスペクトル間の相関の高い音源ごとにグループ化する．繰り返し発生している音に対して定位を行うため，3 回以上音オンセットを検出した音源グループの中で，スペクトル間の相関の平均が最高の音源を定位する．音源定位には，P. Aarabi ら [66] の手法のように，雑音環境に頑強なものもあるが，ここでは音が良好であるとして **2.2** に述べた手法を用いる．このとき，音源グループ内の各音オンセットの定位角の平均を対象音源の定位角とする．

(b) 視覚情報を利用した音オンセット検出

視野内に運動物体が存在し画像の輝度が良好な場合，視覚情報を手掛かりにして対象音の音オンセット検出を行う．雑音環境下における対象音の音オンセット検出のために，視覚における物体の運動方向が変化した時刻を中心に，三角窓の時間軸上での窓関数（高さは 1，底辺の長さは運動方向の変化の平均時間間隔に係数 ($\alpha = 1$) を掛けたもの）を形成する．そして，フーリエ変換 (サンプリング 512 点，シフト量 256 点) を行い，0 [Hz] から重なりなく 1 [kHz] ずつ 8 つの周波数帯域に分けた各周波数帯域において，時間軸上での窓関数を入力音に対して乗じる．これにより，**3.3.1(b)** の音オンセット検出法では難しい，視覚事象に関係ない雑音の影響を低減することができる．

また，雑音環境下での対象音検出のために，スペクトルサブトラクション (SS) を行い [67]，混合音から推定した雑音を周波数領域において除去する．雑音は，映像中の物体が動き出す直前の 10 フレーム (1 フレーム 512 点) の平均から推定する．映像中の物体の動き検出は **4.3.2(c)** において述べる．

なお，実験では，フーリエ変換のフレーム長を 512 点，シフト量をフレーム長の半分とし，窓関数にハミング窓を用いる．周波数帯域ごとの計測時間内の音パワーの平均値 P_{ave} と最大値 P_{max} から次式のように音オフセット検出閾値 P_{off} と，音オンセット検出閾値 P_{on} を決定する．

$$\left.\begin{array}{l} P_{off} = (P_{max} - P_{ave}) \times 0.2 + P_{ave} \\ P_{on} = (P_{max} - P_{ave}) \times 0.3 + P_{ave} \end{array}\right\} \tag{4.1}$$

4.3.2 視覚的注意による対象物検出

(a) 視線方向調節による対象物定位

音源定位により，発生した音事象の方向は推定されるが，空間分解能が低いため，誤差を含んだ定位となる．しかし，局所的に輝度が異なる場面では，対象物が存在すると推定される領域に注意を向け，輝度の調節を行い，対象物を検出しなければならない．このとき，より高い空間分解能を持つ画像情報を用いて図 4.1(2) に示すように視線方向 (GD : gaze direction) の調節を行う必要がある．

本実験では，周辺部が明るく，中央部が薄暗い穴の中などに対象物がある環境の場合，視線方向を調節し，輝度値が小さな領域が画像の中央部に撮影されるよう，カメラをパンする．このため輝度が n [bits]，2^n 階調，画像サイズ $(w \times h)$ の時，輝度閾値 $(th_I = 2^{n-3})$ 以下となる画素数の総和が閾値 $(th_d = (w/4) \times (h/3))$ 以上の場合，それ

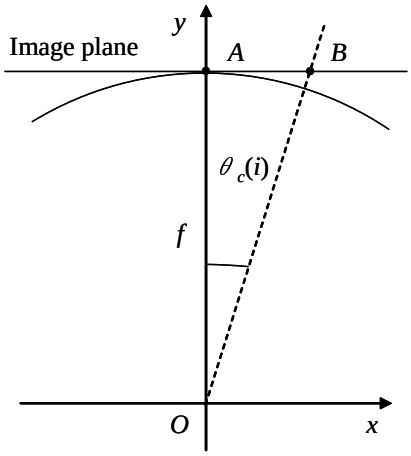


図 4.2: 視線方向の調節
Fig. 4.2 Adjustment of gaze direction

らの画素の x 座標の平均 x_c が画面の x 方向の中心 $(d_c = w/2)$ になるようカメラのパン角を調節する．今回は水平方向のみの視線方向の調節を行うが，垂直方向の視線方向調節も同様に拡張できる．

画面上における画素とカメラの回転角度との関係は次のように求める．**図 4.2** に示すように，原点 O から焦点距離 f だけ離れた画面の中心点を $A(d_c, f)$ とすると，i 回目に測定した画面上における任意の画素点 $B(x_c(i), f)$ を，点 A に移動するために回転するカメラの回転角度 $\theta_c(i)$ は次式で表される．

$$\theta_c(i) = \tan^{-1}(\frac{\epsilon_x}{f}(x_c(i) - d_c)) \tag{4.2}$$

ここで，ϵ_x は補正定数である．A. Bahill ら [68] により，人が眼球運動だけで視標をとらえるのは，せいぜい 15[deg] までであると報告されている．また，音源定位した対象以外に視線方向を向けてしまうことを防ぐため，本実験ではカメラの視線方向の調節を -15〜15[deg] の範囲で行う．そのため，カメラの内部パラメータ ϵ_x/f を求めるとき，カメラ校正を次のように行う．目印となる黒点を中心にカメラを左右方向へ -15〜15[deg] の範囲で 1[deg] ずつ回転させながら画像を撮影し，画像上の黒点の x 座標 $x_c(i)$ と回転角度 $\theta_c(i)$ の関係から ϵ_x/f を最小 2 乗法により算出する．

カメラの実際の調節角度 $\theta_v(i)$ は次式で示すように，$i-1$ 回目までの視線方向の調節指令角度 $\theta_c(i)$ の和となる．

$$\theta_v(i) = \sum_{k=1}^{i-1} \theta_c(k) \tag{4.3}$$

ただし，本実験では物体の運動によって $x_c(i)$ の値が細かく変化するのにしたがって，カメラが変動することを防ぐため，i 回目の調節指令角度 $\theta_c(i)$ の絶対値が 2[deg] 未満の場合はカメラを回転せず，式 (4.3) において調節指令角度 $\theta_c(i)$ を加算しない．$i+1$ 回目に推定される実際の調節角度 $\theta_v(i)$ の絶対値が 15[deg] を超える場合も，カメラを回転せず，調節指令角度 $\theta_c(i)$ を加算しない $\theta_v(i)$ のままとする．

(b) 絞り調節による対象物検出

図 4.1(2) に示す視覚的注意処理を行う．画像の中央部に注視領域を設け，この領域における平均輝度およびエントロピーの計算を行う．竹内は，画像の輝度の特徴を定量的に表すためシャノンの情報量を用いた [69]．本章では，エントロピーを尺度として輝度ヒストグラムの平坦性を評価する．エントロピーは次式で求める [70]．

$$H = -\sum_{k=1}^{K} p(k)\log_2 p(k) \tag{4.4}$$

ここで，$p(k)$ は輝度値 k の画素の全画素に対する割合である．画像の階調数を K とすると，エントロピー H は 0 から $\log_2 K$ までの範囲の値をとり，画像全体が 1 つの輝度値をとる場合，最小となり，画像にすべての輝度値が同じ割合で現れる場合，最大となる．輝度ヒストグラムが平坦になり，より明確に対象物を抽出できるようにカメラの絞りを適応的に変化させる．

カメラの絞りを以下の処理により調節する．輝度が 2^n 階調（n [bits]），画像サイズ $(w \times h)$ の時，注視領域は画像の中央部 $((w/4) \times (h/3))$ とする．注視領域内におけるエントロピーを H，平均輝度を I とする．ここで，入力のカラー画像から輝度画像への変換には，**3.3.2(a)** の式 (3.2) を用いる．I の値に対して絞り調節 cnt (control) を次式のように行う．

$$cnt = \begin{cases} 1 & : \quad I < 2^{n-1} \\ 0 & : \quad I = 2^{n-1} \\ -1 & : \quad I > 2^{n-1} \end{cases} \tag{4.5}$$

$cnt = 1$ のとき，画像全体の輝度値が 3[dB] 上昇するよう絞りを開ける．$cnt = 0$ のとき，絞り調節を行わない．$cnt = -1$ のとき，画像全体の輝度値が 3[dB] 低下するよう絞りを閉じる．次の**条件** 1 を満たさないとき，cnt の値により絞り調節を行う．

条件 1：$|\frac{I(i)-2^{n-1}}{2^{n-1}}| < th_I$

上記の条件を満たし，かつ下記に示す**条件** 2〜6 のいずれかにあてはまる時，次の 1 回は絞り調節を行わない．

条件 2：$H(i) \geq n(1 - th_H)$
条件 3：$|\frac{I(i)-I(i-1)}{I(i)}| < th_I$
条件 4：$|\frac{H(i)-H(i-1)}{H(i)}| < th_H$
条件 5：過去 3 フレームでの $|H-n|$ の値の昇順 (1 位：1，2 位：0，3 位：-1）と $|I-2^{n-1}|$ の値の昇順 (1 位：1，2 位：0，3 位：-1）の和で 2 番目が最小
条件 6：過去 3 フレームでの cnt が $(1, -1, 1)$ または，$(-1, 1, -1)$
また，最大限明るく，または暗く，絞り調節をした後は，それよりも明るく，または暗くする絞り調節は行わない．これにより，エントロピー H を大きくし，周囲の環境の輝度が異なる状況でも適応的に対象を検出する．なお，i 回目に計測した平均輝度を $I(i)$，エントロピーを $H(i)$ とし，輝度ヒストグラム平均閾値 $th_I = 0.2$，エントロピー閾値 $th_H = 0.2$ は予備実験により求めた．

(c)　視野内の運動物体の検出

視野内に運動物体が存在するかを判断するため，まず連続した 3 フレームの間の差分領域の画素数の変化を求め [59]，運動物体の検出を行う．計測時間内の差分画素数の最大値が，運動検出閾値 $(th_m = (w/6) \times (h/4))$ 以上となったとき，計測時間内の差分画素数の平均値 th_d から，運動物体を検出するオンセット閾値 $(D_{on} = 1.5 \times th_d)$ とオフセット閾値 $(D_{off} = 1.1 \times th_d)$ を決定する．差分画素数が D_{off} 未満から D_{on} 以上となった時刻を物体の運動時刻とする．

3 回以上物体の運動が検出された場合，視野内に運動物体が存在すると判断し，音源定位結果に関わらずロボットヘッドを回転しない．運動物体が存在しない場合，運動物体は視野外にあるとして，音源定位推定値に基づいてロボットヘッドを回転する．

(d)　輝度ヒストグラムの線形変換

注視領域の画像の輝度が良好であるかを平均輝度で判断する．注視領域の絞り調節後の平均輝度が，n [bits]，2^n 階調において $2^{n-1} \pm 2^{n-3}$ の範囲内であるとき，輝度は良好であるとする．輝度が良好であれば，運動方向変化の時刻を検出し，聴覚部における音オンセット検出時に利用する．そうでなければ，注視領域内における輝度値ごとの画素数の頻度である輝度ヒストグラムを線形変換 (LT : linear transformation) する．

線形変換を行うことで輝度ヒストグラムの分布を広げ，差分画像が得られやすくする．輝度が 2^n 階調の時，変換前の輝度値を $I_B(i, j)$，階調値の最小，最大を f_{min}，f_{max}

とすると，変換後の輝度値 $I_A(i,j)$ は次式で表される．

$$I_A(i,j) = 2^n \times \frac{I_B(i,j) - f_{min}}{f_{max} - f_{min}} \tag{4.6}$$

ここで，f_{min}，f_{max} は，輝度ヒストグラムがヒストグラム閾値 th_f 以上となる最小と最大の輝度値である．ヒストグラム閾値 th_f は，変換時にヒストグラムを伸張し過ぎて情報量を減らすことを防ぐために，変換前の輝度ヒストグラムの最頻値 h_{max} を用いて，$th_f = 0.02 \times h_{max}$ とする．

(e) 運動方向変化の検出

3 章では，運動発生の検出を運動物体の軌跡の運動方向変化の時刻から求める方法を提案した [59]．本章ではこの手法を用い，物体の運動領域を求めるときの差分画像の閾値は予備実験から求めた．ここで，対象物をたたいて事象を発生させる場合では，対象物が存在する場所における片側の運動方向変化の時系列を抽出し，対象物を振って事象を発生させる場合では，運動方向の変化ごとに音が鳴るとし，両側での運動方向の変化ごとに時系列を検出することとした．

視覚部では，運動の変化する時系列を両側，右側，左側の 3 通りについて求め，統合部において，聴覚事象と対応付けられる最適な運動方向変化の時系列を求める．

4.3.3 視聴覚事象の対応付け

以上の処理により得られた時系列，すなわち音オンセット時系列 $A_k(t)$ と映像における運動方向変化の時系列 $V_l(i)$ を“ 同時性 ”と繰り返しの“ 類似性 ”に着目し，対応付ける [71]．

聴覚と視覚のサンプリング周波数は，それぞれ 16 [kHz]（周波数領域における音オンセット検出時は 62.5 [Hz]）と 30[Hz] で異なるので，統一的に評価するため，聴覚のサンプリング周波数をより小さい視覚のサンプリング周波数にそろえる．時系列 $A_k(t)$，$V_l(i)$ に対し，各信号時点を中心に最大値が 1，底辺の長さが運動方向の変化から求めた平均時間間隔に係数 ($\beta = 0.4$) を掛けた時間になるように，三角波の重みを付ける．そして，0 秒から 1/30 秒ごとの三角波の値により，$\hat{A}_k(i)$，$\hat{V}_l(i)$ を得る．

これら視聴覚の時系列を“ 同時性 ”と“ 類似性 ”を手掛かりとして対応付けるため，相関関数を評価尺度として用いる．計測誤差，および信号抽出処理における誤差の影響を低減するため，視聴覚情報間の相関係数を時間シフトを考慮して求める．そのた

め，次式のように視聴覚信号間の相互相関関数の最大値を求める．

$$Corr(A_k, V_l) = \max_{-4 \leq s \leq 4} \frac{Cov(\hat{A}_k(i), \hat{V}_l(i+s))}{\sqrt{Var(\hat{A}_k) Var(\hat{V}_l)}} \tag{4.7}$$

ここで，i は視聴覚事象が発生している区間，s はシフト量であり，$-4 \leq s \leq 4$ (時間に換算すると ± 0.133 秒) とし，相互相関関数の最大値を算出する．視聴覚事象の相関に対して，相関が 0.5 以上かつ最大となる視聴覚事象を対応付ける．対応付けの成功は，次の 3 条件をすべて満たすときである．(1) 視聴覚事象の対応が正確，(2) 視聴覚事象の相関が 0.5 以上，(3) 運動方向変化の位置と視聴覚事象の発生位置の対応が正確である [59]．

なお, 聴覚的注意（時間軸上での窓関数）により, 両側で視聴覚事象が発生しているにも関わらず, 左右片側ずつで検出された視聴覚事象の相関がより高くなる場合がある．そのため, 両側, 右側, 左側の時間軸上での窓関数で検出された各々の最初の音オンセットにおけるスペクトル間の相関を求め, 相関係数の最小値が 0.5 以上で，かつ両側，右側，左側の各々で検出された視聴覚事象間の相関係数の最小値が 0.7 以上のとき, 両側で発生した視聴覚事象とする．

4.4　視聴覚的注意実験

実験では，実際に起こりうるケースとして，以下の 3 通りの状況を想定する．第一に，視覚情報が良好に検出できるが，聴覚情報が検出しにくい場合，第二に，聴覚情報が良好に検出できるが，視覚情報が検出しにくい場合，そして，これらの比較として第三に，視覚情報と聴覚情報が共に良好に検出できる場合である．システムは，各場面に応じて判断を行い，選択的に対象へ聴覚的，視覚的注意を向け，事前知識なしで視聴覚事象の対応付けを行う．

4.4.1　実験装置

実験装置は，2 個のマイクロフォン（RION，UC-30），1 台のカメラ（SONY，EVI-G20）を備えた左右方向に回転可能なロボットヘッド，ロボットヘッドの制御のための計算機（VAIO, VGN-S70B），マイクロフォンとカメラからの信号を計測し，データを処理する計算機（Dell, PowerEdge SC420）を使用する．

4.4.2 実験条件

聴覚的注意の実験では，視覚情報により得られる物体の運動方向変化の時刻を基に作成した時間軸上での窓関数を用い，音オンセットを検出し，得られた視聴覚信号の対応付けを行う．視覚情報が良好に得られるような明るく開放された空間内において，対象物はベル，ドラム，メトロノームであり，雑音は白色雑音と女性の声である．対象物に対するロボットヘッドの回転角度 $\theta = 0$ [deg] の状態から，雑音の強度を変化させ視聴覚事象を対応付ける．

視覚的注意の実験では，測定した音源の音オンセットから音源定位により，音が発生した方向にカメラが向くようにロボットヘッドを回転し，視覚的注意により，対象物をより正確にとらえる．実験では，視野外の視聴覚事象を音源定位し，カメラ，マイクロフォンを搭載したロボットヘッドを回転させ物体にカメラを向ける．実験において，対象物は**図 4.3**(a)，(b) に示すように周辺部が明るく中央部が暗い穴の中にあると想定し，穴の形状を円形と不規則形の 2 通り用意した．対象物の正面から角度 $\theta = \{-45, -30, 0, 30, 45\}$ [deg] だけロボットヘッドを回転させた状態から視聴覚事象を対応付ける．

