深層学習に基づく音源情報推定のための確率論的目的関数の研究

電気通信大学大学院 情報理工学研究科
博士（工学）の学位申請論文

小泉 悠馬

2017年9月

深層学習に基づく音源情報推定のための確率論的目的関数の研究

博士論文審査委員会

主査 羽田 陽一 教授

委員 南 泰浩 教授

委員 柳井 啓司 教授

委員 庄野 逸 教授

委員 橋本 直己 准教授

A Research on the Design of Statistical Objective Functions for Estimating Acoustic Information using Deep Learning

Yuma Koizumi

Abstract

This research aims to estimate "acoustic information", which is information related to a target sound source such as the source signal, the direction and the state, from the acoustic signal observed by microphones. Two engineering problems are investigated in this thesis: "sound source enhancement" which is the source signal estimation problem, and "anomalous sound detection" which is the state estimation problem.

In recent years, deep learning (DL) has been applied to acoustic information estimation and estimation accuracy has been greatly improved. In DL-based approach, a neural network is used as a nonlinear mapping function from observation signals to target acoustic information. In most cases, the neural network is trained with supervised learning; the neural network is trained to maximize/minimize the "objective function" which evaluates the estimation accuracy of acoustic information such as mean squared error (MSE).

The design of the objective function in acoustic information estimation is equivalent to defining the properties and the estimation accuracy of the target acoustic information. Therefore, it is difficult to estimate acoustic information which cannot be defined its properties and estimation accuracy with a deterministic objective function such as squared error and cross-entropy. For example, the target source which maximizes perceptual sound quality cannot be estimated, because MSE-based objective function does not necessarily improve perceptual quality. As another example, anomalous state of the sound source cannot be also estimated, because anomalous sound due to danger situation rarely occurs and it is difficult to collect label data.

We deal through this paper with the problem of designing the objective function to estimate acoustic information with deep learning. We tackle to design objective functions by defining the statistical properties of the target acoustic information such as probability density function.

In Chapter 3, to estimate target source with a small deep neural networks (DNNs), an objective function to select informative acoustic-features for collecting target sources in noisy environments is proposed. Wiener filtering is a powerful framework for sound-source enhancement. In order to estimate Wiener-filter with a small DNNs, it is essential to find informative acoustic features that provide effective cues for Wiener-filter estimation. In this study, we measured the informative-ness of acoustic features using mutual information between acoustic features and supervised Wiener-filter parameters, e.g., prior signal-to-noise ratios, and developed a method for automatically selecting informative acoustic features from a large number of feature candidates. To automatically select optimum features, we derived a differentiable objective function in proportion to mutual information based on the kernel method. Since the higher-order correlations between acoustic features and Wiener-filter parameters are calculated using the kernel method, the statistical dependence of these variables is accurately calculated; thus, only meaningful acoustic features are selected. Through several experiments conducted on a mock sports field, we confirmed that the signal-to-distortion ratio score improved when various types of target sources were surrounded by loud cheering noise.

In Chapter 4, to improve sound quality of output signals of DNN-based sound source enhancement, an objective function to maximize sound quality measure is proposed. Conventional DNN-based sound source enhancement is trained with mean-squared-error based objective function. Since MSE-based objective function does not necessarily improve perceptual quality, the quality of output signal is degraded. In Chapter 4, we apply a quantitative metric that reflects a human's perceptual score (perceptual score) to objective function, instead of explicitly giving label data. In the objective tests, we confirmed that DNN was trained to maximize audibility perceptual score by using proposed objective function. In the subjective tests, it was confirmed that the output sound quality of proposed method outperformed the conventional method. These result suggest that the proposed method enable to train DNN with various quantity metrics such as auditory score which could not be used in conventional objective function.

In Chapter 5, to detect anomalous sound whose label data cannot be collected, an objective function for anomalous sound detection (ASD) is proposed. Most ASD systems adopt outlier-detection techniques because it is difficult to collect a massive amount of anomalous sound data. To improve the performance of such outlier-detection-based ASD, it is essential to extract a set of efficient acoustic features that is suitable for identifying anomalous sounds. However, the ideal property of a set of acoustic features that

maximizes ASD performance has not been clarified. By considering outlier detection-based ASD as a statistical hypothesis test, we defined optimality as an objective function that adopts Neyman-Pearson lemma; the acoustic feature extractor is optimized to extract a set of acoustic features which maximize the true positive rate under an arbitrary false positive rate. The variational auto-encoder is applied as an acoustic feature extractor and optimized to maximize the objective function. We confirmed that the proposed method improved the F-measure score from 0.02 to 0.06 points compared to those of conventional methods, and ASD results of some real-environment show that the proposed method is effective in identifying anomalous sounds.

深層学習に基づく音源情報推定のための確率論的目的関数の研究

小泉 悠馬

要旨

本研究は，マイクロホンで観測した音響信号から，源信号や音源の種類や状態などの音に関係する情報である「音源情報」を推定する研究である．音源情報推定の題材として，源信号と雑音が重畳した観測信号から源信号を推定する「音源強調」と，観測信号に含まれる環境音の種類や状態を推定して周囲の危険を予測/察知する「異常音検知」に焦点を当てる．音源の種類や状態などの潜在的な音源情報を考慮しながら音源強調ができれば，大歓声に包まれたサッカースタジアムで，特定の選手の声やボールのキック音を推定でき，まるでサッカースタジアムに潜り込んだようなコンテンツ視聴の方法をユーザに提供可能になる．観測信号に含まれる環境音の種類や状態を推定する異常音検知が実現すれば，機器の動作音から，その機器の動作が正常か異常か（状態）を推定できるようになり，製造/保守業務の効率化ができる．

音源情報を推定するための手法として，統計的機械学習に基づくアプローチが研究されており，近年では深層学習を音源情報推定に適用することで，その推定精度が大きく向上している．深層学習に基づく音源情報推定では，ニューラルネットワークを観測信号から所望の音源情報への非線形写像関数として用いる．そしてニューラルネットワークを音源情報の推定精度を評価する「目的関数」の値を最大化/最小化するように求める．多くの深層学習において目的関数には，二乗誤差関数や交差エントロピー関数などの決定論的な目的関数が用いられる．

音源情報推定において目的関数の設計とは，所望の音源情報の性質や推定精度を定義することと等価である．音源情報の中は，決定論的な目的関数では音源情報の性質や推定精度を定義できないものや，もしくは定義することが妥当ではないものも存在する．例えば，人間の主観的な音質評価を最大化する源信号や，異常音（ラベルデータ）が収集できない音源の状態の推定のための目的関数には，決定論的な目的関数は採用できない．この問題を解決するためには，ネットワークの構造だけでなく，ニューラルネットワークの学習に用いる目的関数を高度化しなくてはならない．

本研究では，決定論的な関数で目的関数を設計できない音源情報を推定するために，深層学習に基づく音源情報推定のための目的関数の研究を行う．所望の音源情報の性質

や推定精度を，推定したい音源情報の特性や解きたい問題に応じて入出力値がとるべき値の確率分布や集合として定義し，ニューラルネットワークの入出力が満たすべき統計的な性質を目的関数として記述するという着想からこの問題に取り組む．

3 章では，スポーツの競技音など，ラベルデータが十分に存在しない源信号を強調するための手法を提案する．少量の学習データでニューラルネットワークを学習するためには，事前に設計/選択した音響特徴量を観測信号から抽出し，小規模なニューラルネットワークで音源強調を行う必要がある．3 章では，所望の音源を強調するための適切な音響特徴量を，相互情報量最大化に基づき選択する方法を検討した．この際，特徴量候補の次元数が大きい音響特徴量選択に相互情報量を正確に計算する “カーネル次元圧縮法” を適用することを考え，スパース正則化法に基づく微分可能な目的関数を導出し，大量な音響特徴量候補から適切な音響特徴量を勾配法により選択できる音響特徴量選択法を提案した．定量評価試験では，従来の音響特徴量選択法と比べ SDR が向上することを示し，また主観評価試験では，提案法を用いて音響特徴量を選択することで従来法と比べ源信号の明瞭性が向上することを示した．この成果により，これまで推定が困難とされていた，学習データが十分に得られないような源信号や，これまで源信号の推定対象とされてこず，適切な音響特徴量が未知な源信号も推定できるようになった．

4 章では，音源強調の出力音の主観品質を向上させるために，ラベルデータを一意に定めることができず，二乗誤差などの目的関数で推定精度を定義することが妥当でない源信号を強調するための手法を提案する．従来の深層学習に基づく音源強調では，源信号の振幅スペクトルなどをラベルデータとし，ニューラルネットワークの出力とラベルデータの二乗誤差を最小化するように学習をしてきた．このため，出力音に歪が生じて主観品質が低下するという問題があった．そこで 4 章では，ラベルデータを用意する代わりに主観評価値と相関の高い音質評価値（聴感評点）を最大化するようための目的関数を提案した．定量評価試験では，提案する目的関数を利用することで，聴感評点を最大化するようにニューラルネットワークを学習できることを確認した．また主観評価試験では，提案法は従来の二乗誤差最小化に基づく目的関数を利用した音源強調よりも高い主観品質で音源強調できることを示した．この成果により，これまで音源強調の学習に利用できなかった聴感評点や人間の評価などの，より “高次” な評価尺度を目的関数として利用できるようになり，ニューラルネットワークを用いた音源強調の応用範囲を広げることができる．

5 章では，モーターの異常回転音やベアリングのぶつかり音などの普段発生しない音（異常音）を検知し，機器動作の状態が正常か異常かを判定することで機器の故障を検知する「異常音検知」の実現を目指す．この問題の難しさは，機器の故障頻度がきわめて低いため，機器の異常動作音（ラベルデータ）が収集できず，一般的な識別のためのニュー

ラルネットワークの目的関数である交差エントロピーが利用できない点にある．そこで5章では，正常音が従う確率分布と統計的に差異がある音を異常音と定義することで異常音検知を仮説検定とみなし，異常音検知器を最適化するための目的関数として，仮説検定の最適化基準であるネイマン・ピアソンの補題から“ネイマン・ピアソン指標”を導出した．定量評価試験では，従来法と比べ調和平均が向上したことから，提案法が従来法よりも安定して異常音検知できることを示した．また実環境実験では3Dプリンタや送風ポンプの突発的な異常音や，ベアリングの傷などに起因する持続的な異常音を検知できることを示した．この成果により，異常音データの集まらない状態識別問題を安定的に解くことが可能になり，銃声検知や未知話者検出などのセキュリティのための音源情報推定技術など，負例データの収集が困難な様々な音源情報推定へと応用ができる．

目次

第 1 章

序論

1.1 研究背景

本研究は，マイクロホンで観測した音響信号（以降，観測信号と呼ぶ）から，音に関係する情報である「音源情報」を推定する研究である．これまで，観測信号から音源の位置，方向などの物理的かつ顕在的な音源情報を推定することで，音声会議システム [1, 2, 3] や放送 [4, 5, 6]，音響監視システム [7, 8, 9, 10] など様々な音情報処理技術が研究，開発，また実用化されてきた．本研究では，より知的な音情報処理を実現するために，従来研究における音源情報の定義 [11] を拡張し，源信号に加えて音源の種類や状態などの潜在的な音源情報を推定することを目指す．潜在的な音源情報推定の題材として，源信号と雑音が重畳した観測信号から源信号を推定する音源情報推定である「音源強調」と，観測信号に含まれる環境音の種類や状態を推定して周囲の危険を予測/察知する「異常音検知」に焦点を当てる．

音源情報推定は，着目した音源、すなわち源信号が発する音から情報を抽出する処理であるため，観測信号から源信号そのものを推定する「音源強調」は，全ての音源情報推定問題において重要な役割を担う技術である．音源の種類や状態などの潜在的な音源情報を考慮しながら源信号を推定できれば，沢山の話者の中から特定の話者の声を選択的に抽出して音声認識ができる [12, 13, 14]．また，大歓声に包まれたサッカースタジアムで特定の選手の声やボールのキック音だけをクリアかつ選択的に抽出することで，まるでサッカースタジアムに潜り込んだようなコンテンツ視聴をユーザに提供できる [4, 5, 6, 11, 15, 16]．

「異常音検知」は，観測信号から周囲の危険を予測/察知する音響監視システムを実現するための必須技術として，産業界からの期待の大きい技術である．観測信号に含まれる環境音の種類や状態を推定する異常音検知が実現すれば，空港や街中で悲鳴や銃声などの危険な音を迅速な察知でき，安心安全な社会が実現できる [7, 8, 9, 10]．また産業機器の動作音からその機器の動作が正常か異常かの状態を判定できれば，製造/保守業務の効率化ができる [17, 18]．

本研究は，こういった知的な音情報処理の実現のために，音源情報推定を，マイクロホンを用いて電気信号に変換された音波の時間的な変化を，源信号や音源の種類，状態などの情報に変換する写像関数の設計問題と捉え，潜在的な音源情報を推定するための写像関数を設計する研究を行う．

1.2 従来の音源情報推定

本研究で焦点を当てる「音源強調」と「異常音検知」はそれぞれ，観測信号を源信号の波形へと変換する写像関数と，観測信号を監視対象の音源の状態（正常/異常）へと変換する写像関数の設計問題である．これまで，音源情報推定に関する多くの研究では，写像関数を物理法則に基づき記述してきた [19, 20, 21, 22]．これは，音の発生や伝搬は物理現象であるため，物理的な音源情報の入出力関係を正しく表現するためには，音の物理的な性質を表現することが不可欠なためである．

音源強調の代表的な手法であるアレイ信号処理 [11, 23] では，音の空気伝搬の物理特性を利用する．ここで音の空気伝搬の物理特性とは，複数のマイクロホンの観測信号間で生じる振幅差や位相差である．アレイ信号処理の代表的な手法であるビームフォーミングに基づく線形フィルタリングでは，源信号の到来方向に起因する振幅差や位相差を利用して源信号を強調するフィルタを設計し，観測信号と畳み込むことで源信号を抽出する [23]．また線形フィルタリングの音源強調性能の向上を目的とする非線形フィルタリング [24] では，線形フィルタの出力から源信号のパワースペクトルを推定し，Wiener filter などの時間周波数マスクを観測信号に乗ずることで源信号を抽出する．しかし，源信号と雑音が同じ方向から到来する場合，空気伝搬の物理特性から解析できる音の到来方向だけでは，源信号と雑音を区別することができない．源信号と同じ方向から到来する雑音を抑圧し，源信号を抽出するためには，潜在的な音源情報である源信号の種類や状態を見分けて音源強調する必要がある．

異常音検知においても，古典的な手法では，監視対象物の物理的な構造を利用する [25, 26]．例えば卵のひび割れの打音検査では，運動方程式に基づき楕円状の物質の周波数応答をモデル化し，実測の打音とのかい離を検知することで異常音検知している [25]．しかしこれらの手法は，監視対象物の物理的な構造が既知かつ単純でないと採用することができないし，また未知の異常音も検知できない．未知の異常音を検知し，音源の状態を推定するためには，監視対象物の物理構造に依存しない検知手法が必要である．

このように，音源の種類や状態などの潜在的な音源情報は，物理的にその性質を定義することが困難なため，物理法則に基づく写像関数では推定が難しい．そこで，より情報論的に写像関数を設計する方法として，統計的機械学習に基づく音源情報推定の研究がされてきた [27, 28, 29, 30, 31, 32, 33, 34]．統計的機械学習に基づく音源情報推定で

は，事前に収集された観測信号のデータと所望の音源情報のデータの統計的な関係性を表現するように写像関数を設計/学習するため，潜在的な音源情報も扱うことが可能である．この手法で音源情報を精緻に推定するためには，入出力の関係性を精緻に表現できる柔軟な写像関数が必要である．そのため写像関数に関する研究が広く行われており，入出力の関係性をベイズ統計理論に基づき記述する方法 [27] や，入出力の高次相関を再生核ヒルベルト空間で扱うカーネル法 [35] など，様々な手法が検討されている．次節では，高度な非線形写像関数の設計法の一つである深層学習を応用した音源情報推定について概説する．

1.3 深層学習に基づく音源情報推定

近年，入出力の関係性を精緻に表現する非線形写像関数を学習する枠組みとして，多層構造化したニューラルネットワークを非線形写像関数として用いる深層学習 [36, 37] が大きな成功を収めている．音源情報推定においても，深層学習は成功を収めており，特に観測信号から話者の発話内容を推定する音源情報推定課題である音声認識では，深層学習を適用することで従来の音声認識システムを大きく上回る推定精度を実現した [38]．

深層学習に基づく音源情報推定では，ニューラルネットワークを観測信号から所望の音源情報への非線形写像関数として用いる．そしてその写像関数を，事前に収集した観測信号と所望の音源情報のラベルデータ（例えば，源信号のスペクトルや音源の種類/状態のラベル）を用いて，二乗誤差関数などの音源情報の推定精度を評価する「目的関数」の値を最大化/最小化するように教師あり学習する．また近年では，非線形写像の柔軟性を高めるために，生物の脳の視覚野における神経細胞の受容野の局所性をヒントにした畳み込みニューラルネットワーク（CNN: convolutional neural network）[39, 40] や，時系列情報の時間方向の関係性をモデル化する再帰型ニューラルネットワーク（RNN: recurrent neural network）[41, 42] などのネットワーク構造の研究を応用した研究もおこなわれている．これらの研究では，ネットワーク構造に局所的なスペクトルの変化 [43] や時間方向の関係性の解析機構 [44]，またスペクトルの非負性などの物理的な制約 [45, 46] を加味させる改良を行うことで，音源情報推定の精度が向上することが知られている．

音源情報推定において目的関数の設計とは，所望の音源情報の性質や推定精度を定義することと等価である．多くの教師あり学習に基づく音源情報推定では，目的関数を二乗誤差関数や交差エントロピー関数などの決定論的な関数として記述し，観測信号と所望の音源情報のラベルが対になった大量の学習データを用いて，目的関数の値を最大化/最小化するように学習する [36, 37]．そのため，ラベルデータが収集困難な音源情報や，ラベルが一意に定まらない音源情報は，ニューラルネットワークの構造を工夫するだけでは推定ができないといった問題があった．この問題のために，これまで推定が困難と

されてきた音源情報の例を 3 つ挙げる．

1. 大量のラベルデータが集まらない源信号（音源強調）: スポーツフィールド上の競技音（サッカーボールのキック音や野球のバッティング音）をクリアに収音し，スポーツフィールドに潜り込んだような音響体験を実現したい．しかし競技音の源信号は，音声のように無響室など理想的な環境で学習データを収集することが困難である．ゆえに従来の音源強調 [47, 48] のように，観測信号と源信号を大量に収集して二乗誤差を最小化するようにニューラルネットワークを教師あり学習することができないため，音源情報の推定が困難である．