音はサンプリング周波数 16 [kHz]，量子化 8 [bits]，映像はサンプリングレート 30[Hz]，輝度は $n = 8$ [bits] で 256 階調，画像全体のサイズは 320×240 [pixels] であり，注視領域のサイズは 80×80 [pixels] とした．カメラ校正により式 (4.2) におけるカメラの内部パラメータ ϵ_x/f は，$\epsilon_x/f = 2.480 \times 10^{-3}$ と求められた．2 つのマイクロフォン間の距離は 0.3m であり，音速 $V_{sound} = 347$ [m/s]（気温 26°C）の環境において，音，映像共に 14 秒間計測した．視野内に運動物体が存在するかの判定や，視聴覚事象の対応付けは，カメラの視線方向と画像の輝度を調節した後半の 10 秒間のデータを用いた．物体は，たたかれているドラム，振られているベル，動作中のメトロノームである．いずれも動作と共に音が発生している．

4.5 結果と考察

4.5.1 聴覚的注意

音源定位の精度を調べるため，ドラムがたたかれている状況でロボットヘッドをドラムの方向から −60〜60 [deg] の範囲で回転させて音を測定した．検出した音源グループの中で，音オンセットのスペクトル間の相関の平均が最も高いグループの定位結果は**表 4.1** である．表には，音源定位の真値，音源グループ内の定位角度の平均を Mean

(a) Circle　　(b) Irregular

図 4.3: 対象範囲の形
Fig. 4.3 Shape of object region

表 4.1: ドラムへの音源定位の真値と推定値
Table 4.1 True and values in sound localization of drum

True value [deg]	-60	-45	-30	-20	-10
Mean [deg]	-51.5	-38.5	-25.6	-16.8	-5.7
SD [deg]	2.22	2.53	0.00	0.00	2.01
CORR	0.874	0.950	0.930	0.860	0.886

0	10	20	30	45	60
0.5	12.5	25.6	35.2	47.7	59.9
1.37	0.00	0.00	0.00	2.74	0.00
0.939	0.952	0.928	0.952	0.907	0.928

[deg]，標準偏差を SD[deg]，グループ内の基準となるオンセットと他のオンセットとの相関の平均を CORR(correlation) として示す．音源定位の推定値と真値との誤差は，2 乗平均平方根で 4.62[deg]，すべての角度で 10[deg] 未満であり，音源のおよその方向が推定できている．

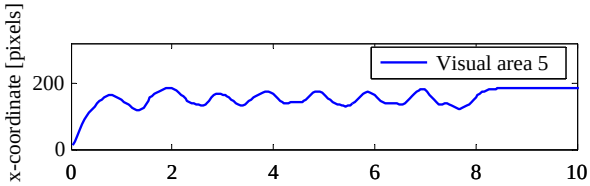
次に，視野内に対象物を検出した後，視覚情報が良好に得られるが，聴覚情報が検出しにくい場合について，聴覚的注意による目的音の検出を次のように示す．例として，視覚情報が良好に得られるような明るく開放された空間内において，ドラムがばちでたたかれていて，ロボットヘッドが対象物から回転角度で 0[deg] の位置からの聴覚的注意による視聴覚事象の対応付けについて説明する．ここで，目的音に付加した雑音は，女性の声である．視覚情報による時間軸上での窓関数を用いて，音オンセットを検出する．**図 4.4** は，視覚情報とそれを用いた音オンセット検出のためのフィルタである．(a) は，視覚から得られた運動軌跡であり音オンセット検出のための手掛かりとなる．(b) は，(a) の軌跡から求められた運動方向が変化する時刻を中心とした三角窓の時間軸上での窓関数である．運動軌跡および運動方向変化の時刻の検出は，**3.3.2**

表 4.2: 雑音を加えたときの視聴覚事象間の対応付け
Table 4.2 Correspondence between audio-visual when noise is added

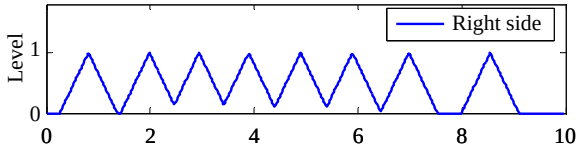
Object	Noise	Non window function				Window function			
		SNR [dB]							
		∞	0	-5	-10	∞	0	-5	-10
Bell	W	2	2	2	2	2	2	2	2
	V		2	2	2		2	2	2
Drum	W	2	1	0	1	2	1	2	2
	V		0	0	1		2	2	2
Met	W	2	2	2	2	2	2	2	2
	V		2	2	2		1	2	2
White noise		6/6	77.8% (14/18)			6/6	94.4% (17/18)		
Voice			72.2% (13/18)				94.4% (17/18)		
Total		78.6% (33/42)				95.2% (40/42)			

に述べた方法を用いる．(c) は目的音であるドラムの音である．(d) は，(c) に音声を雑音として付加して，SNR(signal to noise ratio) が -5[dB] となるようにした混合音である．(e) は，(d) の混合音に時間軸上での窓関数を通した 8 つの周波数帯域の中の 0〜1 [kHz] の周波数帯域のパワーと，このスペクトルより求めた音オンセットと音オフセットの検出閾値 P_{on}，P_{off} である．(f) は，0〜1 [kHz] の周波数帯域において検出した音オンセットである．8 つの周波数帯域ごとで音オンセットを検出し，スペクトルの類似性からグループ化した結果が (g) の音オンセット時系列となった．このように，視覚情報を手掛かりとした時間軸上での窓関数を通すことによって，良好に音オンセットが検出できていることが分かった．

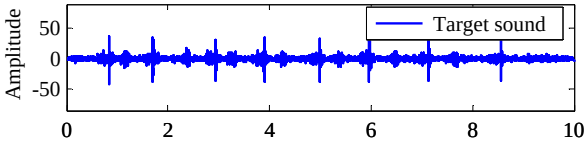
雑音の種類と強度を変えた場合の，聴覚的注意による目的音の検出を行ったときの視聴覚事象の対応付け結果を**表 4.2** に示す．対象物はベル，ドラム，メトロノームであり，雑音は白色雑音 (W) と女性の声 (V) である．SNR=∞ は目的音に雑音を加えない場合であり，SNR=0, – 5, – 10[dB] においてベル，ドラム，メトロノームの各 2 回ずつの 6 回の実験を行い，表中の数字はそれぞれ 2 回の実験中の視聴覚事象の対応付けが成功した回数を示す．成功率は，聴覚的注意を向けない（時間軸上での窓関数を用いない）場合の 78.6% から，聴覚的注意を向ける（時間軸上での窓関数を用いる）場合の 95.2% へ上昇し，本手法の有効性が確認できた．



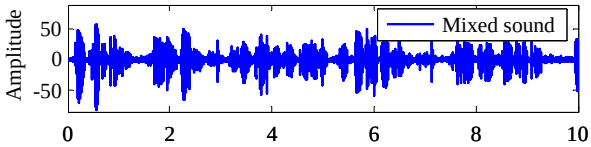
(a) Motion locus obtained by vision



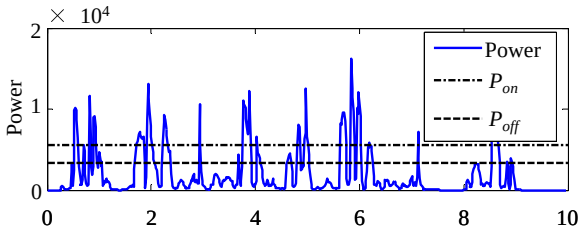
(b) Window function (Right side)



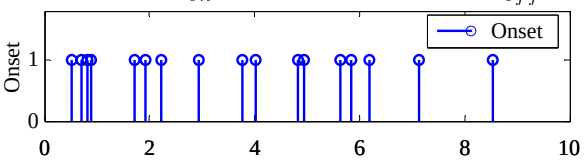
(c) Target sound



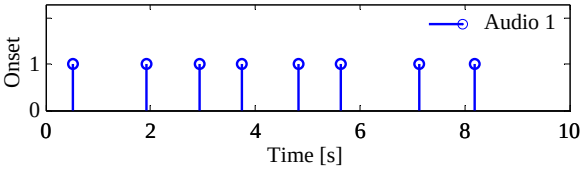
(d) Mixed sound of target sound and voice



(e) Power in frequency band of 0-1 [kHz], onset threshold P_{on} and offset threshold P_{off}



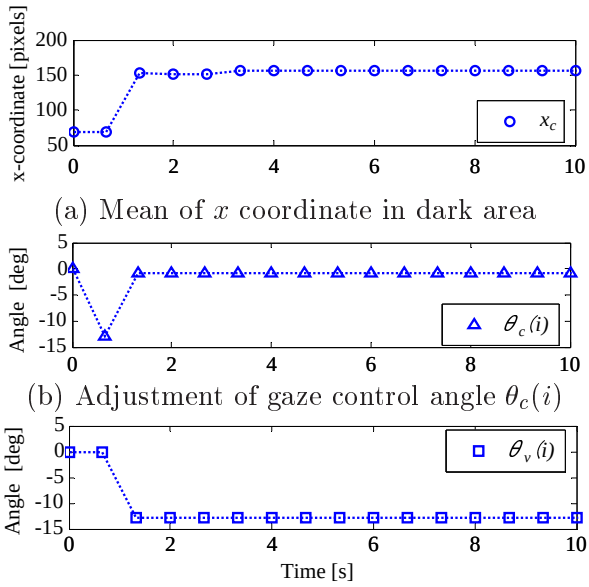
(f) Detected sound onsets in frequency band of 0-1 [kHz] by window function (Right side)



(g) Grouped sound onsets of target sound

図 4.4: 視覚情報を用いた音オンセット検出

Fig. 4.4 Onset detection using visual information



(a) Mean of x coordinate in dark area

(b) Adjustment of gaze control angle $\theta_c(i)$

(c) Change of real gaze angle $\theta_v(i)$

図 4.5: 視線方向の調節

Fig. 4.5 Adjustment of gaze direction

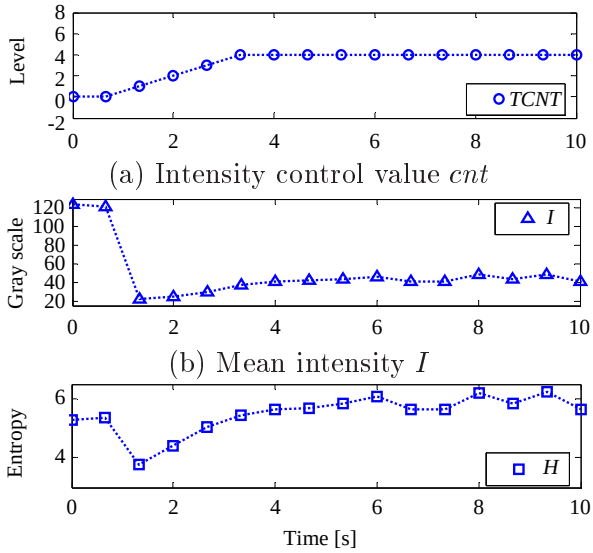
4.5.2 視覚的注意

音を発生した事象へ視覚的注意を向ける実験結果について，図 4.3(b) で示した不規則な形の穴の中で，ドラムがばちでたたかれている場合を例として説明する．

視覚的注意実験 I

ロボットヘッドは，対象物から回転角度で +30[deg] の位置にあり，注意による視聴覚事象の対応付けした結果を**図 4.5〜4.10** で示す．まず，視野内に物体があるかを判定した結果, 物体は検出されず，また音源定位角度は $\theta_a = -23.7$ [deg] で，音源グループ内のスペクトル同士の相関係数の平均は 0.885 となり，ほぼ対象物について音源定位が行えている．以下に対応付けができるまでの過程を示す．

図 4.5 はカメラの実際の調節値 $\theta_v(i)$ の時間的変化を示す．カメラ制御プログラムが起動して 0.67 秒後に (a) の暗い部分の x 座標の平均値 x_c が 69[pixels] となっており，画面の左端に映し出された暗い部分にカメラを向ける．このときの視線方向の調節指令角度は $\theta_c(i) = -12.9$[deg] であり，カメラが回転し終わるまで次の x_c の計測は行わない．1.33 秒後に暗い部分が画面中心に映し出され，暗い領域の中心は $x_c = 156$ [pixels]，調節指令角度は $\theta_c(i) = -0.9$ [deg] となる．$\theta_c(i) < 2$[deg] となるため，視線方向の調節は行わず，実際の調節角度は $\theta_v(i) = -12.9$ [deg] である．



(a) Intensity control value cnt

(b) Mean intensity I

(c) Entropy H

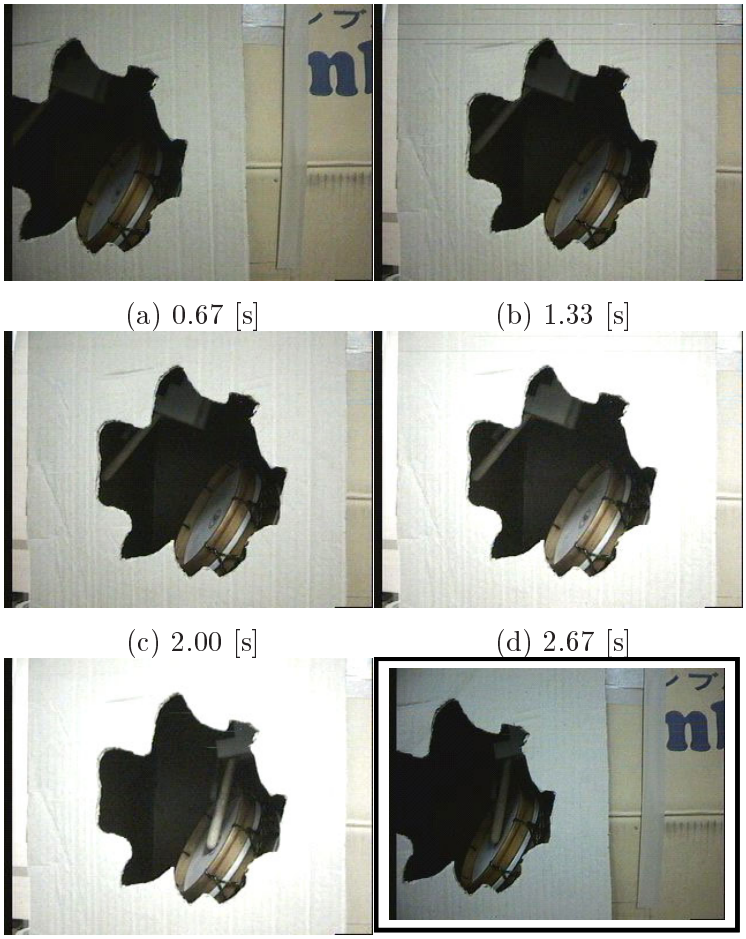
図 4.6: カメラ絞り調節

Fig. 4.6 Adjustment of intensity

図 4.6 は，カメラの絞り調節における調節値の変化とそれに基づく輝度とエントロピーの変化である．(a) は，画像の輝度の制御値 cnt の累積値 $TCNT$ (total control) を示す．絞り調整は，カメラ制御プログラムを起動してから 4 秒後には制御値が $cnt = 4$ となり，終了している．これは絞り調節を最大限に明るくした結果である．(b) は，平均輝度 I であり，0.67 秒まで明るい部分を画面中心にしているためで $I = 122$ と高い値となるが，1.33 秒までに視線方向を対象物がある穴部分に向けると $I = 22$ まで下がる．4 秒以降はわずかに上昇するが，暗い背景部分の影響により輝度 $I = 43$ と低い値となる．(c) は，エントロピー H の時間的変化を示し，1.33 秒後の $H = 3.76$ から増加して $H = 5.68$ となり，目標値である 8 に近づいている．

図 4.7(a)〜(e) は視線方向と画像の輝度の調整を行っているときの各時刻における映像である．カメラ制御プログラムを起動してから 0.67 秒後の (a) においては暗い部分が中央に来ていないが，1.33 秒後の (b) では画像のほぼ中心に調節されている．しかし，対象物は暗く検出されない．画像の中央部の輝度の調整を行うことによって 3.33 秒後の (e) では対象物が検出される．(f) は (a) と同じ状態から視線方向の調節をせず，画像の輝度の調節のみを行った 3.33 秒後の画像であり，(e) と比較して対象物が暗く，運動を検出するのは難しい．

表 4.3 は図 4.7(a)〜(e) に対応した視線方向の調節角度 $\theta_v(i)$，注視領域の平均輝度



(a) 0.67 [s] (b) 1.33 [s]

(c) 2.00 [s] (d) 2.67 [s]

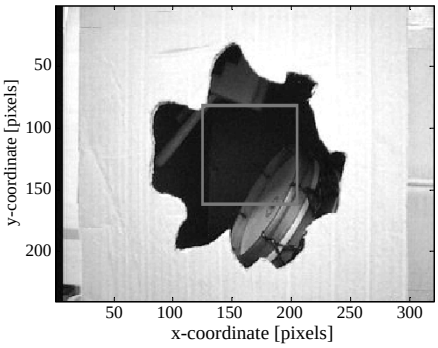
(e) 3.33 [s] (f) 3.33 [s]