2. ラベルデータが一意に定まらない源信号（音源強調）: 高品質な音声通信を実現するために，主観品質を最大化するように源信号を収音したい．しかし源信号の推定結果の二乗誤差の大きさと人間が知覚する音質の劣化の大きさは必ずしも比例しないことが知られている [49]．そのため二乗誤差最小化などでニューラルネットワークを学習して源信号を推定しても，主観品質を最大化する源信号を推定することはできない．聴覚フィルタなどを利用し，主観評価値と相関の高い音質評価値 [50, 51, 52] を計算することはできるが，評価値からそれを最大化するラベル（例えば，時間周波数マスク）は一意に定めることができない．ゆえにニューラルネットワークを教師あり学習することができないため，音源情報の推定が困難である．

3. 異常音（ラベルデータ）が収集できない音源の状態（異常音検知）: 多くの機器は故障する前に，モーターの異常回転音やベアリングのぶつかり音などの普段発生しない音（異常音）が発生する．こういった音を検知することで機器動作の状態が正常か異常かを判定する技術が研究されている [17, 18]．しかし異常音は正常音と比べ発生する頻度がきわめて少なく，音データの収集が困難である．ゆえに音声認識のように，正常音と異常音を大量に収集し，それの識別率を最大化するようにニューラルネットワークを教師あり学習することができないため，音源情報の推定が困難である．

これらの音源情報推定に共通する課題は，二乗誤差関数や交差エントロピー関数などの決定論的な目的関数では，音源情報の性質や推定精度を定義できない，もしくは定義することが妥当ではない点にある．この問題を解決するためには，ネットワークの構造だけでなく，ニューラルネットワークの学習に用いる目的関数を高度化しなくてはならない．

1.4 研究の位置づけ

本研究では，決定論的な関数で目的関数を設計できない音源情報を推定するための，深層学習に基づく音源情報推定のための目的関数の研究を行う．目的関数が決定論的に定義できない音源情報においても推定したい音源情報の特性や解きたい問題に応じて，入出力値がとるべき値の確率分布や集合は定義できるはずである．本研究ではこの着眼点のもと，ニューラルネットワークの入出力が満たすべき統計的な性質を所望の音源情報の性質や推定精度として定義し，目的関数を設計するという着想からこの問題に取り組む．教師あり学習に基づく音源情報推定がラベルデータを用意して決定論的に音源情報推定結果を評価した一方で，本研究では解きたい問題の性質に合わせて推定された音源情報が持つべき統計的な性質を定義し，確率論的に音源情報推定結果を評価する目的関数を設計する．このような性質から，本研究で設計する目的関数を総称して「確率論的目的関数」と呼ぶことにする．

なお，音源情報の性質やその推定精度の評価方法は，音源情報の種類や，解きたい問題に応じて変化するはずである．種々の音源情報推定問題の中で本研究では,「音源強調」と「異常音検知」に焦点を当て，以下の問題を解決する研究を行う．

1. 音源強調（少量のラベルデータが得られる場合）: MMSE を目的関数としたニューラルネットワークの学習を行いたいが，競技音の強調のように，源信号のラベルデータが少量しか得られない状況を考える．過学習を抑えるためにニューラルネットワークによる非線形写像の自由度を抑えたい．そこで源信号をクリアに推定するための音響特徴量を事前に取捨選択するために，MMSE を最小化する音響特徴量の統計的な性質を定義し，最適な音響特徴量を選択するための目的関数の設計について提案する．

2. 音源強調（ラベルデータが一意に定まらない場合）: 主観評価値と相関の高い音質評価値 [50, 51, 52] が計算できるとき，この評価値を最大化する源信号を推定したい．ニューラルネットワークを用いて源信号を推定するために，強化学習の考え方に基づいた目的関数の設計について提案する．

3. 異常音検知（ラベルデータがない場合）: 外れ値検知 [53] の考え方を適用した異常音検知では，正常音が従う確率分布と統計的に差異がある音を異常音と仮定する．この仮定の下で，ニューラルネットワークの出力が統計的に満たすべき性質を定義し，その目的関数の設計について提案する．

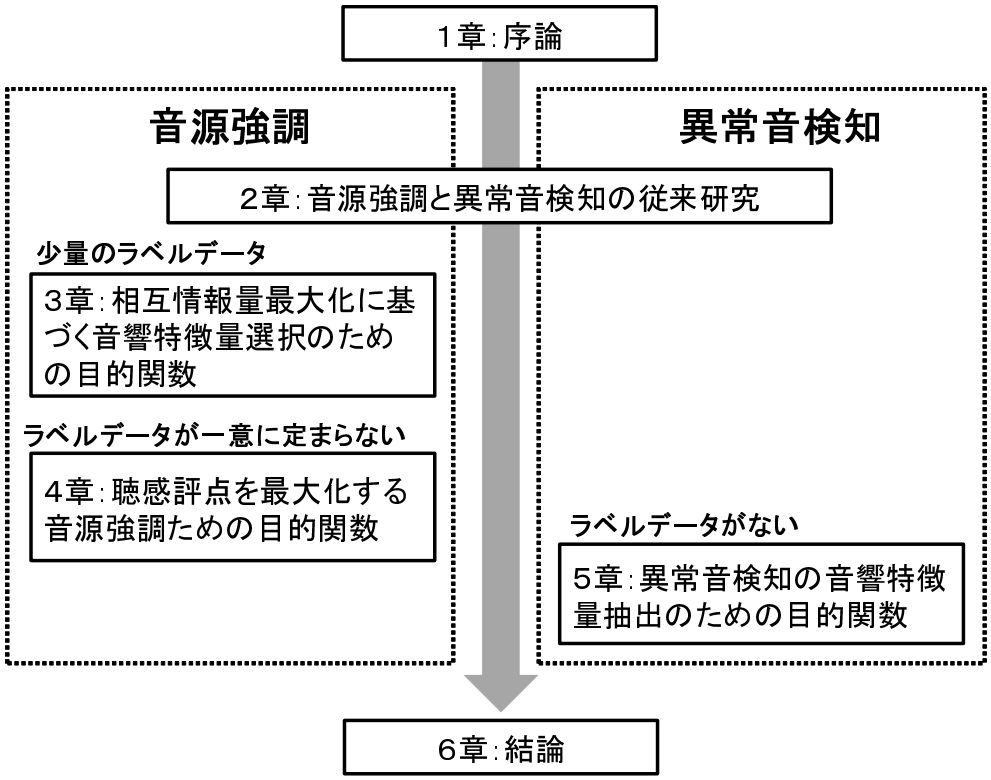


図 1.1: 本論文の構成

1.5 本論文の構成

図 1.1 に，本論文の構成を示す．2 章では，音源強調と異常音検知の従来研究を説明し，その後，深層学習，および深層学習を用いた音源強調と異常音検知の従来研究について説明する．3 章では，少量のラベルデータから，音源情報の推定二乗誤差を最小化する最適な音響特徴量を選択するための目的関数の設計について提案する．4 章では，主観評価値を最大化する音源強調を実現するための目的関数とその最適化法を提案する．5 章では，外れ値検知に基づく異常音検知の精度を最大化する，音響特徴量抽出のための目的関数の設計について提案する．6 章では，本論文の結論を述べる．

第 2 章

音源強調と異常音検知の従来研究

本章では，音源強調と異常音検知の従来研究について説明する．まず，音源強調と異常音検知についてそれぞれ 2.1 節と 2.2 節で概説および定式化したのち，2.3 節で両技術を音源情報推定として一般化する．次いで 2.4 節では深層学習について概説する．その後，2.5 節では深層学習を用いた音源強調の従来研究とその問題点，2.6 節では深層学習を用いた異常音検知の従来研究とその問題点について説明する．

2.1 音源強調

音源強調は，M 本のマイクロホンを用いて観測した信号に含まれている所望の音源の源信号を推定する音源情報推定問題である．音源強調はその応用範囲の広さから古くから取り組まれており，音声会議システム [1, 2, 3]，雑音下での音声認識 [12, 13, 14]，スポーツの生中継 [4, 5, 6] など，音を用いたアプリケーションを実環境で頑健に動作させるためのフロントエンド処理として高度化が期待されている．

音源強調の定式化のために，まず観測信号をモデル化する．m 番目のマイクロホンの観測信号を数十 ms 分の長さで切り出し，時間フレームごとに短時間フーリエ変換（STFT: short-time Fourier transform）した $X_{m,\omega,\tau} \in \mathbb{C}^{M\times\Omega\times T}$ を，$S_{\omega,\tau} \in \mathbb{C}^{\Omega\times T}$ と K 個の雑音 $N_{k,\omega,\tau} \in \mathbb{C}^{K\times\Omega\times T}$ が重畳されたものとして以下のように記述する．

$$X_{m,\omega,\tau} = V_{m,0,\omega}S_{\omega,\tau} + \sum_{k=1}^{K} V_{m,k,\omega}N_{k,\omega,\tau} \tag{2.1}$$

ここで $\omega = \{1, 2, ..., \Omega\}$ と $\tau = \{1, 2, ..., T\}$ は，周波数と時間のインデックスを表す変数である．また，$V_{m,0,\omega}$ は所望の源音源から m 番目のマイクロホンまでの，また $V_{m,k,\omega}$ は k 番目の雑音から m 番目のマイクロホンまでの伝達特性を表す．つまり音源強調は，観測信号 $X_{m,\omega,\tau}$ を入力とし，源信号 $S_{\omega,\tau}$ や源信号を推定するためのパラメータを推定する問題である．

音源強調の代表的な手法として，複数のマイクロホンを用いて観測信号を解析するアレイ信号処理がある [11, 54, 55, 56, 23, 57]．多くのアレイ信号処理では，複数の観測信号間で生じる振幅差や位相差を利用し音源強調や残響除去を実現する．ビームフォーミングに基づく線形フィルタリングでは，源信号の方向情報をもとに源信号を強調するフィルタを設計し，観測信号と畳み込むことで源信号を抽出する [23]．時間領域の畳み込みは，周波数領域では乗算となるため，線形フィルタリングは以下のように記述できる．

$$Y_{\omega,\tau} = \sum_{m=1}^{M} F_{m,\omega} X_{m,\omega,\tau} \tag{2.2}$$

ここで $F_{m,\omega}$ は線形フィルタ，$Y_{\omega,\tau}$ は線形フィルタリングの出力音である．線形フィルタの設計法には，焦点形成法，死角形成法，MINT (Multiple-input INverse Theorem) 法などがある [11]．

線形フィルタリングの音源強調性能を高めるために，数十チャネル以上のマイクロホンを用いる大規模なマイクロホンアレイの研究 [58, 59] やマイクロホンアレイの構造に関する研究 [60, 61] も盛んに行われている．また源信号に関するあらゆる情報が未知の場合に源信号を抽出する方法として，源信号と雑音の独立性 [62] や時間周波数方向の低ランク性 [63]，またその両方 [64] の仮定の下で観測信号をモデル化し線形フィルタを設計する，ブラインド音源分離の研究も行われている [65]．

2.1.1 時間周波数マスクを用いた音源強調

線形フィルタリングの音源強調性能の向上や，単一のマイクロホンの観測信号からの音源強調を目的として，非線形フィルタリングの研究も盛んに行われている [66, 67, 68, 24]．2.1.1 節では表記の簡単化のために，$M = 1$ 本のマイクロホンで観測した信号からの音源強調を仮定し，また伝達特性 $V_{m,k,\omega}$ の記述とマイクロホンのインデックス m を省略して，観測信号を以下のように記述する．

$$X_{\omega,\tau} = S_{\omega,\tau} + \sum_{k=1}^{K} N_{k,\omega,\tau} \tag{2.3}$$

本論文で用いる非線形フィルタリングは，各時間周波数成分ごとにゲインを調整する時間周波数マスクに基づく処理である．時間周波数マスクに基づく音源強調では，0 から 1 の値を持つ時間周波数マスク $G_{\omega,\tau} \in [0, 1]$ を観測信号 $X_{\omega,\tau}$ に掛け合わせることで，源信号が強調された信号 $\hat{S}_{\omega,\tau} \in \mathbb{C}^{\Omega \times T}$ を得る（図 2.1）．

$$\hat{S}_{\omega,\tau} = G_{\omega,\tau} X_{\omega,\tau} \tag{2.4}$$

つまり，時間周波数マスクを用いた音源強調は，観測信号を時間周波数マスク $G_{\omega,\tau}$ やその設計のためのパラメータへの写像関数を求める問題に帰着する．

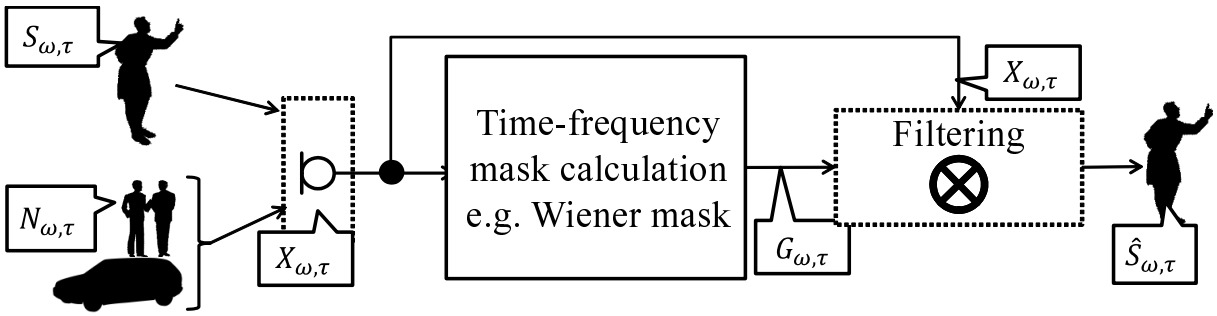


図 2.1: 時間周波数マスクに基づく音源強調.

$G_{\omega,\tau}$ の代表的な計算法には，ウィナーマスク [69]，理想比率マスク (IRM: Ideal ratio mask)[70]，バイナリマスク [67, 71] がある．以下では，各時間周波数マスクを概説する．

ウィナーマスク [69] は，源信号と全ての雑音が互いに無相関かつ定常である時に $S_{\omega,\tau}$ と $\hat{S}_{\omega,\tau}$ の MMSE を最小化するマスクである．しかし，源信号や雑音は非定常で有ることが多いため，実用上は以下の時変ウィナーマスクを用いることが多い．

$$G^{\mathrm{WF}}_{\omega,\tau} = \frac{|S_{\omega,\tau}|^2}{|S_{\omega,\tau}|^2 + |\sum_{k=1}^{K} N_{k,\omega,\tau}|^2} \tag{2.5}$$

ウィナーマスクを計算するためには，源信号の振幅スペクトル $|S_{\omega,\tau}|$ と雑音の振幅スペクトル $|\sum_{k=1}^{K} N_{k,\omega,\tau}|$ の両方を推定しなくてはならない．実用上は計算量や推定する値の数を少なくするために，以下のように源信号と雑音の加法性がパワースペクトル領域でも成り立つと仮定し，

$$|X_{\omega,\tau}|^2 = |S_{\omega,\tau}|^2 + \left|\sum_{k=1}^{K} N_{k,\omega,\tau}\right|^2 \tag{2.6}$$

源信号と雑音のどちらか片方を推定し，以下の形で近似的にウィナーマスクを計算することが多い．

$$G^{\mathrm{WF}}_{\omega,\tau} \approx \frac{|S_{\omega,\tau}|^2}{|X_{\omega,\tau}|^2} \tag{2.7}$$

$$\approx \frac{|X_{\omega,\tau}|^2 - |\sum_{k=1}^{K} N_{k,\omega,\tau}|^2}{|X_{\omega,\tau}|^2} \tag{2.8}$$

式 (2.7) が源信号の振幅スペクトルを推定した場合，式 (2.8) が雑音の振幅スペクトルを推定した場合のウィナーマスクの近似計算式である．また以下の式変形を行うことで，振幅スペクトルの代わりに，源信号と雑音のパワースペクトルの比率である事前信号雑音比（SNR: signal-to-noise ratio）$\xi_{\omega,\tau}$ を推定することで，ウィナーマスクの近似計算を

避ける方法もある．

$$G^{\mathrm{WF}}_{\omega,\tau} = \frac{\xi_{\omega,\tau}}{1+\xi_{\omega,\tau}} \tag{2.9}$$

$$\xi_{\omega,\tau} = \frac{|S_{\omega,\tau}|^2}{|\sum_{k=1}^{K} N_{k,\omega,\tau}|^2} \tag{2.10}$$

ウィナーマスクが源信号と観測信号のパワースペクトルの比率のマスクである一方で，IRM は源信号と観測信号の振幅スペクトルの比率のマスクである．

$$G^{\mathrm{IRM}}_{\omega,\tau} = \frac{|S_{\omega,\tau}|}{|S_{\omega,\tau}| + |\sum_{k=1}^{K} N_{k,\omega,\tau}|} \tag{2.11}$$

振幅スペクトル上で，源信号と雑音の加法性が成り立つ場合，ウィナーマスクよりも IRM の方が，源信号の推定精度が向上することも知られている [73]．IRM もウィナーマスクの場合と同様に，以下のように源信号と雑音の加法性が振幅スペクトル領域でも成り立つと仮定すれば，

$$|X_{\omega,\tau}| = |S_{\omega,\tau}| + \left|\sum_{k=1}^{K} N_{k,\omega,\tau}\right| \tag{2.12}$$

源信号と雑音のパワースペクトルどちらか一方から近似計算できる．

$$G^{\mathrm{IRM}}_{\omega,\tau} \approx \frac{|S_{\omega,\tau}|}{|X_{\omega,\tau}|} \tag{2.13}$$

$$\approx \frac{|X_{\omega,\tau}| - |\sum_{k=1}^{K} N_{k,\omega,\tau}|}{|X_{\omega,\tau}|} \tag{2.14}$$

バイナリマスク [67, 71] は，$G^{\mathrm{BM}}_{\omega,\tau}$ を $\{0,1\}$ の二値に制限することで，時間周波数マスクの設計やマスキングを高速に動作させるマスクである．バイナリマスクでは，源信号と雑音は時間周波数上で重なりなく，スパースに存在していることを仮定している．バイナリマスクは例えば以下のように設計できる [57]．

$$G^{\mathrm{BM}}_{\omega,\tau} = u\left(|S_{\omega,\tau}| - \left|\sum_{k=1}^{K} N_{k,\omega,\tau}\right|\right) \tag{2.15}$$

$$u(x) = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases} \tag{2.16}$$

バイナリマスクも，ウィナーマスクや IRM と同様に，IRM も源信号と雑音のパワースペクトルどちらか一方から近似計算できる．

$$G^{\mathrm{BM}}_{\omega,\tau} \approx u\left(|S_{\omega,\tau}| - (|X_{\omega,\tau}| - |S_{\omega,\tau}|)\right) = u\left(2|S_{\omega,\tau}| - |X_{\omega,\tau}|\right) \tag{2.17}$$

$$\approx u\left(\left(|X_{\omega,\tau}| - \left|\sum_{k=1}^{K} N_{k,\omega,\tau}\right|\right) - \left|\sum_{k=1}^{K} N_{k,\omega,\tau}\right|\right) = u\left(|X_{\omega,\tau}| - 2\left|\sum_{k=1}^{K} N_{k,\omega,\tau}\right|\right) \tag{2.18}$$