図 4.7: 視線方向と画像の輝度の調節
Fig. 4.7 Adjustment of gaze direction and intensity

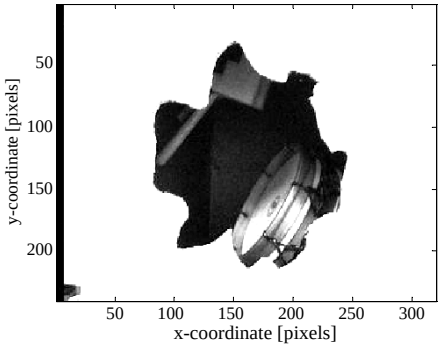
I とエントロピー H の値を示す．調節指令角度 $\theta_c(i)$ は，0.67 秒後の -12.9 [deg] から 1.33 秒後に -0.9 [deg] となり，0[deg] に近づいていることから，視線方向の調節ができていることがわかる．1.33 秒後の時刻において平均輝度は $I = 22$ と低い値であるが，時間とともに増加し，3.33 秒後では $I = 37$ となっている．エントロピー H も 1.33 秒後の $H = 3.76$ から 3.33 秒後の $H = 5.45$ まで増加する．背景部分の輝度が低いため注視領域の平均輝度 I は小さいが，エントロピーの増加からもカメラの絞り調節により，対象物についての情報量が増加していることがわかる．

表 4.3: 視線方向の調節角度，平均輝度とエントロピー
Table 4.3 Adjustment angle in line of sight, mean intensity and entropy

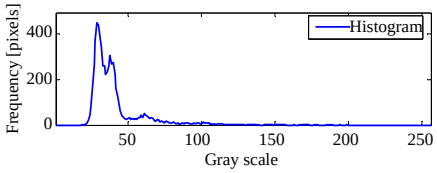
Time [s]	0.67	1.33	2.00	2.67	3.33
$\theta_c(i)$	-12.9	-0.9	-0.9	-0.9	-0.9
I	122	22	25	30	37
H	5.36	3.76	4.41	5.05	5.45



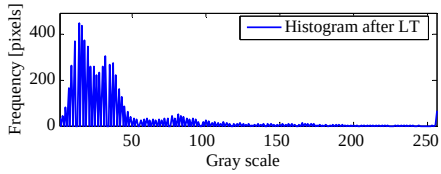
(a) Input image



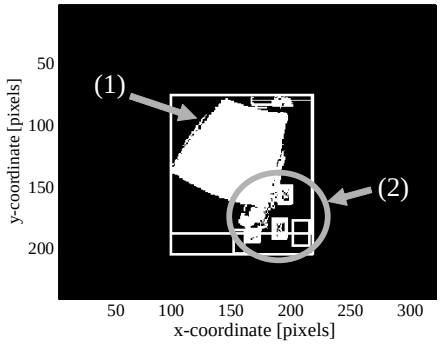
(b) Image after LT



(c) Histogram of the input image

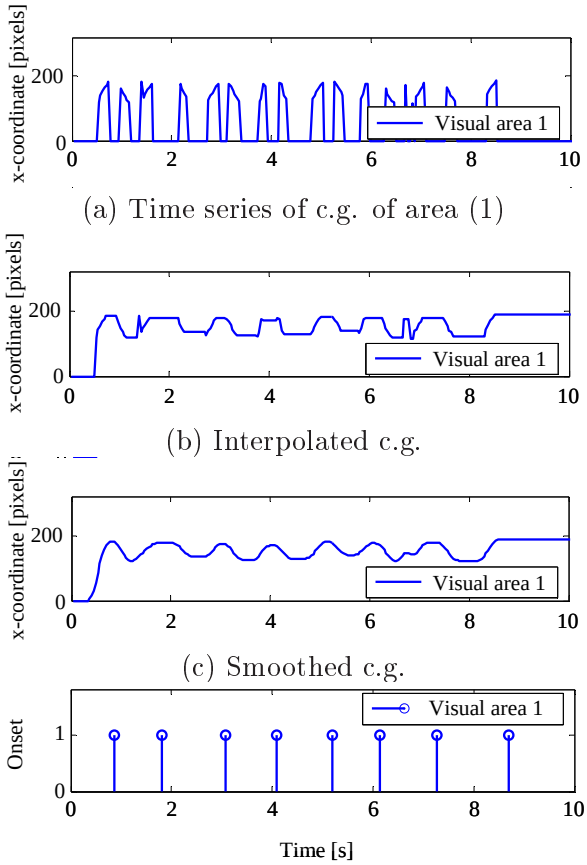


(d) Histogram after LT



(e) Motion area of the object

図 4.8: 検出した対象物
Fig. 4.8 Extracted object



(a) Time series of c.g. of area (1)

(b) Interpolated c.g.

(c) Smoothed c.g.

(d) Motion direction change time series at the right side

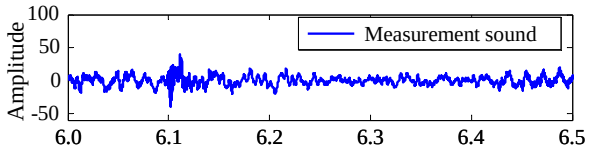
図 4.9: 運動範囲 (1) における運動方向の変化
Fig. 4.9 Motion direction change of area (1)

画像の輝度の判断において，図 4.6(b) のように平均輝度の絞り調節後における平均 $I = 43$ は，128 ± 32 の範囲に入っていないため，輝度ヒストグラムの線形変換をすることによって視覚情報を得られやすくなる．輝度ヒストグラムの線形変換を行うことによって，**図 4.8** に示す (a) の原画像から (b) のように線形変換され，それに対応して輝度分布は (c) から (d) へ広がり，対象物がより明確に見られるようになる．図 4.8(a) 中に，注視領域を中央の実線で示す．(b) の画像列を用いて (e) に示す運動範囲が検出される．図 (e) 中の (1) は対象物体の運動範囲であり，(2) はドラムをたたいたとき，ドラムがわずかに動いた結果のノイズである．

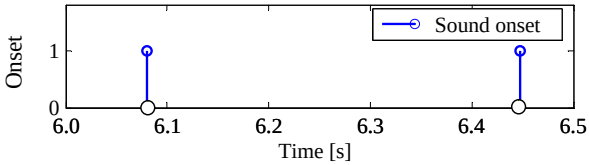
図 4.9 は，図 4.8(e) に示した運動範囲 (1) における，運動方向の変化を示す．(a) は対象物の運動範囲 (1) 内の x 方向の重心座標であり，図中のデータの欠落は，物体の運動軌跡を画像のフレーム間差分により求めているので，運動方向の変化時では重心

軌跡が抽出できないことによる．欠落した時点のデータを一時点前のデータによって補間すると，(b) で示す軌跡が得られる．その後，平滑化処理を行った結果が (c) である．この軌跡に対して右，左，両側における三つの時系列の内，右側の運動方向変化の時系列を (d) に示す．以上のように輝度分布の線形変換によって，差分画像が抽出しやすくなり，(a) のように運動軌跡を求めることができ，(d) のように運動方向変化の時系列を得る．

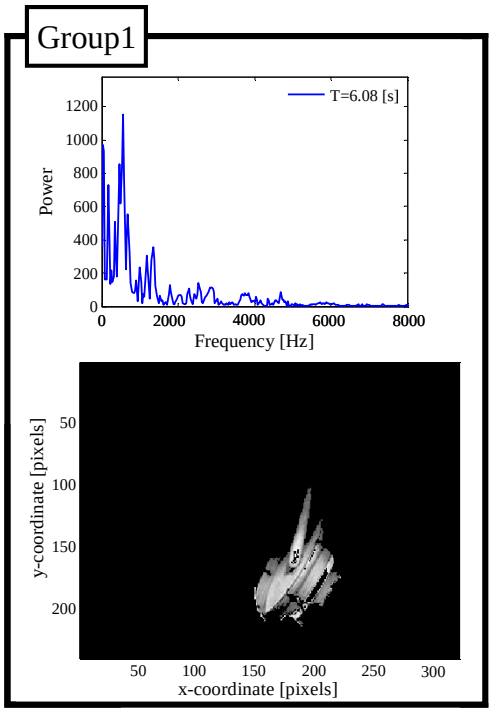
図 4.10 は，計測音と検出した音オンセット，それに対応する音スペクトルと映像である．(a) は計測音，(b) は検出した音オンセットである．(c) は，グループ 1 に属する音スペクトルと対応付けられた映像である．(d) は，グループ 2 に属する音スペクトルである．(c) の結果を見ると，対象物の運動方向が変化したときの映像と対象物によって発生した音が対応付けられており，視聴覚事象の対応付けが成功したことを確認できる．



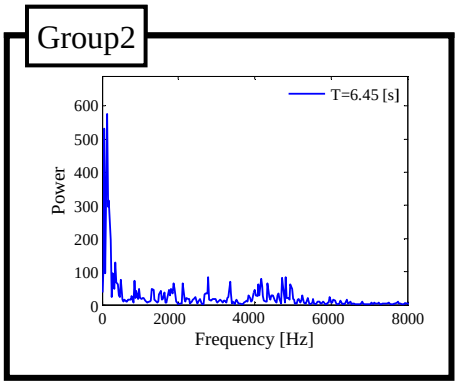
(a) Measurement signal



(b) Onset time series



(a) Spectrum of sound onset and image belonging to Group 1



(b) Spectrum of sound onset belonging to Group 2

図 4.10: 対応付けられた視聴覚事象

Fig. 4.10 Correspondence of one movement and two sounds

表 4.4: 対象物に関する音源定位誤差と視聴覚事象間の対応付け
Table 4.4 Correspondence between audio-visual events relating to objects

Object	Sound localization				Correspondence	
	DP [deg]	RMS	SD	CORR	(A,V)	SR
Bell	−45	16.3	32.5	0.919	-	0/2
	−30	16.7	21.9	0.812	0.888	2/2
	0	13.7	21.8	0.862	0.868	2/2
	30	11.3	39.9	0.858	0.850	2/2
	45	9.9	22.5	0.843	0.903	2/2
	Subtotal	13.8	27.7	0.859	0.850	8/10
Drum	−45	5.3	1.1	0.925	0.653	2/2
	−30	2.9	2.3	0.882	0.548	2/2
	0	8.0	2.2	0.880	0.700	2/2
	30	10.5	1.4	0.903	0.722	2/2
	45	7.0	1.0	0.891	0.672	2/2
	Subtotal	7.2	1.6	0.896	0.668	10/10
Met.	−45	7.5	10.2	0.800	0.678	2/2
	−30	8.8	1.7	0.884	0.687	2/2
	0	5.5	0.9	0.877	0.665	2/2
	30	1.1	0.0	0.881	0.693	2/2
	45	0.0	0.0	0.934	0.782	2/2
	Subtotal	5.7	2.6	0.875	0.701	10/10
Total		9.6	10.6	0.877	0.742	93.3% (28/30)

視覚的注意実験 II

表 4.4 は，対象物としてベル，ドラム，メトロノーム用い，対象物の正面から角度 $\theta = \{-45, -30, 0, 30, 45\}$ [deg] だけロボットヘッドを回転させた状態からの視聴覚事象の対応付け結果である．対象ごとに 5 つの角度で 2 回ずつ実験を行った，計 10 回の平均を表す．初期状態を IP(initial position)，音源定位における推定値と真値との誤差を 2 乗平均平方根 RMS(root mean square) によって表し, 同一音源グループ内の音オンセットごとの音源定位推定値の分布を標準偏差 SD により示し，グループ化された音源の第一音オンセットと他の音オンセットのスペクトル間の相関を CORR によって説明する．視聴覚事象の対応付け結果は，Correspondence として視聴覚の相関を (A,V)，対応付けの成功率を SR(success rate) で表す.

表 4.4 を見ると，音源定位の誤差 RMS はメトロノーム, ドラム, ベルの順に大きくなっており, またベルの標準偏差が大きくなっているのは，視聴覚事象が左右で発生しているためである．音源定位の推定値と真値との誤差は平均で 9.6 [deg] であり，視線方向の調節可能な範囲である．また音源のグループ化における最初の音源とグループ内の他の音源との相関の平均は 0.877 であり，目的の音源について定位ができていることがわかる．視聴覚事象の対応付けは 30 回の実験で 28 回成功し，成功率は 93.3%

であり，成功した実験の視聴覚事象の相関の平均は 0.742 である．画像の輝度を調節した後の平均輝度は，30 回の平均において 54.0 であり，すべて輝度ヒストグラムの線形変換を行った．なお，輝度ヒストグラムの線形変換を行わない場合，視聴覚事象の対応付けは 17 回成功し，成功率は 56.7%，視聴覚の相関の平均は 0.657 である．

成功率の目標値を定めるために，明るく雑音が小さい場所に対象物があり，視聴覚事象共に良好に抽出できる場合について実験を行った．対象物の正面から角度 $\theta = \{-45, -30, 0, 30, 45\}$ [deg] だけロボットヘッドを回転させた初期状態より，三つの対象（ベル，ドラム，メトロノーム）ごとに，5 通りの角度で 1 回ずつの計 15 回の視聴覚事象の対応付け実験を行った．音源定位の推定値と真値との誤差は 2 乗平均平方根で 9.4 [deg] であり，画像の輝度を調節した後の平均輝度は 15 回中の平均において 125.5 であった．この結果，音源定位，画像の輝度の調節が良好に行えたことを確認した．視聴覚事象の対応付けは 13 例成功し，視聴覚事象間の相関は平均で 0.823 であり，成功率は 86.7% である．対応付けが失敗したのは，ベルとドラムが 1 例ずつであり，運動方向が変化する両側，左右の位置の検出が間違っていたことによる．このことから，画像の輝度が良好な場合は，音源定位のみで視野外の物体への注意が行え，運動情報を抽出することができ視聴覚の対応付けを行うことができた．以上の結果から，本実験の成功率の目標値を 85% とし，視覚情報の抽出が困難な場合にも視線方向や絞りの調整，輝度ヒストグラムの線形変換により，視聴覚事象の対応付けの成功率が 85% を超えており，対応付けが行えることを確認した．

本章では，視聴覚による選択的注意を目標としており，実験の簡略化のため音源定位およびカメラの視線方向の調節は，左右方向のみとした．本システムの配置を鉛直方向にすることで，上下方向の音源定位や視線方向の調節へ対応できる．しかし，頭部運動を 3 次元的にとらえることにより，人が 3 次元空間上において視線をどのように制御しているかについての報告もなされており [7]，聴覚と組合せた 3 次元的な視線方向の制御については検討が必要である [72]．また，周辺部が明るく中央部が暗い部分に注意を向けたが，逆に周辺部が暗く中央部が明るい場合にも輝度の閾値の設定により対応できる．

なお，視覚心理物理実験に基づく知見として，注視している場所と注意を向けている場所が異なる場合がある [73–80]．本研究では，注視領域外の画像領域の情報も用いて音オンセットの検出を行っているので，この機能を実現していると考えられる．

4.6　おわりに

本章では選択的注意として，聴覚的注意では音源定位および雑音に対して頑健な音オンセット検出を行い，また視覚的注意では局所的に輝度が異なる場面における対象物の検出手法について提案した．

実験の結果，視覚事象と聴覚事象が共に良好に抽出できる場合の視聴覚事象の対応付けの成功率が 86.7% であったことから，注意による視聴覚情報の対応付けにおける成功率の目標値を，85% 以上とした．聴覚的注意において，視覚情報による時間軸上での窓関数を用いて音オンセットを検出した場合，視聴覚事象の対応付けの成功率は，注意を向けないときの 78.6% から 95.2% へ上昇し，聴覚的注意手法の有効性を確認した．また，視覚的注意において，周辺が明るく，中央部が暗い場所における対象物の視聴覚事象の対応付け実験の成功率は 93.3% であり，このことから局所的に輝度が異なる場面でも視覚的注意により，対象物の視覚情報を検出できることを確認した．これらの技術は，人間のパートナーとして期待されるロボットにとって視聴覚処理の効率化と正確さの向上において有益である．

本章では，視聴覚による選択的注意を目標としており，音源定位およびカメラの視線方向の調節は，左右方向のみとした．本システムの配置を鉛直方向にすることで，上下方向の音源定位や視線方向の調節へ拡張できる．また，周辺部が明るく中央部が暗い部分に注意を向けたが，逆に周辺部が暗く中央部が明るい場合にも輝度の閾値の設定により対応できる．さらに，注意によって得られた視聴覚事象の対応付け結果から学習することで，概念を獲得できると考えられる．これらについては今後の課題である．

第5章　視聴覚事象の統計的関係を用いた概念の獲得

5.1　はじめに

人は視聴覚事象を対応付けし，事例として蓄積したものを体系化することによって，概念として獲得すると考えられる．心理学的に，概念は，個別的具体的なものから共通する一般的抽象的なものへつくりあげられたものをいう [81]．実世界には非常に多様な事物があり，また時間的に変化するので，実世界に関する概念をすべてロボットやシステムに事前に与えることは不可能である．そのため，システムが環境に存在する事物に関する概念を自律的に学習することが必要となる．