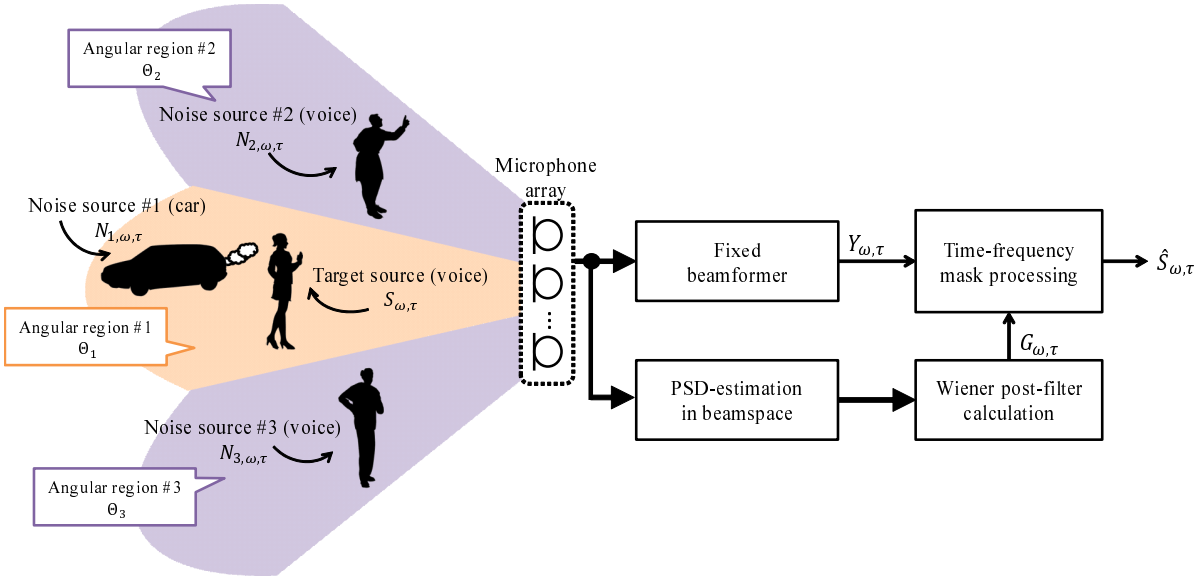


図 2.2: 局所 PSD 推定法による時間周波数マスク設計.

なおバイナリマスクの設計法には式 (2.15) の他にもさまざまな方法が提案されており，代表的な方法として，信号の 2 チャネルへの到達レベル差を利用してバイナリマスクを設計する SAFIA (sound source Segregation based on estimating incident Angle of each Frequency component of Input signals Acquired by mlultiple microphones)[67] などがある．

2.1.2 時間周波数マスクの推定

ここまで代表的な時間周波数マスクについて概説してきた．各時間周波数マスクの計算式から明らかなように，時間周波数マスクの計算のためには源信号や雑音の振幅スペクトルやパワースペクトル，事前 SNR を推定する必要がある．これらの代表的な推定法には，雑音の定常性を仮定して源信号の振幅スペクトルを求めるスペクトル減算法 [74]，複数のマイクロホンを用いて拡散性雑音を求める方法 [66, 68]，また線形フィルタリングにおけるビームフォーマの感度行列を利用して干渉性雑音を求める局所 PSD（power spectral density）推定法 [24] などがある（図 2.2）．ここでは時間周波数マスクの計算のためのパラメータ推定法の従来法の一例として，源信号や雑音のパワースペクトル密度（PSD）を求める手法である，局所 PSD 推定法 [24] を説明する．

局所 PSD 推定法は，図 2.2 に示すように，複数のビームフォーミングの出力を利用して非線形マスクを設計する方法である．今，音源が存在する領域を重なりのない L 個の領域 $\Theta_{1,\ldots,L}$ に分割する．そして，l 番目の領域に存在する音源群を強調する線形フィル

タを $F_{l,m,\omega}$ とし，その出力を以下のように記述する．

$$Y_{l,\omega,\tau} = \sum_{m=1}^{M} F_{l,m,\omega} X_{m,\omega,\tau} \tag{2.19}$$

$$= \sum_{m=1}^{M} F_{l,m,\omega} \left(V_{m,0,\omega} S_{\omega,\tau} + \sum_{k=1}^{K} V_{m,k,\omega} N_{k,\omega,\tau} \right) \tag{2.20}$$

すると，l 番目のビームフォーマにおける k 番目の音源への感度は以下のように記述できる．

$$D_{\omega}^{(l,k)} = \left| \sum_{m=1}^{M} F_{l,m,\omega} V_{m,k,\omega} \right|^2 \tag{2.21}$$

ここで，源信号と各雑音は無相関であると仮定すると，各ビームフォーマの出力のパワースペクトル密度は，各音源のパワースペクトル密度に感度 $D_{\omega}^{(l,k)}$ を乗じた値の和で表現できる．ここで，全ての時間フレームでこの関係性が成り立つと仮定することで，各ビームフォーマの出力のパワースペクトルは以下のように記述できる．

$$|Y_{l,\omega,\tau}|^2 = D_{\omega}^{(l,0)} |S_{\omega,\tau}|^2 + \sum_{k=1}^{K} D_{\omega}^{(l,k)} |N_{k,\omega,\tau}|^2 \tag{2.22}$$

この関係性を行列形式で記述すると以下のように記述できる．

$$\underbrace{\begin{bmatrix} |Y_{1,\omega,\tau}|^2 \\ |Y_{2,\omega,\tau}|^2 \\ \vdots \\ |Y_{L,\omega,\tau}|^2 \end{bmatrix}}_{\boldsymbol{\Phi}_{Y,\tau}} = \underbrace{\begin{bmatrix} D_{\omega}^{(1,0)} & D_{\omega}^{(1,1)} & \cdots & D_{\omega}^{(1,K)} \\ D_{\omega}^{(2,0)} & D_{\omega}^{(2,1)} & \cdots & D_{\omega}^{(2,K)} \\ \vdots & \vdots & \ddots & \vdots \\ D_{\omega}^{(L,0)} & D_{\omega}^{(L,1)} & \cdots & D_{\omega}^{(L,K)} \end{bmatrix}}_{\boldsymbol{D}_{\omega}} \underbrace{\begin{bmatrix} |S_{\omega,\tau}|^2 \\ |N_{1,\omega,\tau}|^2 \\ \vdots \\ |N_{K,\omega,\tau}|^2 \end{bmatrix}}_{\boldsymbol{\Phi}_{S,\tau}} \tag{2.23}$$

ここで $\boldsymbol{D}_{\omega}$ はビームフォーマの感度行列である．したがって，源信号および雑音のパワースペクトルは，以下の逆問題を解くことで推定することができる．

$$\boldsymbol{\Phi}_{S,\tau} = \boldsymbol{D}_{\omega}^{+} \boldsymbol{\Phi}_{Y,\tau} \tag{2.24}$$

ただし，上付き文字の $+$ は一般化逆行列を表す．源信号および雑音のパワースペクトルが求まれば，式 (2.5) でウィナーマスクを設計することができるため，源信号を強調することができる．

この手法は，ビームフォーマの出力と感度行列 $\boldsymbol{D}_{\omega}$ から源信号と雑音のパワースペクトルを推定している．すなわち，音源の位置や方向などの物理的な性質を利用した音源強調である．ゆえに源信号と雑音が空間的に離れて位置している場合は精度よく源信号を推定できるが，図 2.2 のように，所望の源信号と雑音 (Noise source # 1) が空間的に近接した位置に存在する場合，感度行列がランク落ちを起こすため一般化逆行列の計算

が不安定となり，源信号を推定することができない．これを解決するためには，音源の物理的な性質だけでなく，音源の種類などの潜在的な音源情報も加味しながら時間周波数マスクのパラメータを推定する必要がある．

2.2 異常音検知

異常音検知は，観測信号 $X_{\omega,\tau} \in \mathbb{C}^{\Omega\times T}$ が，その音を発した物体/環境の正常な状態に起因して発せられたもの（正常音）か，異常な状態に起因して発せられたもの（異常音）かを推定する，音源の状態推定問題である．つまり異常音検知は，観測信号 $X_{\omega,\tau}$ を入力とし，監視対象の状態を出力するような写像関数を設計する問題である．応用例には，空港や街中の普段の音を正常音として，悲鳴や銃声などの危険な音を検知する「音響監視システム」[7, 8, 9, 10] や，産業機器の普段の動作音を正常音として，故障に起因するめったに発生しない機械音を検知することで機器の状態を判定する「機器検査/保守」[17, 18] などがある．

異常音検知の古典的な手法では，監視対象物の物理的な構造を利用して異常音を検知する [25, 26]．例えば卵のひび割れの打音検査では，運動方程式に基づき楕円状の物質の周波数応答をモデル化し，実測の打音とのかい離を検知することで異常音検知している [25]．しかしこれらの手法は，監視対象物の物理的な構造が既知かつ単純でないと採用することができないし，また未知の異常音も検知できない．複雑な物理構造の音源の異常音やを検知するために，ヒューリスティックに判別ルールを記述することもできるが，多くの場合，どのような異常音が発生するかは未知のため現実的ではない．そこで，監視対象物の物理構造に依存せずに音源の状態を推定するために，統計的アプローチに基づく異常音検知が研究されている．

2.2.1 統計的アプローチに基づく異常音検知

異常音検知の難しさは，異常音がめったに発生しない音であるため，異常音データが集まらない点にある [53]．そのため音声認識や音響イベント検出で採用されるような，教師あり学習に基づく識別アプローチを採用することができない．ゆえに異常検知の分野では，正常音のデータだけから検知器の学習が可能な外れ値検知を応用してきた [75, 76, 77, 78, 79]．音を用いた異常音検知においても，外れ値検知に基づく手法が研究されている [80]．外れ値検知では，異常音の定義を「普段は発生しないような音」としている．より詳しくは，正常音と同じ分布から生成されたと考えらない，統計的に有意に差のある音と定義している．

外れ値検知に基づく異常音検知は，正常音の学習データから抽出された音響特徴量が

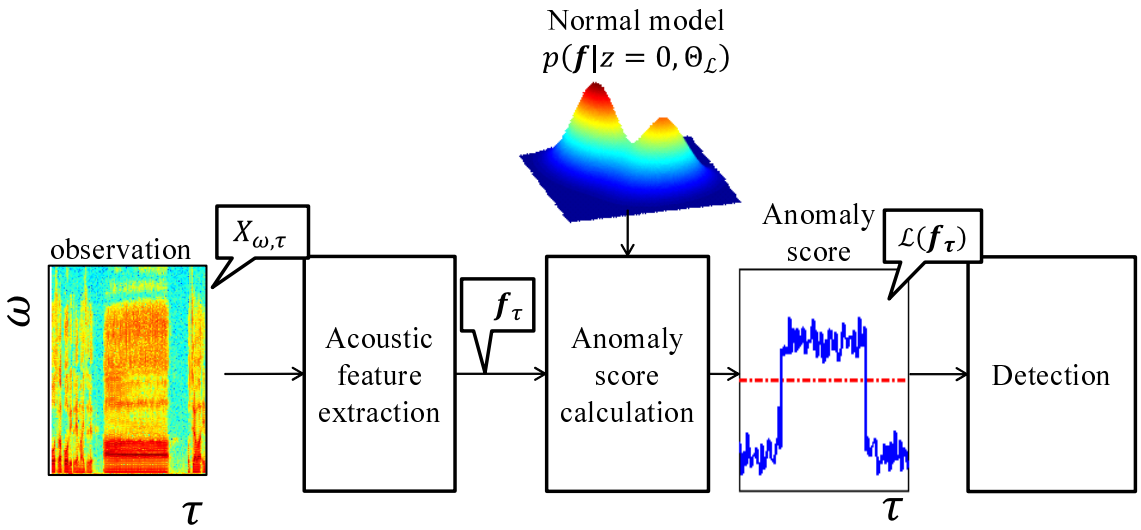


図 2.3: 外れ値検知に基づく異常音検知.

従う確率密度関数（正常モデル）を利用し，新たな観測信号から抽出された音響特徴量の負の対数尤度（異常度）を計算する．そしてその異常度が事前に設定した閾値以上であれば，観測信号は正常音の学習データとは同じ分布から生成されたとはいえないと考えられるため観測信号を異常音と判定する．図 2.3 は外れ値検知に基づく異常音検知の処理フローである．以下では処理フロー全体を概説する．

まず，観測信号から音響特徴量 $\boldsymbol{f}_\tau \in \mathbb{R}^{\mathsf{D}}$ を抽出する．

$$\boldsymbol{f}_\tau = \mathcal{M}(\boldsymbol{x}_\tau|\Theta_{\mathcal{M}}) \tag{2.25}$$

ここで $\mathcal{M}$ は音響特徴量を抽出する関数（音響特徴量抽出関数）であり，$\Theta_{\mathcal{M}}$ はそのパラメータである．音響特徴量 $\boldsymbol{f}_\tau$ には例えば，メル周波数ケプストラム係数（MFCC: mel-frequency cepstrum coefficient）や線スペクトル対（LSP: line spectral pairs）などを利用することが一般的である．次に正常モデル $p(\boldsymbol{f}|z=0,\Theta_{\mathcal{L}})$ を用いて，異常度 $\mathcal{L}(\boldsymbol{f}_\tau)$ を正常モデルの負の対数尤度として計算する．

$$\mathcal{L}(\boldsymbol{f}_\tau) = -\ln p(\boldsymbol{f}_\tau|z=0,\Theta_{\mathcal{L}}), \tag{2.26}$$

ここで離散変数 $z \in \{0,1,...,\infty\}$ は観測信号が正常音である時に $z=0$ となる確率変数，$p(\boldsymbol{f}_\tau|z=0,\Theta_{\mathcal{L}})$ は正常モデルであり $\Theta_{\mathcal{L}}$ は分布のパラメータである．代表的な正常モデルの実装例は混合ガウス分布（GMM: Gaussian mixture model）であり以下のようになる．

$$p(\boldsymbol{f}|z=0,\Theta_{\mathcal{L}}) = \sum_{c=1}^{\mathsf{C}} w_c \mathcal{N}\left(\boldsymbol{f}|\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c\right), \tag{2.27}$$

ここで C は混合数，w_c，$\boldsymbol{\mu}_c$，$\boldsymbol{\Sigma}_c$ はそれぞれ混合比と c 番目のガウス分布の平均ベクトルと共分散行列を表す．つまり $\Theta_{\mathcal{L}} = \{w_c,\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c|c=1,...,\mathsf{C}\}$ である．最後に異常度 $\mathcal{L}(\boldsymbol{f}_\tau)$

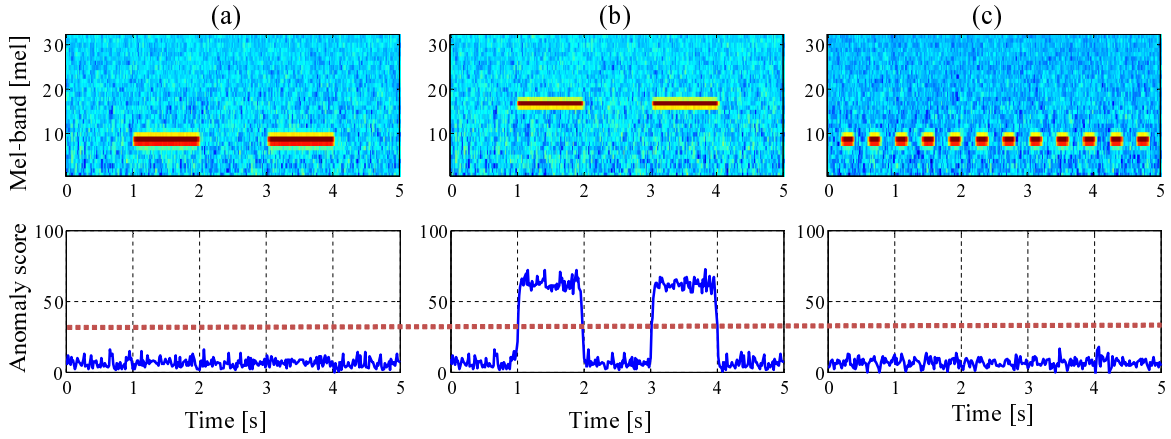


図 2.4: MFCC を用いた異常音検知の例．(a) 正常音，(b) 異常音 1，(c) 異常音 2，赤線が閾値を表す．上図は観測信号の対数メルフィルタバンク出力を表し，下図の青線は異常度，赤破線は異常判定の閾値を表す．外れ値検出に基づく異常音検知では，不適切な音響特徴量を用いると異常音を検知することができない．

が事前に設定した閾値 ϕ を越えたら，観測信号を異常音と判定する．

$$\mathcal{H}(\mathcal{L}(\boldsymbol{f}_\tau), \phi) = \begin{cases} 0 \text{ (正常音)} & \mathcal{L}(\boldsymbol{f}_\tau) \leq \phi \\ 1 \text{ (異常音)} & \mathcal{L}(\boldsymbol{f}_\tau) > \phi \end{cases} \tag{2.28}$$

2.2.2 異常音検知と音響特徴量設計

外れ値検知に基づく異常音検知において具現化すべき関数は，音響特徴量抽出関数 $\mathcal{M}$ と正常モデル $p(\boldsymbol{f}_\tau | z = 0, \Theta_\mathcal{L})$ である．後者の正常モデルは確率分布であるため，GMM などの音源情報推定に限らない汎用的な確率分布を用いることができる．ゆえに，異常音検知において設計が重要となるのは音響特徴量抽出関数 $\mathcal{M}$ である．例えば，正常音と異常音で大きく異なる音の特徴が音色であるならば，MFCC などの音響特徴量を抽出すればよい．また正常音と異常音で音の時間構造が大きく異なるなら，振幅スペクトルの差分（Δ 特徴量）などを抽出すればよい．しかし異常音検知では，異常音データが事前に収集できないため，正常音と異常音の間でどのような音響的な違いが発生するかが未知であり，異常音の特性に合わせて音響特徴量をヒューリスティックに設計することは困難である．また，不適切な音響特徴量を用いて異常音検知を行うと，その検知精度が著しく低下する．

図 2.4 に，外れ値検出に基づく異常音検知の実行例を示す．正常音は周波数が 500Hz かつ持続時間が 1.0 秒の単一正弦波とした（図 2.4, (a)）．異常音は周波数が 1200Hz かつ持続時間が 1.0 秒の単一正弦波（異常音 1，図 2.4, (b)）と高さが 500Hz かつ持続時間

が 0.2 秒の単一正弦波（異常音 2，図 2.4, (c)）した．式 (2.25) で抽出する音響特徴量は各フレームの観測音の MFCC とし，正常モデルは混合数が 2 の GMM でモデル化した．MFCC はスペクトル包絡（スペクトルの概形）に関する音響特徴量のため，正常音と周波数の異なる単一正弦波である異常音 1 は検知できていることがわかる．一方，単体フレームの MFCC では音の時間構造を表現することができないため，正常音と長さの異なる異常音 2 は検知できていない．このことからわかるように，外れ値検出に基づく異常音検知では，従来のように不適切な音響特徴量を用いると異常音を検知することができない．ゆえに異常音検知において中心となる問題は，観測信号から音響特徴量を抽出する関数 $\mathcal{M}$ を設計することであり，異常音検知の精度を向上させるためには，観測信号を音響特徴量へ写像する関数を適切に設計しなくてはならない．