生後 18〜20 週の幼児は発話と口の動きの対応を認識し，言語を獲得している [8]．また，統計的な情報を用いて幼児が言語の学習を行うことが確認されている [9]．このような機能を実現するため，複数モダリティの統計的な相関関係を基に，音や画像入力から概念を自律的に獲得するシステムを構築することが望まれる．

そして，画像や音といった複数の情報源を持つマルチモーダルなシステムが，それらの情報を統合して概念を学習するための枠組みについて研究が行われている [82]．栗田ら [83] の研究では，印象語と絵画といった定量的に比較できない関係を正準相関分析を行うことで，定性的に関係付け，印象語から絵画や類似画の検索，未知の絵画に対する印象語の推定を行った．井手ら [84] は，教師映像データから映像内容（ニュース映像中の字幕）と画像特徴量の関係を統計的に学習して対応付け，未知の映像の画像特徴量から映像内容を推測した．自律的な知能システムの実現のため，知識の獲得と学習について，モデル化がなされてきた [85–89]．石黒ら [90] は，静止画像と，その画像を表す音声のマルチモーダル情報の入力から，カーネル正準相関分析によってモダリティ間の統計的相関関係を得た．

3 章では，運動情報，音の発生，物体の映像上での動きに対して，同時性と類似性の手掛かりを用いて，能動的に視聴覚事象を対応付ける手法を示した [59]．また，4 章では対象物へ選択的に注意を向けることで視聴覚事象を対応付ける手法を示した [91]．次に，対応付けた事例をどのように分類し，どのような事例であるかを理解すること

が必要となる．また，経験し記憶した視聴覚事象に基づいて，新たな視聴覚事象がどの事例かの識別や，視覚，聴覚のどちらか一方の情報を得たときに，それに対応する聴覚または視覚の事象を想起することも必要である．

本章では，対応付けた事例を共通する特徴によって分類し，それら事例の集合体および中心的な事例によって表されるものを概念とする．視聴覚事象についての概念を自律的に獲得するため，対応付けられた画像と音の事例の集合の統計的関係を学習する．すなわち，正準相関分析により画像と音の特徴との相関関係を学習し，正準空間を作成する．この正準空間において，特徴ベクトルに対して教師なし学習を行い，概念を獲得する．その後，新たに示された物体の音（画像）をその物体を表す概念の音（画像）への識別を可能とする．または，物体の画像（音）に対して，その物体を表す概念の音（画像）を想起できるようにする．

本章の構成は次のとおりである．まず **5.2** では概念の獲得について説明する．**5.3** においては，視聴覚における概念の獲得，事例の識別と想起処理について述べる．**5.4** では実験方法を，**5.5** では実験結果を示すとともに考察を行う．最後に，**5.6** で本章をまとめる．

5.2　概念の獲得とは

認知心理学において，概念が果たしている機能として，分類，理解と説明，予測，推論，概念の組合せ，事例の生成等が挙げられる [92]．また，概念学習と概念表現のモデルは，以下に挙げるものを含め，様々なものが考えられている．(1) プロトタイプモデル：概念が，多くの所属事例に共通する特徴からなるプロトタイプ，または最も典型性の高い事例を中心として構成されているとするモデル [93]，(2) 事例モデル：概

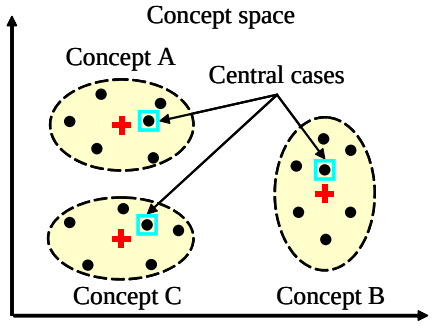


図 5.1: 概念と中心的な事例の関係
Fig. 5.1 Relation between concepts and central cases

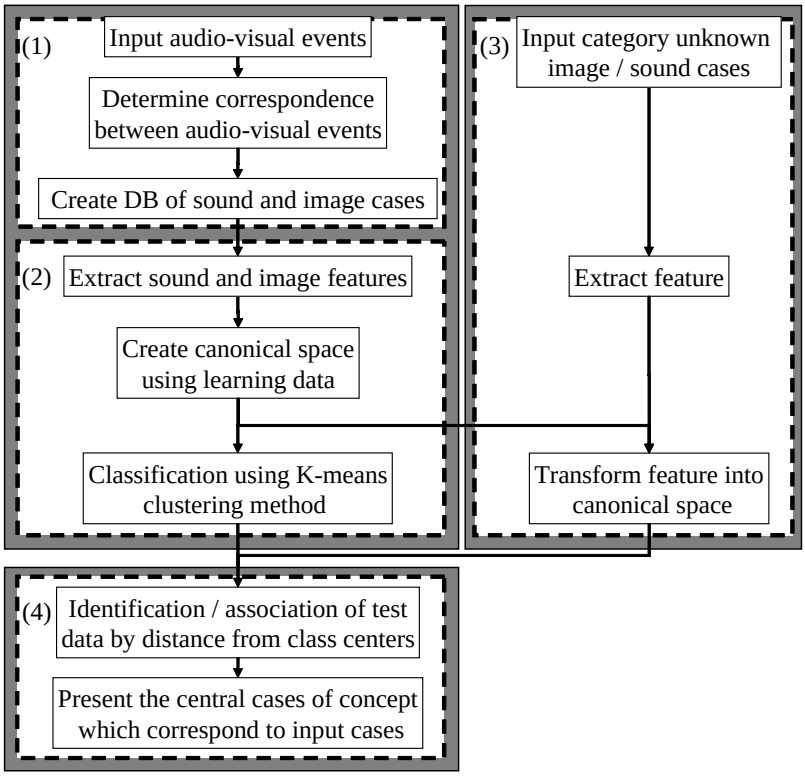


図 5.2: 概念獲得と識別，想起処理
Fig. 5.2 Process of concept acquisition and identification, association

念は所属する個々の事例によって表現されているとするモデル.

本研究では，概念が果たしている機能において，分類と事例の生成の 2 つを取り上げ，概念は，共通の特徴を持つ事例の集合体からなり，中心的な事例を概念に属する最も典型性の高い事例として表現する．概念空間は，それぞれの概念にとって，本質的な特徴を表すための空間とする．**図 5.1** に示すように，概念空間において，“ • ”が事例，“ + ”が獲得された概念のクラス中心であるとき，中心的な事例とは，クラス中心から最も距離が近い事例である．得られた概念が何を表現しているのかは，この中心的な事例によって代表される．本研究では，このような概念を事前知識（カテゴリ情報）なしで学習することを概念の獲得とする [94–96].

5.3 概念獲得と識別および想起の処理

図 5.2 は，概念獲得と識別，想起処理の概要である．概念獲得処理は，(1) の視聴覚事象のパターン抽出部と，(2) のクラスタリング部によって行われる．また，概念獲

得後の事例の識別と想起処理は，同図の (3) の事例の変換部，(4) の識別と想起部で行われる．(1) の視聴覚事象のパターン抽出部において，視聴覚事象を対応付け，対応付けられた画像と音のデータベース (以下，DB) を作成する．(2) のクラスタリング部において，音と画像の特徴抽出と得られた特徴量に対して正準相関分析 (以下，CCA : canonical correlation analysis) を行い，音と画像の特徴量を正準空間へ写像する変換行列を学習する．この正準空間が概念空間となる．そして，作成した正準空間において K-means 法によりクラスタリングを行う．

(3) の事例の変換部では，(2) のクラスタリング部において求めた変換行列により，入力事例の特徴を正準空間へ写像する．写像は，次の (4) の識別と想起部において，識別する場合と想起する場合の 2 通り行う．識別する場合では，音（画像）の入力事例を音（画像）の正準空間へ写像する．想起する場合は，音（画像）の入力事例を画像（音）の正準空間へ写像する．

最後に，(4) の識別と想起部では，入力事例と (2) において獲得された概念の各クラス中心との距離を比較し，距離が最小のクラスの中心的な事例を示す．ここで，識別では入力事例（音，画像，音と画像）に対応する概念の中心的な事例（音，画像，音と画像）を示す．想起では，入力事例（音，画像）に対応する概念の中心的な事例（画像，音）を示す．

5.3.1　特徴抽出

(a)　聴覚特徴

聴覚特徴として，**2.3** で述べた線形予測分析によって得られる線形予測係数 (LPC) を用いる．LPC は，音が調音フィルタの出力であることを前提として，効率的に音のスペクトルの概形を線形予測分析によって求められる．本研究では，1 次〜10 次の 10 次元の LPC を音特徴ベクトルとする．

図 5.2 の (2) クラスタリング部における音の特徴量の抽出処理では，映像と対応付けられた音を音オンセット時刻から $N(= 1024)$ 点の信号とした．音オンセットとは，**3.3.1(b)** において述べたように，背景雑音以外の音がない状態で，音の大きさが急激に変化した時刻からの短時間の音とする [59]．

(b) 視覚特徴

視覚特徴として，**2.4** で述べた Hu モーメントを用いる．これは，画像内のオブジェクトの平行移動や回転移動，スケール変化の影響を受けない量であり，正規重心モーメントを組み合わせた式で，7 種類定義されている [39]．この抽出した各画像の色相，輝度に対して，7 つの Hu モーメントを計算し，計 14 次元の画像特徴ベクトルを抽出する．ここで，入力のカラー画像から輝度画像への変換には，**3.3.2(a)** の式 (3.2) を用いる．また，色相 (Hue) は，色の様相の相違を表し，HSV 色空間を彩度 (Saturation) と明度 (Value) と共に構成する．座標 (x, y) における色相 $h(x, y)$ は，入力のカラー画像の RGB 各画素の値 $f_R(x, y)$，$f_G(x, y)$，$f_B(x, y)$ より，次式のように求める．

$$h(x,y) = \left[\alpha - \arctan\left\{\frac{2f_R(x,y)-f_G(x,y)-f_B(x,y)}{\sqrt{3}(f_G(x,y)-f_B(x,y))}\right\}\right]/2\pi$$

$$\begin{cases} \alpha = \frac{\pi}{2} & (f_G(x,y) > f_B(x,y)) \\ \alpha = \frac{3\pi}{2} & (f_G(x,y) < f_B(x,y)) \\ h(x,y) = 1 & (f_G(x,y) = f_B(x,y)) \end{cases} \tag{5.1}$$

これまで，音と対応付いた物体の映像を，3 フレームの輝度画像間における差分画像によって求めていた [59]．しかし，この処理では動いている部分のみを抽出するため，物体の一部分の映像しか得られなかった．ここでは，3 フレーム間差分画像をシードとして 8 近傍の膨張処理を行うことによって物体全体の画像を抽出する．処理は次のように行う．

Step 1 画像上座標 (x, y) における，物体および背景が撮影されている原画像 (original image) の RGB 画素値 $f_o(x, y)$ から，輝度値 $g_o(x, y)$ と色相 $h_o(x, y)$ を求める．3 フレーム間差分画像 (difference image) [59] の領域に RGB 各画素を加えた画像を $f_d(x, y)$ の初期値とし，$f_d(x, y)$ から輝度値 $g_d(x, y)$ と色相 $h_d(x, y)$ を求め，画面上を左上から右下へ走査する．

Step 2 $g_d(x, y) = 0$ かつ画素点 (x, y) を中心とする 8 近傍の画素点 $(x + i, y + j)$，$-1 \le i \le 1$，$-1 \le j \le 1$ において，以下の条件

(a) $g_o(x, y) \ge th_{il}$かつ $g_d(x + i, y + j) > 0$

(b) $|g_d(x + i, y + j) - g_o(x, y)| \le th_{id}$

(c) $h_o(x, y) \le th_{hl}$かつ $th_{hu} < h_o(x, y)$

(d) $|h_o(x + i, y + j) - h_d(x, y)| \le th_{hd}$

をすべて満たす画素が 1 つでもある場合，物体画像に含めて $f_d(x,y) := f_o(x,y)$ とし，$g_d(x,y)$，$h_d(x,y)$ を更新して走査を続ける．

Step 3 走査を完了しても，$f_d(x,y)$ が更新された画素がひとつもない場合は処理を終了し，$f_d(x,y)$ を物体画像とする，そうでない場合は **Step2** へ戻る．

ここで，画像の横と縦のサイズを $w(=320)$[pixels] と $h(=240)$[pixels] とする．また，画像の階調値については，RGB はそれぞれ 0〜255，輝度は 0〜255，色相は 0〜1 とする．上述の条件 (a)〜(d) の閾値は，予備実験により，それぞれ $th_{id}=150$，$th_{il}=70$，$th_{hu}=0.8$，$th_{hl}=0.4$，$th_{hd}=0.5$ とする．

(c)　視聴覚データのサンプリング

対応付けられた視聴覚特徴量の組を視聴覚データと呼ぶ．視聴覚データ数は，観測毎で異なる．また，本システムはロボットへの実装を想定しており，あらかじめ決定した 12 箇所の異なる視点から物体を観測するため，次のように視聴覚データをサンプリングする．

Step 1 視聴覚事象を対応付けた各物体の中で，最小の視聴覚データ数をサンプリング数 N とし，視点の位置番号を i，各視点でのデータ番号を j，視聴覚事象を対応付けた各物体の聴覚または視覚データを $O_{i,j}$ とする．ここで，入力順序に依存しないよう，視点 i におけるデータ $O_{i,j}$ のデータ番号 j を無作為に入れ替える．

Step 2 全データ O_{ij} に対して，各視点でのデータ番号 j を第 1 キー，視点番号 i を第 2 キーとして，昇順にソートし，上位 N 個を抽出する．

5.3.2　正準相関分析による概念空間の作成

5.3.1 において求めた視覚の原特徴 $\boldsymbol{x}$，聴覚特徴 $\boldsymbol{y}$ に対して，**2.5** に述べた正準相関分析を行い，求められた視聴覚の正準主成分 $\boldsymbol{a}$，$\boldsymbol{v}$ によって構成される正準空間を作る．これらによって得られる正準空間が概念空間となる．

5.3.3　クラスタリングによる概念の獲得

正準空間におけるクラスタリング処理について述べる．音，映像を物体ごとにまとめるためクラスタリングを行い，中心的な事例を決定する．本研究では，**2.6** に述べた

クラス数 K を与えてクラスタリング処理を行う K-means 法を用いる．K-means 法は，初期クラス中心位置と点列データの入力順序にある程度依存する．そのため，**2.7** に述べた初期クラスタとクラスタ数の決定法により，まずパターンの入力順序が異なるデータ列を用いてクラスタリングを 10 回行い，最も分離精度の良い結果を採用する．さらに，未知のクラス数 K を決定するため，**2.7** のクラスタ数の決定法を用いる．

中心的な事例は，クラスタリング後のクラス j の中心 $\boldsymbol{c}_j$ とのマハラノビス距離が最小のパターンとする．

5.3.4 入力事例の識別と想起

概念を獲得した後に，カテゴリ未知の入力事例に対応する概念の中心的な事例を提示することを，事例の識別とする．また，カテゴリ未知の入力事例に対応する概念の異なるモダリティにおける中心的な事例を提示することを事例の想起とする．

(a) 識別

獲得した概念に基づく，入力事例の識別を以下の方法で行う．ここで，CCAK は本章における識別の提案手法であり，他手法は CCAK と識別成功率を比較するためのものである．

OFC (original feature category) は原特徴を学習データとし，どのカテゴリ (物体) に含まれるかという事前知識を与えてクラス中心を求め，原特徴空間においてテストデータを識別する．

OFK (OF k-means) は原特徴を学習データとし，K-means 法でクラスタリングし，原特徴空間においてテストデータを識別する．

CCAC (canonical correlation analysis category) は正準空間へ変換した特徴を学習データとし，各々の学習データがどのカテゴリ (物体) に含まれるかという事前知識を与えてクラス中心を求め，正準空間においてテストデータを識別する．

CCAN (CCA number) は正準空間へ変換した特徴を学習データとし，各々の学習データのカテゴリ数を事前知識として与えてクラスタリングし，正準空間においてテストデータを識別する．

CCAK (CCA k-means) は正準空間へ変換した特徴を学習データとし，K-means 法によってクラスタリングし，正準空間においてテストデータを識別する．

CCAKAV (CCAK audio-visual) は CCAK と同様の処理によってクラスタリングを行うが，音（画像）のテストデータと後述の式 (5.5) で推定される画像（音）特徴を合成し，正準空間においてテストデータを識別する．

獲得した概念のクラス数が K であるとき，概念のクラス中心を $\boldsymbol{a}_c = Z\boldsymbol{x}_c$，$\boldsymbol{v}_c = W\boldsymbol{y}_c$ $(c = 1, \cdots, K)$ で表し，テストデータの原特徴を $\boldsymbol{x}_\alpha^{(t)}$，$\boldsymbol{y}_\alpha^{(t)}$，それらの正準変量を $\boldsymbol{a}_\alpha^{(t)} = Z\boldsymbol{x}_\alpha^{(t)}$，$\boldsymbol{v}_\alpha^{(t)} = W\boldsymbol{y}_\alpha^{(t)}$ で表す．OFC，OFK の識別には，音（画像）のテストデータ $\boldsymbol{x}_\alpha^{(t)}$（$\boldsymbol{y}_\alpha^{(t)}$）と音（画像）のクラス中心 $\boldsymbol{x}_c$（$\boldsymbol{y}_c$）との距離を用いる．音と画像がともに入力されるときは，$[\boldsymbol{x}_\alpha^{(t)}, \boldsymbol{y}_\alpha^{(t)}]$ と $[\boldsymbol{x}_c, \boldsymbol{y}_c]$ との距離を用いる．

CCAC，CCAN，CCAK の識別には，音（画像）のテストデータ $\boldsymbol{a}_\alpha^{(t)}$（$\boldsymbol{v}_\alpha^{(t)}$）と音（画像）のクラス中心 $\boldsymbol{a}_c$（$\boldsymbol{v}_c$）との距離を用いる．音と画像がともに入力されるときは，$[\boldsymbol{a}_\alpha^{(t)}, \boldsymbol{v}_\alpha^{(t)}]$ と $[\boldsymbol{a}_c, \boldsymbol{v}_c]$ との距離を用いる．

CCAKAV では，音（画像）のテストデータ $\boldsymbol{a}_\alpha^{(t)}$（$\boldsymbol{v}_\alpha^{(t)}$）と，後述する式 (5.5) で推定される画像（音）特徴 $\tilde{\boldsymbol{v}}_\alpha^{(t)}$（$\tilde{\boldsymbol{a}}_\alpha^{(t)}$）による $[\boldsymbol{a}_\alpha^{(t)}, \tilde{\boldsymbol{v}}_\alpha^{(t)}]$（$[\tilde{\boldsymbol{a}}_\alpha^{(t)}, \boldsymbol{v}_\alpha^{(t)}]$）とクラス中心 $[\boldsymbol{a}_c, \boldsymbol{v}_c]$ との距離を識別に用いる．ただし，音と画像がともに入力されるときは $[\tilde{\boldsymbol{a}}_\alpha^{(t)}, \tilde{\boldsymbol{v}}_\alpha^{(t)}]$ とクラス中心 $[\boldsymbol{a}_c, \boldsymbol{v}_c]$ との距離を用いる．

(b)　想起

獲得した概念に基づく，入力事例の想起は以下の方法で行う．変量 $\boldsymbol{a}_\alpha$，$\boldsymbol{v}_\alpha$ の相関行列 R_A，R_V は次式となる．

$$R_A = \sum_{\alpha=1}^{n} \boldsymbol{a}_\alpha \boldsymbol{a}_\alpha^t / n = ZR_X Z^t = I_s \tag{5.2}$$

$$R_V = \sum_{\alpha=1}^{n} \boldsymbol{v}_\alpha \boldsymbol{v}_\alpha^t / n = WR_Y W^t = I_s \tag{5.3}$$

ここで，I_s は s 次元の単位行列である．

また，式 (2.50) より次式が求められる．

$$R_{XY} W^t = R_X Z^t \Lambda \tag{5.4}$$

相互相関行列 R_{AV} は次式となる．

$$\begin{aligned} R_{AV} &= \sum_{\alpha=1}^{n} \boldsymbol{a}_\alpha \boldsymbol{v}_\alpha^t / n = ZR_{XY} W^t = ZR_X Z^t \Lambda \\ &= \Lambda = R_{VA}^t \end{aligned} \tag{5.5}$$

ここで，Λ は，相関係数 r_k を対角要素とする行列である．よって，$\boldsymbol{a}_\alpha$ から $\tilde{\boldsymbol{v}}_\alpha$ へ，あるいは $\boldsymbol{v}_\alpha$ から $\tilde{\boldsymbol{a}}_\alpha$ への線形回帰式は，

$$\tilde{\boldsymbol{v}}_\alpha = \Lambda \boldsymbol{a}_\alpha,\ \tilde{\boldsymbol{a}}_\alpha = \Lambda \boldsymbol{v}_\alpha \tag{5.6}$$

で与えられる（証明は**付録 A** 参照）.

未知パターン $\boldsymbol{x}_\alpha(\boldsymbol{y}_\alpha)$ から推定された $\tilde{\boldsymbol{v}}_\alpha = \Lambda Z \boldsymbol{x}_\alpha(\tilde{\boldsymbol{a}}_\alpha = \Lambda W \boldsymbol{y}_\alpha)$ と概念のクラス中心 $\boldsymbol{v}_c = W\boldsymbol{y}_c$ $(\boldsymbol{a}_c = Z\boldsymbol{x}_c)$ との間のマハラノビス距離が最小となる概念を想起し，概念の中心的な事例によって示す.

想起の処理を以下のように行う.

CCACAS (CCAC association) は CCAC と同様の方法でクラス中心を求め，正準空間においてテストデータの画像 (音) から音 (画像) の事例を想起する.

CCAKAS (CCAK association) は CCAK と同様の方法でクラスタリングし，CCACAS と同様に想起する.

CCACAS，CCAKAS の想起には音（画像）のテストデータ $\boldsymbol{a}_\alpha^{(t)}(\boldsymbol{v}_\alpha^{(t)})$ から推定される画像（音）特徴 $\tilde{\boldsymbol{v}}_\alpha^{(t)}(\tilde{\boldsymbol{a}}_\alpha^{(t)})$ と画像（音）のクラス中心 $\boldsymbol{v}_c(\boldsymbol{a}_c)$ との距離を用いる．ここでの提案手法は CCAKAS であり，カテゴリ既知の CCACAS との比較を行う.

5.4 実験

ドラム，ベル，メトロノームについて，運動によって発生した画像と音を群化の法則を基に対応付け [59]，それぞれの特徴量を算出する．正準相関分析により作成した正準空間においてクラスタリングを行う．最後に，カテゴリ未知の入力事例から対応する概念の同じモダリティ（異なるモダリティ）における中心的な事例を提示し，識別（想起）を行う．実験装置として，1 個のマイクロフォンと 1 台のカメラおよびこれらからの信号を計測する計算機とデータ処理のための計算機を使用する [59].

5.4.1 実験環境と条件

対象事象は，振られているベル，たたかれているドラム，動作中のメトロノームである．いずれも動作とともに音が発生している．シーンには動き，音がともに 1 つのみ存在するとする．ここで，対象物の形状と音を抽出しやすくするために，物体には黄色のセロハンを張り，背景は黒色の画用紙（マーメイド）とし，対象物の音の他に

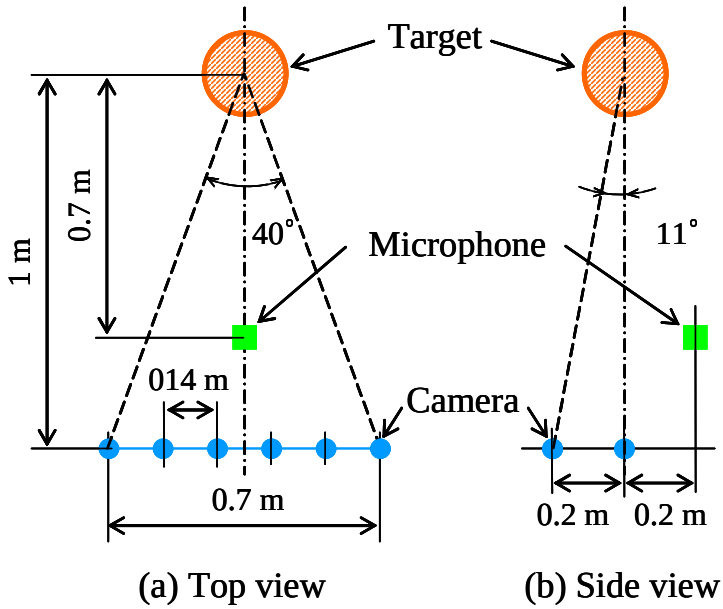


図 5.3: 実験装置配置
Fig. 5.3 Setup of experimental equipment

雑音はないものとする．**図 5.3** は，マイクロフォンとカメラの配置を示す．1 箇所に固定した 1 つのマイクロフォン (RION，UC-30) と，物体を異なる視点で観測するために，カメラ (SONY，EVI-G20) の位置を変え（左右 6 箇所，2 段階の高さの計 12 箇所），各箇所で 2 回，時系列サンプルを計測する．12 の時系列サンプルを 1 組のデータとし，1 回目を Data set 1，2 回目を Data set 2 とする．

映像と対応付けられた音は，音オンセット時刻から 1024 点の音信号（サンプリング周波数 16 [kHz]，量子化 8 [bits]）とする．フーリエ変換時には，サンプリングを 512 点，シフト量を 256 点 とする．線形予測分析は，青木 [97] のプログラムを参考にし，分析におけるフレーム長は 400 点とした．

映像は画像サイズ 320 × 240 [pixels] であり，サンプリングレート 30 [Hz]，輝度は 8 [bits]256 階調，色相は 0〜1 に正規化した特徴量を用いた．

5.4.2　概念の獲得実験

学習データの数（総サンプル数）は，最小のサンプル数 N_{min} に時系列数 S をかけた数とする．正準相関分析では，視聴覚の特徴量の正準相関変量間の相関係数 r_k が，$r_k \geq th_c$ である k 次元までの正準変量を用いた．閾値 $th_c = 0.5$ は，原特徴の本質的

表 5.1: 視聴覚事象時系列の相関係数と抽出したパターン数
Table 5.1 Correlation between audio and visual events time series and the number of extracted patterns

Object	Correlation	Data set 1		Data set 2	
		Audio	Visual	Audio	Visual
Bell	0.946	71	70	70	67
Drum	0.924	91	61	90	65
Metronome	0.776	146	99	138	76

な構造を抽出するのに適当である値を，予備実験により求めた．K-means 法では，クラス数 K を 1〜10 まで変えてクラスタリングを行う．

5.4.3 事例の識別と想起実験

事例の識別と想起において，Closed テストとして学習データとテストデータに同じ Data set 1 を用い，Open テストとして学習データに Data set 1，テストデータに Data set 2 を用いる．すべての組の中で，抽出した音と画像が最小である視聴覚データ数を N_{min}，Data set ごとのデータの組数を S としたとき，総サンプル数 N_{all} を $N_{all} = N_{min} \times S$ とする．

5.5 結果と考察

5.5.1 外界からのパターン検出

表 5.1 は音と画像の時系列から視聴覚事象を対応付け，音と画像を抽出した結果である．表中の Correlation は，物体ごとの音と画像の発生時刻における相関係数の平均である．Data set は，抽出した音と物体画像のパターン数を表す．

物体ごとの音と画像の発生時刻における相関係数の平均は，ベルが最も大きく 0.946 で，以下ドラム 0.924，メトロノーム 0.776 である．抽出した Audio と Visual のパターン数は，ベルでは Data set 1 の 71 と 70，Data set 2 の 70 と 67 と差が少ないが，ドラム (91 と 61，90 と 65) とメトロノーム (146 と 99，138 と 76) では，音に比べ画像が少ない．これは画像のばらつきが大きく，音に比べて物体画像の抽出が難しいためである．

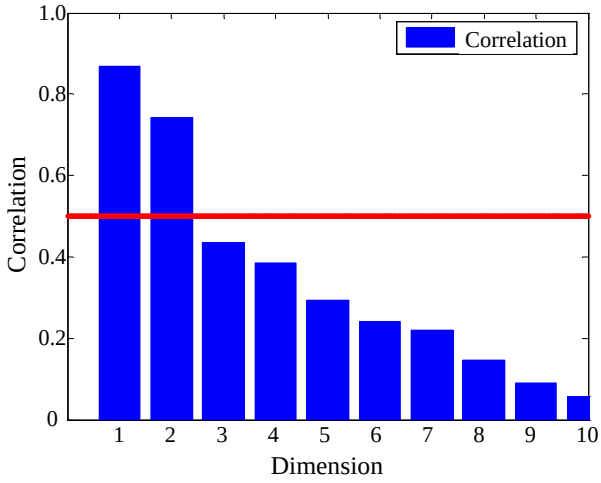


図 5.4: 視聴覚事象間の次元ごとの相関係数
Fig. 5.4 Correlation between audio-visual events for each dimension

5.5.2　概念の獲得

ベル，ドラム，メトロノームの 3 種類についての概念獲得実験の結果を示す．学習データとして，最小のサンプル数 N_{min} はドラムの Data set 1 の視覚情報に合わせ $N_{min} = 61$ とした．学習データに Data set 1，テストデータに Data set 2 を用い，データ組ごとに 61 サンプルの音と画像から時系列数 $S = 3$ 組で計 $N_{all} = 183$ サンプルの特徴量に対して正準相関分析を行った．**図 5.4** は，視聴覚事象間の次元ごとの相関係数を示す．視聴覚間の相関係数は，1 次が 0.867，2 次が 0.743，3 次が 0.434 であった．今回の実験では，相関係数が 0.5 以上である 2 次元までの正準主成分を用いた．学習データに対する K-means 法およびクラスタの有効性分析では，音と画像，画像のみ，音のみのすべてにおいて，クラス数は 3 と判定された．

図 5.5 は，クラスタリングを行った正準空間の特徴量分布を示す．図の (a)，(c)，(e) はカテゴリ既知の場合，(b)，(d)，(f) はクラスタリングを行った場合である．(a)，(b) は聴覚と視覚の 1 次正準変量，(c)，(d) は聴覚の 1，2 次正準変量，(e)，(f) は視覚の 1，2 次正準変量の分布である．

図 5.5 の“ • ”はベル，“ × ”はドラム，“ + ”はメトロノーム，“ □ ”はクラス中心を示す．楕円は各クラス内のパターンの分散共分散行列から求めた正規分布の等確率楕円であり，長半径と短半径は各軸方向での標準偏差である．(a) における直線は，1 次の正準変量において，聴覚を $\boldsymbol{a}^{(1)}$，視覚を $\boldsymbol{v}^{(1)}$ としたときの聴覚 $\boldsymbol{a}^{(1)}$ から視覚 $\tilde{\boldsymbol{v}}^{(1)}$

表 5.2: カテゴリ未知における中心的な事例のカテゴリ既知におけるクラス中心からの距離の近接順位

Table 5.2 Nearness ranks of distance between class centers in category known-case and center cases in category unknown-case

Object	Audio-Visual	Audio	Visual
Bell	1.67	3.20	6.22
Drum	4.00	1.00	5.40
Metronome	1.00	1.30	2.00

(視覚 $\boldsymbol{v}^{(1)}$ から聴覚 $\tilde{\boldsymbol{a}}^{(1)}$) への線形回帰直線である．2 直線の係数は，視聴覚間の 1 次の相関係数 $r_1 = 0.867$ である．

カテゴリ未知の場合の (b)，(d)，(f) は，カテゴリ既知の場合の (a)，(c)，(e) とそれぞれ非常に近い結果である．破線は K-means 法によるクラス中心の移動を示し，破線で結ばれた四角は，クラスタリング過程のクラス中心を示す．初期クラス中心が実際のクラス中心から遠い場合でも，K-means 法により，実際のクラス中心に近い位置に収束していることが分かる．

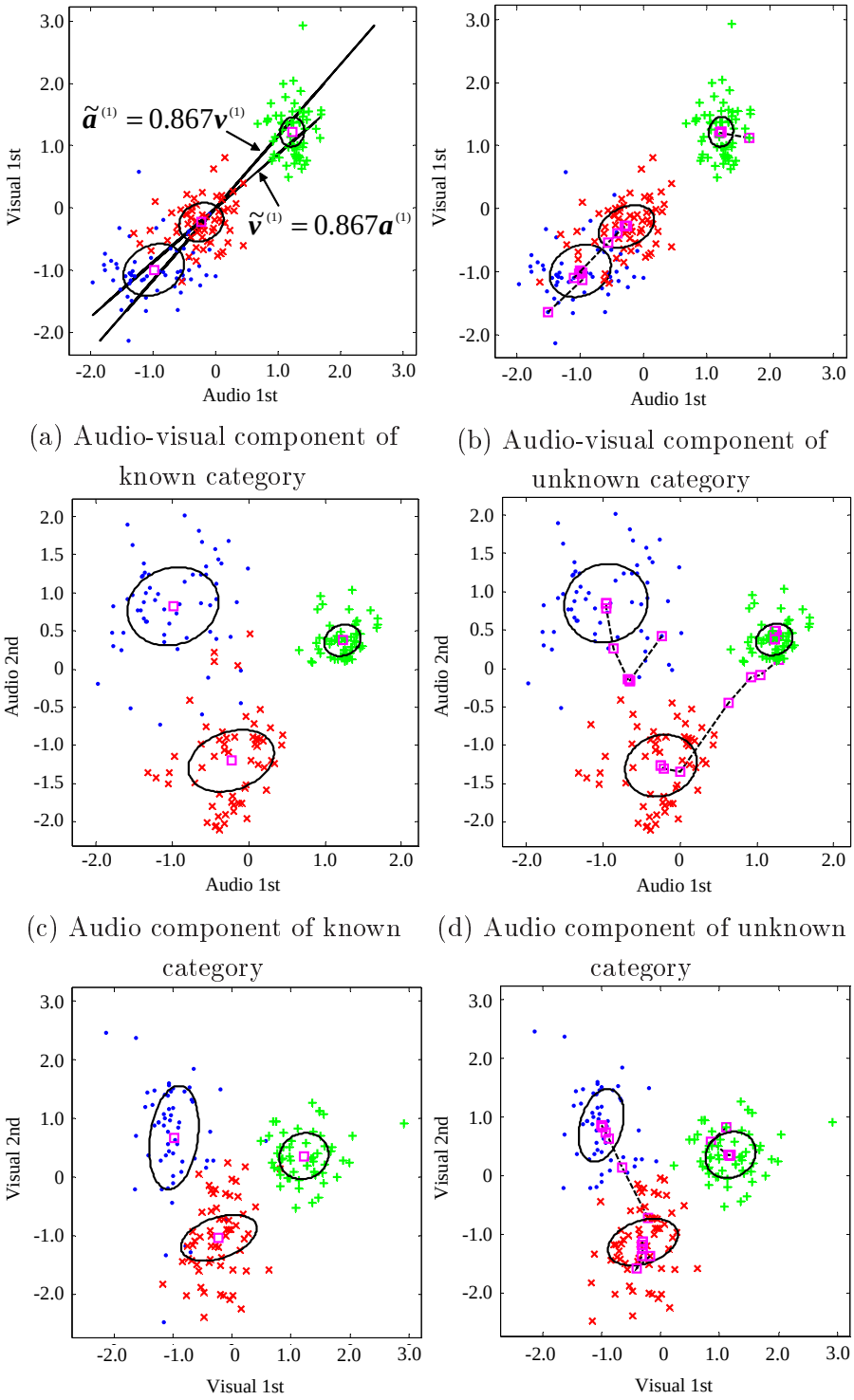
図 5.6(a)〜(f) に獲得された画像と音の概念ごとの中心的な事例を示す．(a)，(c)，(e) はベル，ドラム，メトロノームの画像を示す．(b)，(d)，(f) 中において，緑色の線は FFT スペクトル，赤色の線は音の LPC スペクトルを表す．これらから，中心事例がカテゴリの特徴を表していることが確認された．