2.3 音源情報推定の定式化

2.1 節と 2.2 節では，音源強調と異常音検知の従来技術を解説してきた．そして両技術ともその推定精度を高めるには，観測信号を時間周波数マスクのパラメータや音響特徴量へ写像する関数を適切に設計することが重要であることを述べた．本節ではこれらを踏まえ，音源情報推定を，観測信号 $\boldsymbol{x}$ を入力とし，音源情報 $\boldsymbol{y}$ を出力とするような，写像関数 $\mathcal{M}$ を設計する問題として一般化する．音源強調であれば $\mathcal{M}$ は観測信号を時間周波数マスクのパラメータへ写像する関数であるし，異常音検知であれば $\mathcal{M}$ は観測信号を音響特徴量へ写像する関数である．本研究では，写像関数 $\mathcal{M}$ を設計する手順は

(i) 観測信号 $\boldsymbol{x}$ から音源情報 $\boldsymbol{y}$ への変換を写像関数で記述し

(ii) 音源情報の推定精度の評価値を最大化するように，写像関数のパラメータを学習する

という二つの手順で構成されるとする．それぞれの手続きは，以下のように定式化できる（図 2.5）．

$$\hat{\boldsymbol{y}} = \mathcal{M}\left(\boldsymbol{x}|\Theta_{\mathcal{M}}\right) \tag{2.29}$$

$$\Theta_{\mathcal{M}} \leftarrow \arg\max_{\Theta_{\mathcal{M}}} \mathcal{J}\left(\Theta_{\mathcal{M}}\right) \tag{2.30}$$

ここで $\Theta_{\mathcal{M}}$ は写像関数のパラメータであり，$\mathcal{J}$ は音源情報の推定結果を評価する目的関数である．式 (2.29)(2.30) を実行するために具現化すべき関数は，写像関数 $\mathcal{M}$ と目的関数 $\mathcal{J}$ であり，それぞれの関数は以下の要件を満たすように設計すべきである．

$\mathcal{M}$ の条件: 観測信号 $\boldsymbol{x}$ と音源情報 $\boldsymbol{y}$ の関係性を正しく表現できる写像であること

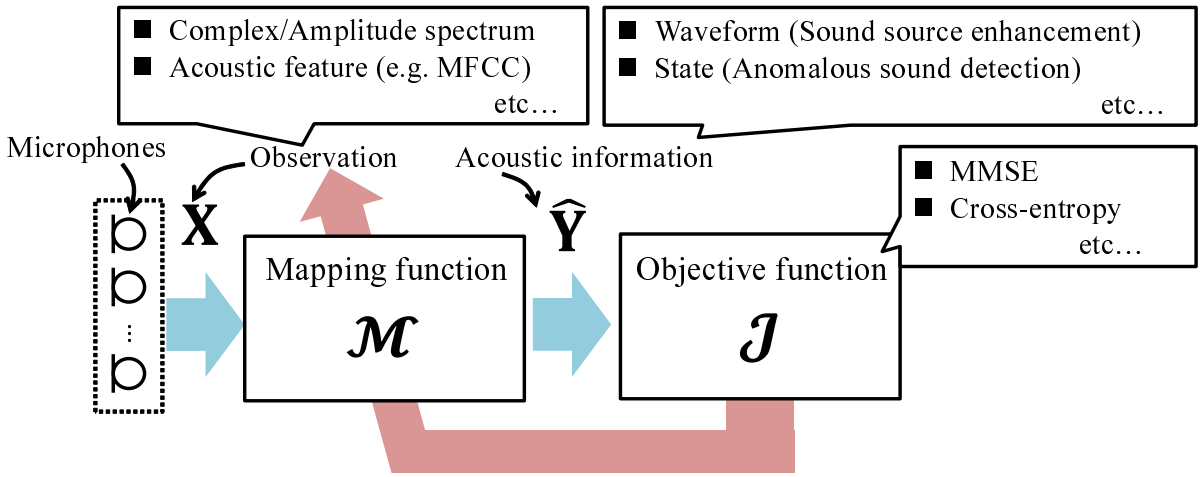


図 2.5: 音源情報推定.

$\mathcal{J}$ **の条件:** 所望の音源情報の性質や推定精度を定義する尺度であること

物理法則に則り記述するだけでは推定が難しい音源情報を推定するために，情報論的に写像関数を設計する方法として，統計的機械学習に基づく音源情報推定の研究がされてきた [29, 30, 31, 32, 33, 34]．統計的機械学習に基づく音源情報推定では，$\mathcal{M}$ を物理的な制約を緩和した非線形な写像関数として記述する．特に教師あり学習に基づく音源情報推定では，事前に収集された観測信号のデータ $\mathcal{X} = \{\boldsymbol{x}_\tau | \tau = 1, ..., T\}$ と所望の音源情報のデータ $\mathcal{Y} = \{\boldsymbol{y}_\tau | \tau = 1, ..., T\}$ の関係性を表現するように $\Theta_\mathcal{M}$ を学習することで複雑な写像を表現する．また目的関数は，音源情報の推定精度が高いほど大きな値を返す評価関数 $\mathcal{R}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ の平均値として表現することが一般的である．

$$\Theta_\mathcal{M} \leftarrow \arg\max_{\Theta_\mathcal{M}} \mathcal{J}(\Theta_\mathcal{M}) \tag{2.31}$$

$$\mathcal{J}(\Theta_\mathcal{M}) = \frac{1}{T} \sum_{\tau=1}^{T} \mathcal{R}(\boldsymbol{y}_\tau, \mathcal{M}(\boldsymbol{x}_\tau | \Theta_\mathcal{M})) \tag{2.32}$$

より複雑な写像を必要とする音源情報の推定のために，統計的機械学習に基づく音源情報推定では，入出力の関係性を精緻に表現する $\mathcal{M}$ に関する研究が広く行われてきた．これまで，入出力の関係性をベイズ統計理論に基づき記述する方法 [27] や，$\boldsymbol{x}$ と $\boldsymbol{y}$ の高次相関を再生核ヒルベルト空間で扱うカーネル法 [35] など，様々な手法が検討されている．深層学習に基づく音源情報推定では，$\mathcal{M}$ を多層構造化したニューラルネットワークで記述することで高度な非線形写像を実現している．次節以降では，まず音源強調と異常音検知について定式化したのちに深層学習について概説し，その後深層学習を用いた音源強調と異常音検知を説明する．

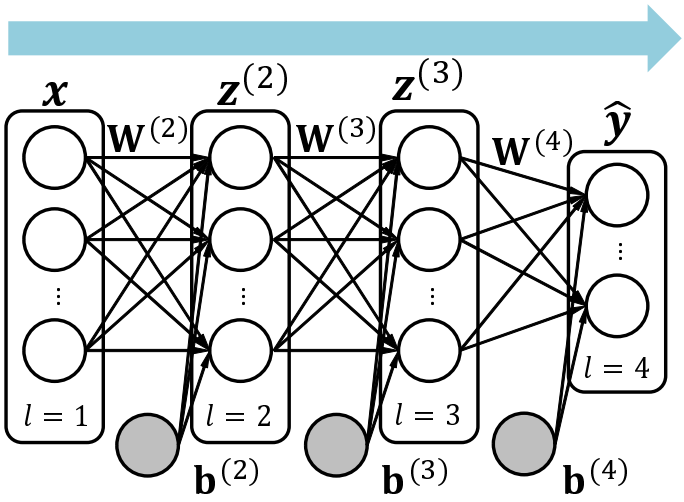


図 2.6: $L = 4$ の全結合多層ニューラルネットワークの構造.

2.4 深層学習

ここまで，音源強調と異常音検知について定式化し，どちらの手法においても，観測信号を別の情報へと写像する関数が必要があることを述べた．そして統計的機械学習に基づく音源情報推定では，入出力の関係性を精緻に表現する非線形写像関数を学習であることが重要であることを述べた．本節では，入出力の関係性を精緻に表現する非線形写像関数を学習するである深層学習について概説する．

2.4.1 多層ニューラルネットワークとその学習

本節では，最も基本的な多層ニューラルネットワークである全結合多層ニューラルネットワークとその学習法について説明する．以降，CNN や RNN との区別および表記の簡単のために，全結合多層ニューラルネットワークのことを DNN（deep neural network）と呼ぶ．図 2.6 に DNN の構造の例を示す．DNN は，各層の全てのユニットが重み行列 $\mathbf{W}$ で結合された，入力層から出力層までが一方向に伝搬するネットワークである．DNN による非線形写像（順伝搬）は，以下のように表現できる．

$$\mathcal{M}\left(\boldsymbol{x}_\tau|\Theta_\mathcal{M}\right) \leftarrow \sigma_\mathcal{M}\left\{\boldsymbol{u}_\tau^{(L)}\right\} \tag{2.33}$$

$$\boldsymbol{z}_\tau^{(l)} = \sigma_\theta\left\{\boldsymbol{u}_\tau^{(l)}\right\} \tag{2.34}$$

$$\boldsymbol{u}_\tau^{(l)} = \mathbf{W}^{(l)}\boldsymbol{z}_\tau^{(l-1)} + \mathbf{b}^{(l)} \tag{2.35}$$