中心的な事例がカテゴリ特徴を表しているか，妥当性を確認するため，カテゴリ未知における中心的な事例を，カテゴリ既知におけるクラス中心からの近接順位によって示す．カテゴリ未知のクラスタリングによって求められた中心的な事例の近接順位が全サンプルの上位であれば，カテゴリ特徴を良く表しているとする．以下で述べる識別および想起の成功は，中心的な事例の妥当性を反映し，カテゴリ未知の視聴覚事例と獲得した概念の中心的な事例との対応付けが正確であるときとする．

表 5.2 は，概念を獲得したときのベル，ドラム，メトロノームのカテゴリ未知の CCAK における中心的な事例を，カテゴリ既知の CCAC におけるクラス中心からのマハラノビス距離の近接順位で示す (詳細なデータは**付録 B 表 B.1** 参照)．近接順位は，クラスタリング結果の平均によって表す．

近接順位は 61 サンプル中，Audio-Visual において Drum の 4.00 位以内，Audio は Bell の 3.20 位以内となり良好であった．Visual は Bell の 6.22 位以内となったが，これは抽出した画像パターンのばらつきが大きかったためである．したがって，カテゴ

リ未知におけるクラスタリングから得られた中心的な事例はカテゴリの特徴を表していることを確認した.



(a) Audio-visual component of known category

(b) Audio-visual component of unknown category

(c) Audio component of known category

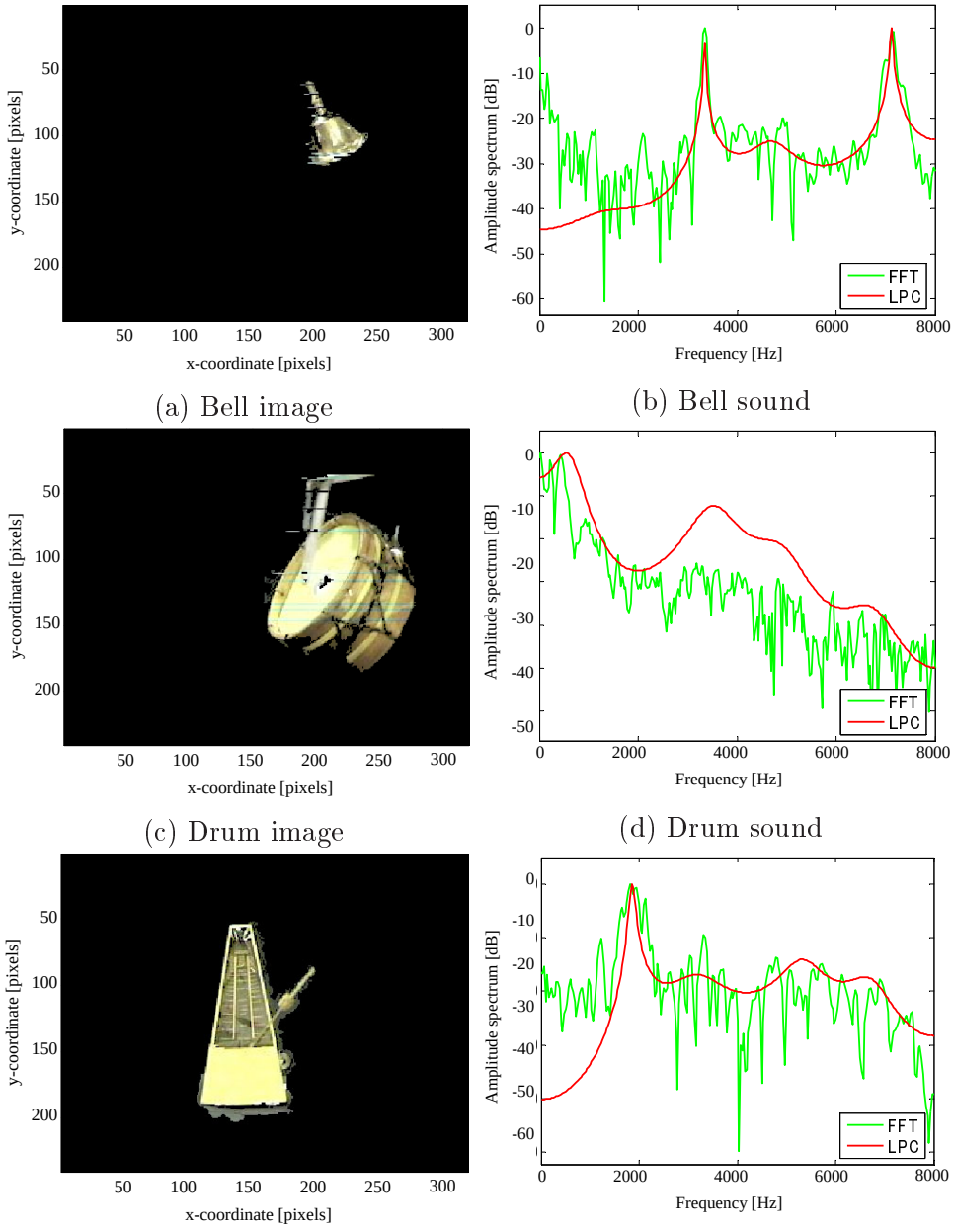
(d) Audio component of unknown category

(e) Visual component of known category

(f) Visual component of unknown category

図 5.5: カテゴリ既知と未知の場合における特徴分布

Fig. 5.5 Feature distribution in case of known category and unknown category



(a) Bell image　(b) Bell sound

(c) Drum image　(d) Drum sound

(e) Metronome image　(f) Metronome sound

図 5.6: 視聴覚事象の中心的な事例

Fig. 5.6 Central cases of audio-visual events

5.5.3 入力事例の識別と想起

(a) 識別

新しい入力事例の識別における識別成功率を**表 5.3** に示す (詳細なデータは**付録 B 表 B.2， B.3** 参照)．表 5.3 において，推定クラス数 CN(estimation class number) は **5.5.2** の概念の獲得処理によって求められたものであり，識別成功率 SR(success rate) との比較のために示す．推定クラス数 ，識別成功率は 10 回の実験の平均で示す．Classifier は，クラスタリング法と識別法である．

OFC における識別成功率は，Closed で Visual の 94.9% 以上，Open では Audio の 89.4% 以上と高い．OFK は，推定クラス数が 2.0 と実際のクラス数 3 からの誤差が大きく，クラス中心や中心的な事例が誤り，識別成功率は Closed において Audio の 16.7% 以上，Open において Audio の 15.8% 以上と低い．

本章における提案手法である CCAK は，推定クラス数が Audio の 2.9 や Visual の 2.8 と実際のクラス数 3 を良好に推定しており，識別成功率は Closed において Visual の 83.7% 以上，Open においても Visual の 83.2% 以上と高い．本研究における概念獲得成功率は，この CCAK における Closed テストの Audio-Visual 結果であり，推定クラス数 2.8，成功率 86.4% である．カテゴリ既知の CCAC において Closed では Visual の 93.5% 以上，Open では Visual の 91.9% 以上であることや，カテゴリ数既知の CCAN において Closed では Audio の 86.4% 以上，Open では Audio の 82.7% 以上であることから CCAK の目標成功率を 80% とした．本手法である CCAK の成功率は，83.2% 以上と目標値を達成している．一方，CCAKAV は，推定クラス数が Visual の 2.5 と実際のクラス数 3 よりも少なく推定しており，Closed において識別成功率は Visual の 65.8% 以上，Open において Visual の 67.3% 以上と CCAK より低い．

したがって，本章における提案手法である CCAK の識別成功率は，目標値を達成しており，また原特徴を用いた OFK や推定した特徴量を合成した CCAKAV に比べ識別成功率は高くなり，本手法の有効性が確認できた．

(b) 想起

表 5.4 は，画像（音）入力から対応する音（画像）の中心的な事例を想起したときの成功率を示す (詳細なデータは**付録 B 表 B.4，B.5** 参照)．成功率は 10 回の実験の平均である．CCACAS の想起成功率は Closed において Visual to Audio の 93.6% 以上，Open において Audio to Visual の 91.9% 以上と高く，CCAKAS は Closed にお

表 5.3: 識別成功率
Table 5.3 Success rate of identification

Test	Classifier	Audio-Visual		Audio		Visual	
		CN	SR	CN	SR	CN	SR
Closed	OFC	3.0	99.1%	3.0	97.4%	3.0	94.9%
	OFK	2.0	36.1%	2.0	16.7%	3.0	33.4%
	CCAC	3.0	96.9%	3.0	96.2%	3.0	93.5%
	CCAN	3.0	86.4%	3.0	86.4%	3.0	92.9%
	CCAK	2.8	86.4%	2.9	89.8%	2.8	83.7%
	CCAKAV	2.8	79.4%	2.8	86.5%	2.5	65.8%
Open	OFC	3.0	94.4%	3.0	89.4%	3.0	93.8%
	OFK	2.0	34.9%	2.0	15.8%	3.0	38.1%
	CCAC	3.0	95.2%	3.0	92.2%	3.0	91.9%
	CCAN	3.0	87.0%	3.0	82.7%	3.0	92.1%
	CCAK	2.8	87.2%	2.9	86.0%	2.8	83.2%
	CCAKAV	2.8	84.2%	2.8	89.2%	2.5	67.3%

表 5.4: 想起成功率
Table 5.4 Success rate of association

Test	Classifier	Visual to Audio	Audio to Visual
Closed	CCACAS	93.6%	96.5%
	CCAKAS	86.9%	85.6%
Open	CCACAS	92.0%	91.9%
	CCAKAS	85.8%	81.5%

いて Audio to Visual の 85.6% 以上，Open において Audio to Visual の 81.5% 以上と高くなっている．このことから想起が良好に行われたことが確認できた．

5.6　おわりに

本章では，概念獲得として分類，事例の生成について取り上げ，対応付けた事例を共通する特徴によって分類し，それらの分布および中心的な事例によって表されるものを概念とした．視聴覚事象についての概念を自律的に獲得するため，物体に対して，1 箇所に固定した 1 個のマイクロフォンと，異なる外観を得るために 12 箇所に順に移動させて配置した 1 台のカメラによって物体を計測する．音の発生，物体の映像上での動きに対して，同時性と類似性の手掛かりを用いて視聴覚事象を対応付け，特徴を抽出する．

対応付けられた画像と音の事例の集合を用い，両者の統計的関係を学習するため，まず正準相関分析により画像と音の特徴との相関関係を学習し，正準空間を作成する．この正準空間が概念空間となる．次に，正準空間において，特徴ベクトルに対して教

師なし学習を K-means 法を用いたクラスタリングと情報量を基準としたクラスタの有効性分析によって行い，概念を獲得する．

概念を獲得した後に，カテゴリ未知の入力事例に対応する概念の中心的な事例を提示することを，事例の識別とした．また，カテゴリ未知の入力事例に対応する概念の異なるモダリティにおける中心的な事例を提示することを事例の想起とした．識別および想起の成功は，カテゴリ未知の視聴覚事例と獲得した概念との対応付けが正確である条件を満たすときとした．

概念獲得成功率は 86.4% であり，本実験の識別および想起の成功率の目標値を 80% としたことに対し，識別成功率が 83.2% 以上，想起の成功率が 81.5% 以上となり，目標値を達成している．このことから，正準相関分析において，画像と音の特徴ベクトルの相関係数の大きい成分から順に正準変量を求め，正準空間の次元を減らすことができた．以上の結果から，本手法の有効性が確認された．

第6章　結論

本論文では，物体操作による視聴覚事象の対応付け，選択的注意による視聴覚事象の対応付け，そして視聴覚事象の対応付けに基づく概念の獲得の方法を提案し，その有効性を実験により示した．

対象物に働きかけ，物体の視聴覚事象を能動的に対応付ける方法では，視聴覚事象の対応付けに物体特有の情報ではなく，一般的な法則（ゲシュタルトの群化の法則）を用いる．物体を操作しようとする脳から筋への信号（遠心性）の変化と，その時，視聴覚によって知覚される信号（求心性）である音の変化および映像の変化の“ 同時性 ”，物体操作と視聴覚事象の間における繰り返しの“ 類似性 ”，を手掛かりとして対応付けた．実験では，マニピュレータにより物体を振り，マイクロフォンとカメラを用いて観測し，運動と視聴覚情報の対応付けを行った．

視聴覚情報の対応付けにおける成功率の目標値は，対応付け条件が厳しいため，70% 以上とした．対応付けの成功は，次の 3 条件をすべて満たすときである．(1) 視聴覚事象の対応が正確である，(2) 運動に関する視聴覚事象の相関が最大，(3) 運動方向変化の位置と視聴覚事象の発生位置の対応が正確である．実験の結果，運動情報を用いた視聴覚事象の対応付けの成功率は 73.8% と目標値を達成しており，本手法の有効性を示す．よって，メトロノームのような音や動きがある他の事象が存在する環境であっても，ドラムをたたく，ベルを振るといった物体操作を行うことで，運動に関する視聴覚事象の対応付けができることを確認した．

選択的注意として，聴覚的注意では音源定位および雑音に対して頑健な音オンセット検出を行い，また視覚的注意では局所的に明るさが異なる場面における対象物の検出手法について提案した．

実験の結果，視覚事象と聴覚事象が共に良好に抽出できる場合の視聴覚事象の対応付けの成功率が 86.7% であったことから，注意による視聴覚情報の対応付けにおける成功率の目標値を，85% 以上とした．聴覚的注意において，視覚情報による時間軸上での窓関数を用いて音オンセットを検出した場合，視聴覚事象の対応付けの成功率は，注意を向けないときの 78.6% から 95.2% へ上昇し，聴覚的注意手法の有効性を確認し

た．また視覚的注意において，周辺が明るく，中央部が暗い場所における対象物の視聴覚事象の対応付け実験の成功率は 93.3% であり，このことから局所的に明るさが異なる場面でも視覚的注意により，対象物の視覚情報を検出できることを確認した．これらの技術は，人間のパートナーとして期待されるロボットにとって視聴覚処理の効率化と正確さの向上において有益である．

概念獲得として分類，事例の生成について取り上げ，概念は，共通の特徴を持つ事例の集合体からなり，中心的な事例を概念に属する最も典型性の高い事例として表現した [98]．視聴覚事象についての概念を自律的に獲得するため，物体に対して，1 箇所に固定した 1 個のマイクロフォンと，異なる外観を得るために 12 箇所に順に移動させて配置した 1 台のカメラによって物体を計測する．音の発生，物体の映像上での動きに対して，同時性と類似性の手掛かりを用いて視聴覚事象を対応付け，パターンを抽出する．

対応付けられた画像と音の事例の集合を用い，両者の統計的関係を学習するため，まず正準相関分析により画像と音の特徴との相関関係を学習し，正準空間を作成する．この正準空間が概念空間となる．次に，正準空間において，特徴ベクトルに対して教師なし学習を K-means 法を用いたクラスタリングと情報量を基準としたクラスタの有効性分析によって行い，概念を獲得した．

事前知識なしで概念を獲得した後に，カテゴリ未知の入力事例から対応する概念の同じモダリティにおける中心的な事例を提示することを事例の識別とした．また，カテゴリ未知の入力事例から対応する概念の異なるモダリティにおける中心的な事例を提示することを事例の想起とした．識別および想起の成功は，カテゴリ未知の視聴覚事例と獲得した概念との対応付けが正確であるという条件を満たすときである．

本実験の識別および想起の成功率の目標値を 80% としたことに対し，識別成功率が 83.2% 以上，想起の成功率が 81.5% 以上となり，目標値を達成している．このことから，正準相関分析において，画像と音の特徴ベクトルの相関係数の大きい成分から順に正準変量を求め，正準空間の次元を減らすことができた．以上の結果から，本手法の有効性が確認された．

今後の課題として，以下の事柄が挙げられる．

物体操作による視聴覚事象の対応付けでは，実時間での対応付けや視聴覚事象が不良であるときに物体操作を手掛かりとした視聴覚事象間の対応付けを行うことが考えられる．注意による視聴覚事象の対応付けでは，視聴覚事象のどちらか一方が良好でなければならなかったが，両方とも不良な状況において，視聴覚の相互作用によって

対応付け精度が向上することが望まれる．また，概念の獲得研究については，正準相関分析を用いた本手法と，ニューラルネットワーク等の他手法で獲得した概念の比較や，より多くのカテゴリにおける概念の獲得，概念の追加的な獲得 [99]，および概念の階層構造 [100] をロボットシステムに実装することである．これらの技術により，ロボットは，より複雑な知識を獲得し，人のパートナーとして生活を共にすることができると考えられる．

最後に，本研究が，ロボットの知能化の発展に何らかの貢献となることを期待して，本論文の結びとする．

謝辞

本研究を進めるにあたり，多くの適切な御指導と御助言を頂きました名古屋大学大学院情報科学研究科メディア科学専攻 大西 昇 教授に深く感謝致します．大西先生には，指導教員として本研究を進める機会を与えて頂きました．また，明確な研究指導と，熱心かつ的確な御教示，時宜を得た激励，そして絶え間ない援助を頂いたことにより，本論文をまとめることができました．大西先生の御指導，御鞭撻は，今後の研究活動を行う上での支えとなるものです．ここに，心からの感謝の意を表します．

本論文の作成にあたり，本論文を丁寧に読んで頂き，貴重な御意見と激励を頂いた，名古屋大学大学院情報科学研究科メディア科学専攻 末永 康仁 教授，工藤 博章 准教授に深く感謝致します．

本研究を進めるにあたり，多くの方々から有益な助言と討論を頂いたことに感謝致します．また，名古屋大学情報連携統括本部情報戦略室 竹内 義則 准教授，名古屋大学大学院情報科学研究科メディア科学専攻 松本 哲也 助教，並びに大西研究室の皆様には，熱心な討論と有益な意見を頂きました．本研究で実験を行うにあたり，貴重な時間を割いて実験を手伝って頂き，有益な御助言も多数頂きました．本当にありがとうございました．

本研究の一部は，筆者が平成 16 年度から 19 年度にかけて，文部科学省 21 世紀 COE プログラムにおける名古屋大学情報系 COE“社会情報基盤のための音声・映像の知的統合” および平成 20 年度の名古屋大学情報科学研究科プロジェクト “視聴覚事象の対応付けからの知識の獲得” にて，リサーチアシスタントとして携わったものです．この間，平成 18 年 4 月〜5 月に，末永先生，大西先生のご支援のもとオーストラリア Monash 大学の Digital Perception 研究室に滞在させて頂きました．海外での研究活動に直接触れることができ，Adelaide 大学 Computer Science 研究科 Computer Vision 研究室 David Suter 教授（前 Monash 大学 Electrical and Computer Systems Engineering 研究科 Digital Perception 研究室）をはじめ，Digital Perception 研究室の皆様との議論はとても良い刺激となりました．また滞在中の生活の助言をしてくださった元 COE 研究員の皆様には感謝しています．

最後に，長期間に渡る教育を受けることを快く支え，後押しをくれた両親と家族に心から感謝致します.

付 録A　式(5.5)における線形回帰の証明

確率変数 X，Y が標準化により，平均 $\mu_X = \mu_Y = 0$，分散・共分散 $\sigma_X = \sigma_Y = \sigma_{XY} = \sigma_{YX} = 1$ であるとき，2 次元正規分布の確率密度関数は，次式のように表せる．

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\lambda^2}} \times \quad exp\{-\frac{1}{2(1-\lambda^2)}(x^2 - 2\lambda xy + x^2)\} \tag{A.1}$$

ここで，$\lambda = \sigma_{XY}/\sigma_X\sigma_Y$ は X と Y の相関係数である．このとき，X の周辺確率密度関数は，次式となる．

$$f(x) = \frac{1}{\sqrt{2\pi}} exp\big(-\frac{x^2}{2}\big) \tag{A.2}$$

そして，Y の条件付確率密度関数は次式となる．

$$f(y|x) = \frac{f(x,y)}{f(x)} = \frac{1}{\sqrt{2\pi(1-\lambda^2)}} exp\{-\frac{(y-\lambda x)^2}{2(1-\lambda^2)}\} \tag{A.3}$$

したがって，確率変数 X が $X = x$ であるときの確率変数 Y の期待値は次式で与えられる．

$$E(Y|X=x) \quad = \quad \lambda x = \tilde{y} \tag{A.4}$$

同様に，確率変数 Y が $Y = y$ であるときの確率変数 X の期待値は次式で表せる．

$$E(X|Y=y) = \lambda y = \tilde{x} \tag{A.5}$$

付 録B　概念獲得実験における詳細データ

表 5.2 の近接順位の詳細な結果を**表 B.1** に示す．表中の左上の番号は実験の順番である．7 番目の実験で，対応するクラスが生成されなかったものは“ - ”で示す．

表 B.1: 近接順位の詳細データ
Table B.1 Detail of nearness ranks.

Trial	Object	Audio-Visual	Audio	Visual
1	Bell	1	2	8
	Drum	1	1	3
	Met.	1	1	5
2	Bell	1	2	5
	Drum	1	1	1
	Met.	1	1	1
3	Bell	3	4	6
	Drum	1	1	4
	Met.	1	1	3
4	Bell	2	3	6
	Drum	1	1	3
	Met.	1	1	1
5	Bell	2	1	6
	Drum	1	1	4
	Met.	1	4	4
6	Bell	2	1	9
	Drum	6	1	1
	Met.	1	1	1
7	Bell	-	16	-
	Drum	26	-	32
	Met.	-	1	-
8	Bell	2	1	4
	Drum	1	1	2
	Met.	1	1	1
9	Bell	1	1	5
	Drum	1	1	3
	Met.	1	1	1
10	Bell	1	1	7
	Drum	1	1	1
	Met.	1	1	1
Ave	Bell	1.67	3.20	6.22
	Drum	4.00	1.00	5.40
	Met.	1.00	1.30	2.00

表 5.3 に示した概念の識別結果の詳細を**表 B.2** と**表 B.3** に示す．**表 B.2** は，識別実験における Closed テストの結果の詳細である．

表 B.2: 識別成功率の詳細データ（Closed テスト）
Table B.2 Detail of success rate of identification (Closed test)

(a)OFK

Trial	Audio-Visual	Audio	Visual
1	99.5%	98.4%	96.2%
2	100.0%	97.3%	95.6%
3	98.9%	97.3%	94.0%
4	98.9%	97.3%	95.1%
5	98.9%	97.3%	95.6%
6	98.9%	97.3%	95.1%
7	99.5%	97.8%	92.9%
8	98.9%	97.3%	94.5%
9	98.9%	97.3%	94.5%
10	98.9%	97.3%	95.6%
Ave	99.1%	97.4%	94.9%

(b)OFC

Trial	Audio-Visual	Audio	Visual
1	33.3%	33.3%	20.2%
2	33.3%	0.0%	40.4%
3	60.7%	0.0%	72.1%
4	33.3%	33.3%	20.8%
5	33.3%	0.0%	19.7%
6	33.3%	33.3%	21.3%
7	33.3%	0.0%	26.8%
8	33.3%	33.3%	72.1%
9	33.3%	0.0%	20.8%
10	33.3%	33.3%	20.2%
Ave	36.1%	16.7%	33.4%

(c)CCAC

Trial	Audio-Visual	Audio	Visual
1	96.7%	95.1%	92.4%
2	97.8%	95.1%	94.5%
3	95.6%	95.6%	94.0%
4	98.4%	96.2%	92.9%
5	96.2%	96.7%	93.4%
6	98.4%	97.3%	93.4%
7	97.3%	97.3%	92.4%
8	96.7%	96.2%	95.1%
9	95.6%	96.7%	92.9%
10	96.7%	95.6%	94.0%
Ave	96.9%	96.2%	93.5%

(d)CCAN

Trial	Audio-Visual	Audio	Visual
1	95.1%	95.1%	92.4%
2	97.8%	96.2%	94.0%
3	94.0%	94.5%	92.4%
4	96.7%	96.2%	92.9%
5	95.6%	96.7%	92.9%
6	96.7%	97.3%	92.9%
7	0.0%	0.0%	91.3%
8	97.3%	96.2%	95.1%
9	93.4%	96.2%	91.3%
10	97.8%	96.2%	94.0%
Ave	86.4%	86.4%	92.9%

(e)CCAK

Trial	Audio-Visual	Audio	Visual
1	95.1%	95.1%	92.4%
2	97.3%	96.2%	94.0%
3	94.0%	94.5%	92.4%
4	96.7%	96.2%	92.9%
5	95.6%	96.7%	92.4%
6	96.7%	97.3%	92.9%
7	0.0%	33.3%	0.0%
8	97.3%	96.2%	95.1%
9	93.4%	96.2%	91.3%
10	97.8%	96.2%	94.0%
Ave	86.4%	89.8%	83.7%

(f)CCAAV

Trial	Audio-Visual	Audio	Visual
1	86.3%	95.1%	93.4%
2	91.8%	96.7%	94.5%
3	85.2%	97.3%	33.3%
4	88.0%	95.6%	93.4%
5	87.4%	95.1%	92.9%
6	88.5%	96.7%	92.4%
7	0.0%	0.0%	0.0%
8	90.7%	96.7%	30.6%
9	89.1%	96.2%	33.3%
10	86.9%	95.6%	94.0%
Ave	79.4%	86.5%	65.8%

また，識別実験における Open テスト結果の詳細を**表 B.3** に示す.

表 B.3: 識別成功率の詳細データ（Open テスト）
Table B.3 Detail of success rate of identification (Open test)

(a)OFK

Trial	Audio-Visual	Audio	Visual
1	94.0%	89.1%	94.0%
2	94.0%	88.0%	94.0%
3	92.4%	87.4%	93.4%
4	95.6%	89.6%	94.5%
5	94.5%	90.2%	93.4%
6	92.9%	90.7%	92.9%
7	94.0%	89.6%	91.8%
8	95.1%	89.6%	94.5%
9	95.1%	89.1%	94.0%
10	96.7%	90.7%	95.1%
Ave	94.4%	89.4%	93.8%

(b)OFC

Trial	Audio-Visual	Audio	Visual
1	32.8%	31.7%	18.6%
2	31.1%	0.0%	44.3%
3	55.7%	0.0%	77.6%
4	32.8%	31.7%	18.6%
5	31.7%	0.0%	18.0%
6	32.8%	31.7%	19.7%
7	32.8%	0.0%	72.7%
8	32.8%	31.1%	76.0%
9	32.8%	0.0%	18.0%
10	33.3%	31.7%	18.0%
Ave	34.9%	15.8%	38.1%

(c)CCAC

Trial	Audio-Visual	Audio	Visual
1	94.5%	91.3%	92.4%
2	97.8%	93.4%	93.4%
3	95.1%	91.8%	94.5%
4	94.5%	91.3%	91.3%
5	94.5%	92.9%	91.3%
6	95.1%	91.8%	91.8%
7	93.4%	91.8%	88.5%
8	98.4%	94.5%	93.4%
9	94.0%	91.8%	91.3%
10	95.1%	91.8%	91.3%
Ave	95.2%	92.2%	91.9%

(d)CCAN

Trial	Audio-Visual	Audio	Visual
1	94.5%	91.3%	92.4%
2	97.3%	93.4%	93.4%
3	95.6%	90.2%	95.1%
4	95.6%	90.2%	90.2%
5	97.8%	92.9%	90.2%
6	97.8%	91.8%	92.4%
7	0.0%	0.0%	89.1%
8	97.8%	94.5%	92.9%
9	96.7%	91.3%	92.9%
10	97.3%	91.3%	92.4%
Ave	87.0%	82.7%	92.1%

(e)CCAK

Trial	Audio-Visual	Audio	Visual
1	94.5%	91.3%	91.8%
2	98.9%	93.4%	93.4%
3	95.6%	90.2%	95.1%
4	95.6%	90.2%	90.2%
5	97.8%	92.9%	91.3%
6	97.8%	91.8%	92.4%
7	0.0%	33.3%	0.0%
8	97.8%	94.5%	92.9%
9	96.7%	91.3%	92.9%
10	97.3%	91.3%	92.4%
Ave	87.2%	86.0%	83.2%

(f)CCAAV

Trial	Audio-Visual	Audio	Visual
1	91.3%	100.0%	93.4%
2	96.7%	98.9%	97.3%
3	92.9%	98.9%	33.3%
4	92.9%	98.9%	95.6%
5	92.9%	98.9%	97.3%
6	95.6%	98.9%	95.1%
7	0.0%	0.0%	0.0%
8	95.1%	98.9%	32.8%
9	94.5%	99.5%	32.8%
10	90.2%	98.9%	95.1%
Ave	84.2%	89.2%	67.3%

表5.4 に示した想起実験の Closed テストにおける結果の詳細を表 **B.4** に示す．また，Open テストにおける結果の詳細を**表 B.5** に示す．

表 B.4: 想起成功率の詳細データ（Closed テスト）
Table B.4 Detail of success rate of association (Closed test)

(a)CCACAS

Trial	Audio-Visual	Audio	Visual
1	96.7%	92.4%	96.2%
2	97.8%	94.5%	94.5%
3	95.6%	94.0%	96.2%
4	98.4%	94.0%	97.3%
5	96.2%	94.0%	97.3%
6	98.4%	92.9%	97.3%
7	97.3%	92.4%	96.7%
8	96.7%	94.5%	96.2%
9	95.6%	92.9%	97.3%
10	96.7%	94.0%	96.2%
Ave	96.9%	93.6%	96.5%

(b)CCAKAS

Trial	Audio-Visual	Audio	Visual
1	95.1%	91.8%	94.0%
2	97.3%	91.8%	94.0%
3	94.0%	92.9%	94.5%
4	96.7%	93.4%	95.6%
5	95.6%	93.4%	95.1%
6	96.7%	93.4%	96.7%
7	0.0%	33.3%	0.0%
8	97.3%	93.4%	95.6%
9	93.4%	92.4%	95.6%
10	97.8%	93.4%	95.1%
Ave	86.4%	86.9%	85.6%

表 B.5: 想起成功率の詳細データ（Open テスト）
Table B.5 Detail of success rate of association (Open test)

(a)CCACAS

Trial	Audio-Visual	Audio	Visual
1	94.5%	91.8%	91.8%
2	97.8%	94.0%	93.4%
3	95.1%	94.5%	91.8%
4	94.5%	91.3%	89.6%
5	94.5%	90.2%	92.9%
6	95.1%	91.8%	92.4%
7	93.4%	89.6%	90.7%
8	98.4%	93.4%	94.0%
9	94.0%	91.8%	92.4%
10	95.1%	91.8%	89.6%
Ave	95.2%	92.0%	91.9%

(b)CCAKAS

Trial	Audio-Visual	Audio	Visual
1	94.5%	91.3%	89.6%
2	98.9%	91.3%	92.9%
3	95.6%	95.1%	90.2%
4	95.6%	91.8%	88.0%
5	97.8%	89.6%	91.8%
6	97.8%	91.8%	91.3%
7	0.0%	31.1%	0.0%
8	97.8%	92.4%	92.9%
9	96.7%	92.9%	90.7%
10	97.3%	90.7%	88.0%
Ave	87.2%	85.8%	81.5%

参考文献

[1] E. Spelke, “Infants’ Intermodal perception of events,” Cognitive Psychology 8, pp. 553–560, 1976.

[2] 正高信男，“脳から心へ (宮下保司, 下条信輔編) 第 7 章 身ぶりの行動発達学”, 岩波書店, pp. 341–349, 1995.

[3] K. Ejiri and N. Masataka, “Co-occurrence of preverbal vocal behavior and motor action in early infancy,” Developmental Science 4, pp. 40–48, 2001.

[4] 国際電気通信基礎技術研究所編，“視聴覚情報科学-人間の認知の本質にせまる”, オーム社, 1994.

[5] 椎名健，“心理学パッケージ Part 4(小川捷之, 椎名健編) 第 3 章 カクテル・パーティー効果”, ブレーン出版, pp. 25–31, 1984.

[6] M. Grestya, “Coordination of head and eye movements to fixate continuous and intermittent targets,” Vision Res., Vol. 14, pp. 395–403, 1974.

[7] 山田光穂，“2 次元平面上の視標を注視させたときの頭部運動と眼球運動の協調関係の分析”, 電子情報通信学会論文誌 D-II, Vol. J75-D-II, No. 5, pp. 971–981, 1992.

[8] P. Kuhl and A. Meltzoff, “The Bimodal Perception of Speech in Infancy,” Science, Vol. 218, No. 4577, pp. 1138–1141, 1982.

[9] J. Saffran, R. Aslin and E. Newport, “Statistical Learning by 8-Month-Old Infants,” Science, Vol. 274. No. 5294, pp. 1926–1928, 1996.

[10] 早川和宏，鈴木亮，向井利春，大西昇，“物理法則に基づく視聴覚情報の対応付け”，電子情報通信学会技術研究報告 EID98-147, pp. 13–18, 1999.

[11] J. Chen, T. Mukai, Y. Takeuchi, T. Matsumoto, H. Kudo, T. Yamamura and N. Ohnishi, “Finding the Correspondence of Audio-Visual Events Caused by Multiple Movements,” The Journal of The Institute of Image Information and Television Engineers, Vol. 55, No. 11, pp. 1450–1459, 2001.

[12] J. Chen, T. Mukai, Y. Takeuchi, T. Matsumoto, H. Kudo, T. Yamamura and N. Ohnishi, “Relating Audio-Visual Events Caused by Multiple Movements, In the Case that a Moving Object is Out of Camera View,” The Journal of The Institute of Image Information and Television Engineers, Vol. 58, No. 12, pp. 1828–1834, 2004.

[13] K. Nakadai, H. Okuno and H. Kitano, “Robot Recognizes Three Simultaneous Speech By Active Audition,” Proceedings of IEEE Conference on Robotics and Automation, pp. 398–403, 2003.

[14] P. Aarabi and S. Zaky, “Robust Sound Localization Using Multi-Source Audiovisual Information Fusion,” Information Fusion, Vol. 2, No. 3, pp. 209–223, 2001.

[15] 赤松幹之, “視覚と触覚と運動の統合”, 電気情報通信学会誌, Vol. 76, No.11, pp. 1176–1182, 1993.

[16] 橋本浩一, “視覚と制御”, 計測自動制御学会制御部門大会ワークショップ 制御部門大会ワークショップテキスト, pp. 37–68 , 2001.

[17] 高橋弘太, 来海暁, 山口佳子, 山崎弘郎, “聴覚－視覚－聴覚を階層的に融合する知能化センサ”, 計測自動制御, Vol. 27, No. 3, pp. 1127–1137, 1994.

[18] 陶山健仁, 高橋弘太, 岩倉博, “2 段階のデータ選別による複数音源定位”, 電子情報通信学会論文誌 A, Vol. J79-A, No. 6, 1127–1137, 1996.

[19] L. Itti, C. Koch and E. Niebur, “A Model of Saliency-based Visual Attention for Rapid Scene Analysis,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, No. 11, pp. 1254–1259, 1998.

[20] H. Tagare, K. Toyama and J. G. Wang, “A Maximum-Likelihood Strategy for Directing Attention during Visual Search,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23, No. 5, pp. 490–500, 2001.

[21] 陳彬，金子正秀，“複数の異なるモダリティ情報の統合に基づく移動ロボットの行動選択”, 電気学会論文誌 C 分冊, Vol. 125-C, Np. 5, pp. 765–773, 2005.

[22] 港隆史，浅田稔，“注視機構実現に向けた視覚-行動学習による画像特徴と状態空間の構成”，日本ロボット学会誌，Vol. 21，No. 1，pp. 87-93, 2003.

[23] 長井志江，細田耕，森田章生，浅田稔，“視覚注視と自己評価型学習の機能に基づくブートストラップ学習を通した共同注意の創発”，人工知能学会論文誌，Vol. 19，No. 1，pp. 10–19, 2004.

[24] 大西正輝, 影林岳彦, 福永邦雄，“視聴覚情報の統合による会議映像の自動撮影”, 電子情報通信学会論文誌 D-II, Vol. J85-D-II, No. 3, pp. 537–542, 2002.

[25] 川野隆亮, 竹内義則, 大西昇，“視聴覚情報を用いた複数人物の追跡”, 電子情報通信学会技術研究報告 PRMU2002–263, pp. 123–128, 2003.

[26] 中川聖一，中西宏文，古部好計，板橋光義，“ 視聴覚情報の統合化に基づく概念の獲得”，人工知能学会誌，Vol. 8, No. 4，pp. 499–508, 1993.

[27] 赤穂昭太郎，速水悟，長谷川修，吉村隆，麻生英樹，“EM 法を用いた複数情報源からの概念獲得”，電子情報通信学会論文誌 A，Vol. J80-A，No. 9，pp. 1546–1553, 1997.

[28] D. Roy，“Learning Visually Grounded Words and Syntax of Natural Spoken Language,” Evolution of Communication, 2001.

[29] N. Iwahashi，“A Method for the Coupling of Belief Systems through Human-Robot Language Interaction,” Proceedings of the 12th IEEE Workshop Robot and Human Interactive Communication, pp. 385–390, 2003.

[30] 小島量，長谷川修，“ヒューマノイドロボット上の自己増殖型ニューラルネットワークを用いた視聴覚情報からの能動的・追加的概念獲得”，電子情報通信学会技術研究報告 PRMU2005-57， pp. 35–40, 2005.

[31] J. Clark and A. Yuille,“Data Fusion for Sensory Information Processing Systems,” Kluwer Academic Publishers, 1990.

[32] 山崎弘郎, 石川正俊，“センサフュージョン,” コロナ社, 1992.

[33] L. Klein, “Sensor and Data Fusion: A Tool for Information Assessment and Decision Making,” SPIE Press, 2004.

[34] 大西昇, 杉江昇, “生体情報処理”, 昭晃堂, 2001

[35] 行場次朗, “人工知能学事典 (人工知能学会編)　第 8 章 画像・音声メディア 8-e ゲシュタルト理論”, 共立出版, 2005.

[36] D. Katz, “Gestalt Psychology : Its Nature and Significance,” Robert Tyson, 1951.

[37] 大賀寿郎, 山崎芳男, 金田豊, “ 音響システムとディジタル処理”, 電子情報通信学会, 1995.

[38] 安藤彰男, “リアルタイム音声認識”, 電子情報通信学会, 2003.

[39] M. Hu, “Visual Pattern Recognition by Moment Invariants,” IRE Transactions on Information Theory, Vo. 8, No. 2, pp. 179–187, 1962.

[40] 大津展之, 関田巌, 栗田多喜夫, “パターン認識—理論と応用 (行動計量学シリーズ)”, 朝倉書店, 1996.

[41] 奥野忠一, 久米均, 芳賀敏郎, 吉澤正, “多変量解析法”, 日科技連出版社, 1971.

[42] 柳井晴夫, 高木廣文, “多変量解析ハンドブック”, 現代数学社, 1986.

[43] J. Hartigan, “Clustering algorithms, Chapter 4, The K-means algorithm,” John Wiley & Sons Inc, 1975.

[44] 玉木徹, 山村毅, 大西昇, “画像系列からの人物領域の抽出”, 電気学会論文誌 C 分冊, Vol. 119-C, No. 1, pp. 37–43, 1999.

[45] 池田徹志, 石黒浩, 浅田稔, “相互情報量最大化に基づく信号情報源の移動軌跡の推定”, 電子情報通信学会論文誌 D, Vol. J90-D, No. 2, pp. 535–543, 2007.

[46] H. Nock, G. Iyengar, C. Neti, “Multimodal processing by finding common cause,” Communications of the ACM, Vol. 47, No. 1, 2004.

[47] 向井利春, 石川正俊, “複数センサによる予測誤差を用いたアクティブセンシング”, 日本ロボット学会誌, pp. 715–721, 1994.

[48] 石川正俊，“センサフュージョンの課題”, 日本ロボット学会誌, Vol. 8, No. 6, pp. 735–742, 1990.

[49] 片山正純, 川人光男，“触覚，体性感覚と運動司令を統合する神経回路モデル”, 日本ロボット学会誌, Vol. 8, No. 6, pp. 117–125, 1990.

[50] P. Haggard and B. Whitford, “Supplementary motor area provides an efferent signal for sensory suppression,” Cognitive Brain Research Vo. 19, No. 1, pp. 52–58, 2004.

[51] 吉川雄一郎，細田耕，浅田稔，辻義樹，“複数センサ データの不変性に基づく身体の発見”, 日本ロボット学会誌, Vol. 23, No. 8, pp. 986–992, 2005.

[52] 嶋田総太郎, 開一夫，“自己身体認識 における視覚と体性感覚の時間的整合性について”, 電子情報通信学会技術研究報告, NC2005-42, pp. 33–38, 2005.

[53] A. Klapuri, “Sound onset detection by applying psychoacoustic knowledge,” Proceedings of IEEE International Conference on Acoustics, Speech, and Signal (ICASSP), pp. 3089–3092, 1999.

[54] J. Tranter 著，山形浩生訳，“LINUX マルチメディアガイド”，オライリー・ジャパン，1997.

[55] 飯尾淳, “Linux による画像処理プログラミング”, オーム社，2000.

[56] K. Nishibori, J. Chen, Y. Takeuchi, T. Matsumoto, H. Kudo and N. Ohnishi, “Determining Correspondences between Sensory and Motor Signal,” Proceedings of Pacific-Rim Conference on Multimedia, Part 1, pp. 532–539, 2004.

[57] K. Nishibori, Y. Takeuchi, T. Matsumoto, H. Kudo and N. Ohnishi, “An Attentional Correspondence of Audio-Visual Events,” Proceedings of International Workshop on Advanced Image Technology, pp. 151–156, 2005.

[58] Kento Nishibori, Yoshinori Takeuchi, Tetsuya Matsumoto, Hiroaki Kudo, Noboru Ohnishi, “Determination of Correspondence between Audio and Visual Events through active motion,” Proceedings of International Workshop on Advanced Image Technology, pp. 620–625, 2007.

[59] 西堀研人，竹内義則，松本哲也，工藤博章，大西昇，“物体操作による視聴覚事象の対応付け”, 電気学会論文誌 C 分冊，Vol. 128-C, No. 2, pp. 242–252, 2008.

[60] G. Meyer and J. Mulligan, “Continuous audio-visual digit recognition using N-best decision fusion,” Information Fusion, Vol. 5, No. 2, pp. 91–101, 2004.

[61] M. Jägersand, “Saliency maps and attention selection in scale and spatial coordinates: an information theoretic approach,” ICCV ’95: Proceedings of the Fifth International Conference on Computer Vision, IEEE Computer Society, pp. 195–202, 1995.

[62] L. Garcia, A. Oliveira, R. Grupen, D. Wheeler and A. Fagg, “Tracing patterns and attention, humanoid robot cognition,” IEEE Transactions on Intelligent Systems, Vol. 15, No. 4, pp. 70–77, 2000.

[63] S. Vijayakumar, J. Conradt, T. Shibata and S. Schaal, “Overt visual attention for a humanoid robot,” Proceedings of International Conference on Intelligence in Robotics and Autonomous Systems (IROS 2001), 2001.

[64] D. Zotkin, R. Duraiswami, H. Nanda and L. Davis, “Multimodal Tracking for Smart Videoconferencing,” Proceedings of the 2nd International Conference on Multimedia and Expo, 2001.

[65] 熊田孝恒，菊池正，“脳から心へ (宮下保司, 下條信輔編) 第 4 章 注意とは何か”, 岩波書店, pp. 98–106, 1995.

[66] P. Aarabi and S. Mavandadi, “Robust sound localization using conditional time-frequency histograms,” Information Fusion, Vol. 4, No. 2 pp. 111–122, 2003.

[67] 中川聖一，“音声認識研究の動向”, 電子情報通信学会論文誌 D-II, Vol. J83-D-II, No. 2, pp. 433–457, 2000.

[68] A. Bahill, D. Adler and L. Stark, “Most naturally occurring human saccades have magnitudes of 15 degrees or less,” Investigative Ophthalmology & Visual Science, Vol. 14, pp. 468–469, 1975.

[69] 竹内義則, 大西昇, 杉江昇，“情報理論に基づいたアクティブビジョンシステム”, 電子情報通信学会論文誌 D-II, Vol. J81-D-II, No. 2, pp. 323–330, 1998.

[70] 今井秀樹，“情報理論”, 昭晃堂, 1984.

[71] 藤崎和香，“聴覚情報処理のフロンティア研究と情報通信技術への応用 [II] 視聴覚の情報統合と同時性知覚”, 電気情報通信学会誌, Vol. 89, No. 10, 2006.

[72] 中島弘道, 大西昇, 向井利春，“スペクトルの特徴マップを用いた上下方向音源定位学習システム”, 電子情報通信学会論文誌 D-II, Vol. J87-D-II, No. 11, pp. 2034–2044, 2004.

[73] 小濱剛，新開憲，臼井支朗，“マイクロサッカードの解析に基づく視覚的注意の定量的測定の試み”，映像情報メディア学会誌，Vol. 52，No. 4, pp. 571–576, 1998.

[74] 内川惠二総編集，塩入諭編，“視覚 II 視覚系の中期・高次機能 第 8 章視覚的注意”，朝倉書店, 2007.

[75] 三浦利章，“行動と視覚的注意”，風間書房, 1996.

[76] 横澤一彦，“視覚的注意の基礎”，心理学評論，Vol. 46，No. 3，pp. 353–356, 2003.

[77] 熊田孝恒，“視覚的注意とは何か？ ―いくつかの基本的な概念を中心として”，VISION，8，pp. 195–198, 1996.

[78] 日本視覚学会編，“視覚情報処理ハンドブック 11 章 視覚的注意”，朝倉書店, 2000.

[79] 佐藤俊治，三宅章吾，“スケールスペース理論に基づく注視モデル”，電子情報通信学会論文誌 D-II, Vol. J86-D-II, No. 10, pp. 1490–1501, 2003.

[80] 安西祐一郎，苧坂直行，前田敏博，彦坂興秀，“注意と意識 第 3 章 注意の神経機構”，岩波書店, 1996.

[81] 下中邦彦編，“心理学事典”，平凡社, 1979.

[82] 中川聖一，升方幹雄，“視聴覚情報の統合化に基づく概念と文法の獲得システム”，人工知能学会誌, Vol. 10, No. 4, pp. 619–627, 1995.

[83] 栗田多喜夫, 加藤俊一, 福田郁美, 坂倉あゆみ，“印象語による絵画データベースの検索”, 情報処理学会論文誌, Vol. 33, No. 11, pp. 1373–1383, 1992.

[84] 井手一郎，浜田玲子，坂井修一，田中 英彦，“言語情報を伴う画像の画像的特徴量と語義の統計的対応付け,” Vol. 99，No. 3，情報処理学会研究報告，CVIM, pp. 137–143, 1999.

[85] 山川宏，“パターンベースド知能システム - 学習から見たシンボルグラウンディング問題の検討-”, RWC 情報統合ワークショップ ’95, pp. 167–175, 1995.

[86] 大須賀節雄，佐伯胖，“知識の獲得と学習”，オーム社，1987.

[87] R. Michalski，電総研人工知能研究グループ，“概念と規則の学習”, 共立出版, 1988.

[88] T. Maia and N. Chang, “Grounding the Acquisition of Grammar in Sensorimotor Representations,” Proceedings of AAAI Spring Symposium on Learning Grounded Representations, 2001.

[89] C. Wendelken and L. Shastri,“Acquisition of concepts and causal rules in SHRUTI,” Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society, 2003.

[90] 石黒勝彦，大津展之，國吉康夫，“音声・画像入力からの概念獲得のためのインターモーダル学習”，電子情報通信学会技術研究報告, PRMU, Vol. 104, No.370, pp. 17–24, 2004.

[91] 西堀研人，竹内義則，松本哲也，工藤博章，大西昇，“生物に示唆を得た選択的注意による視聴覚事象の対応付け手法”，映像情報メディア学会誌, Vol. 62，No. 7, pp. 1086–1097, 2008.

[92] 河原哲雄，“人工知能学事典 (人工知能学会編)　第 2 章　知の基礎科学　2-20 概念とカテゴリ”, 共立出版, 2005.

[93] E. Rosch and B. Lloyd, “Cognition and Categorization,” L. Erlbaum Associates, 1978.

[94] R. Michalski, “Knowledge acquisition through conceptual clustering : A theoretical framework and algorithm for partitioning data into conjunctive concepts,” International Journal of Policy Analysis and Information Systems, Vol. 4, pp. 219–243, 1980.

[95] 椹木哲夫, “概念クラスタリング”, 日本ファジィ学会誌, Vol. 8, No. 3, pp. 463–467, 1996.

[96] 神嶌敏弘，“人工知能学事典 (人工知能学会編)　第 5 章　機械学習　5-4　概念クラスタリング”, 共立出版, 2005.

[97] 青木直史，“ディジタル・サウンド処理入門”，CQ 出版，2006.

[98] 西堀研人，竹内義則，松本哲也，工藤博章，大西昇，“視聴覚事象の対応付けによる概念の獲得,” 情報学ワークショップ, pp. 7–12, 2008.

[99] D. Fisher, “Knowledge Acquisition Via Incremental Conceptual Clustering,” Machine Learning, Vol. 2, No. 2, pp. 139-172, 1987.

[100] L. Talavera, J. Bejar, “Generality-based conceptual clustering with probabilistic concepts,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23, No. 2, pp. 196–206, 2001.