ここで L はニューラルネットワークの層数であり，$\mathbf{W}^{(l)}$，$\mathbf{b}^{(l)}$ はそれぞれ l 層目の重み行列とバイアスベクトルである．つまり DNN のパラメータは $\Theta_\mathcal{M} = \{\mathbf{W}^{(l)}, \mathbf{b}^{(l)}|l = 2, ..., L\}$ である．また σ_θ と $\sigma_\mathcal{M}$ は活性化関数と呼ばれる非線形関数であり，シグモイド関数やラ

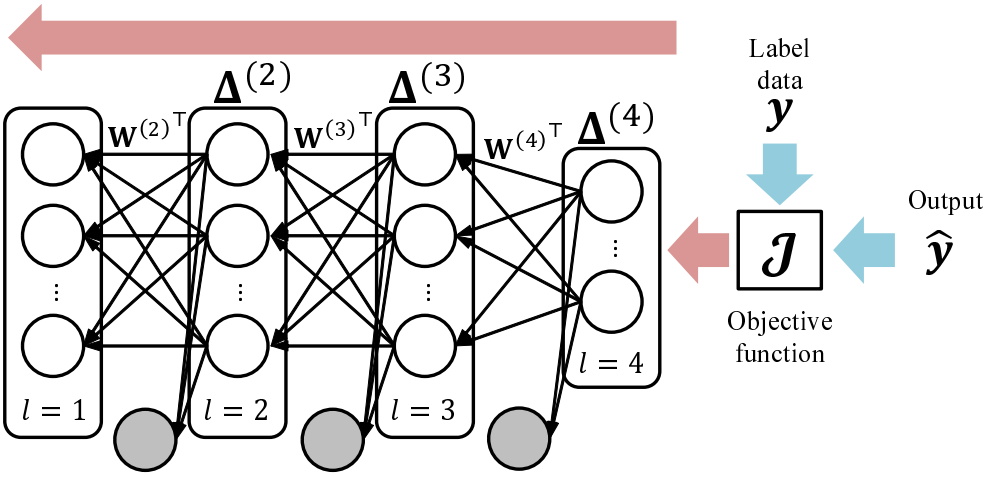


図 2.7: $L = 4$ の全結合多層ニューラルネットワークにおける誤差逆伝搬.

ンプ関数が用いられる．なお $\boldsymbol{z}_\tau^{(1)} = \boldsymbol{x}_\tau$ であり，音源情報推定において入力 $\boldsymbol{x}_\tau$ は，観測信号の周波数情報と時間情報の両方を考慮するために，観測信号 $X_{\omega,\tau} \in \mathbb{C}^{\Omega \times T}$ の時間周波数要素を並べたベクトルとすることが多い [38, 81].

$$\boldsymbol{x}_\tau = (\boldsymbol{X}_{\tau-P_b}, ..., \boldsymbol{X}_\tau, ..., \boldsymbol{X}_{\tau+P_f})^\top \tag{2.36}$$

$$\boldsymbol{X}_\tau = (\ln|X_{1,\tau}|, ..., \ln|X_{\Omega,\tau}|) \tag{2.37}$$

ここで $\omega = \{1, 2, ..., \Omega\}$ と $\tau = \{1, 2, ..., T\}$ は，周波数と時間のインデックスを表す変数であり，$\top$ は転置を表す．また P_b, P_f は考慮する前後の時間フレーム数であり，コンテキスト窓と呼ばれる [81].

教師あり学習に基づく $\Theta_\mathcal{M}$ の学習は，事前に用意された観測信号と所望の音源情報の学習データを用いて，式 (2.32) で行われる．しかし DNN は複雑な関数であり，目的関数 $\mathcal{J}(\Theta_\mathcal{M})$ を最大化する $\Theta_\mathcal{M}$ を解析的に求めるのは困難である．そこで学習には勾配法が用いられる．

$$\Theta_\mathcal{M} \leftarrow \Theta_\mathcal{M} + \lambda \frac{\partial \mathcal{J}(\Theta_\mathcal{M})}{\partial \Theta_\mathcal{M}} \tag{2.38}$$

ここで λ はステップサイズである．勾配法で複雑な非線形写像を学習する際に問題となるのは第二項の勾配計算である．DNN のように非線形写像を行列演算や非線形演算の合成関数で表現する場合，勾配の計算には微分の連鎖則を繰り返す必要がある．この問題を解決するために，ニューラルネットワークの勾配計算を効率的に行う “誤差逆転伝播法 [82]” が提案された．誤差逆伝搬では，図 2.7 のように，$l+1$ 層の誤差を l 層へ逆方向に伝搬させ勾配を計算する．詳細な導出については [37] などの専門書に譲り，ここでは実行のアルゴリズムを簡潔に示す．まず，表記の簡単のために $\mathbf{U}^{(l)} = (\boldsymbol{u}_1^{(l)}, \boldsymbol{u}_2^{(l)}, ..., \boldsymbol{u}_T^{(l)})$,

$\mathbf{Z}^{(l)} = (\boldsymbol{z}_1^{(l)}, \boldsymbol{z}_2^{(l)}, ..., \boldsymbol{z}_T^{(l)})$ とおき，式 (2.34)(2.35) をそれぞれ

$$\mathbf{Z}^{(l)} = \sigma_\theta \left\{ \mathbf{U}^{(l)} \right\} \tag{2.39}$$

$$\mathbf{U}^{(l)} = \mathbf{W}^{(l)} \mathbf{Z}^{(l-1)} + \mathbf{b}^{(l)} \mathbf{1}_T^\top \tag{2.40}$$

と書き換える．ここで $\mathbf{1}_T$ は 1 を T 個並べたベクトルである．まず目的関数内の評価関数 $\mathcal{R}\left(\boldsymbol{y}_\tau, \mathcal{M}\left(\boldsymbol{x}_\tau | \Theta_\mathcal{M}\right)\right)$ を最終層の入力 $\boldsymbol{u}_\tau^{(L)}$ で偏微分した値を並べた行列である $\boldsymbol{\Delta}^{(L)}$ を計算する．

$$\boldsymbol{\delta}_\tau^{(L)} = \frac{\partial \mathcal{R}\left(\boldsymbol{y}_\tau, \mathcal{M}\left(\boldsymbol{x}_\tau | \Theta_\mathcal{M}\right)\right)}{\partial \boldsymbol{u}_\tau^{(L)}} \tag{2.41}$$

$$\boldsymbol{\Delta}^{(L)} = (\boldsymbol{\delta}_1^{(L)}, \boldsymbol{\delta}_2^{(L)}, ..., \boldsymbol{\delta}_T^{(L)}) \tag{2.42}$$

そして各層の，目的関数の l 層の入力 $\boldsymbol{u}_\tau^{(l)}$ による微分を以下で再帰的に計算する．

$$\boldsymbol{\Delta}^{(l)} = \sigma'_\theta \left\{ \mathbf{U}^{(l)} \right\} \odot \mathbf{W}^{(l+1)\top} \boldsymbol{\Delta}^{(l+1)} \tag{2.43}$$

ここで $\odot$ はアダマール積を表し，σ'_θ は σ_θ を $\boldsymbol{u}_\tau^{(l)}$ で要素ごとに微分した値を返す関数である．そして，$\boldsymbol{\Delta}^{(l)}$ を用いて重み行列とバイアスベクトルの勾配を計算する．

$$\partial \mathbf{W}^{(l)} = \frac{1}{T} \boldsymbol{\Delta}^{(l)} \mathbf{Z}^{(l-1)\top} \tag{2.44}$$

$$\partial \mathbf{b}^{(l)} = \frac{1}{T} \boldsymbol{\Delta}^{(l)} \mathbf{1}_T^\top \tag{2.45}$$

最後に重み行列とバイアスベクトルを勾配法で更新する．

$$\mathbf{W}^{(l)} \leftarrow \mathbf{W}^{(l)} + \lambda \partial \mathbf{W}^{(l)} \tag{2.46}$$

$$\mathbf{b}^{(l)} \leftarrow \mathbf{b}^{(l)} + \lambda \partial \mathbf{b}^{(l)} \tag{2.47}$$

なお，アルゴリズムの収束の安定化や高速化のために，式 (2.46)(2.47) の第二項の勾配項を調整する手法として，AdaGrad [83] や Adam [84] などの手法が提案されている．また，$\Theta_\mathcal{M}$ の学習は勾配法で行われるため，求まる解はその初期値に依存する．そこで，$\Theta_\mathcal{M}$ の初期値決定法として，制約付きボルツマンマシン（RBM: restricted Boltzmann machine）[85] や識別的事前学習法（discriminative pre-training）[86] などが提案されている．

2.4.2 深層学習におけるネットワーク構造に関する研究

本節では，ネットワーク構造の高度化に関する研究を概説する．音は物理現象であるため，音源情報を推定する写像関数には物理的な制約や聴覚に関する知見を含めるべきである．本節では，ネットワーク構造を工夫することで物理的な制約や聴覚に関する知見を含める代表的な方法として，畳み込みニューラルネットワーク（CNN: convolutional neural network）[39, 40] と再帰型ニューラルネットワーク（RNN: recurrent neural network）[41, 42] を概説する．

(a) 畳み込みニューラルネットワーク

CNN は，生物の脳の視覚野における神経細胞の受容野の局所性をヒントにしたネットワークである [39, 40]．銃声や悲鳴などの特定の音を検出する音源情報推定を考えたとき，火薬の爆発に起因する突発的音や，声の震えに起因する特定の周波数変化など，ある特定の時間周波数パターンを持った音だけを抽出したいことがある．特定の入力パターンを検知する方法として画像処理の分野では，2 次元畳み込み（convolution）が用いられている．2 次元畳み込みは，入力画像と特定のパターンを表す行列との相関をとることに近く，CNN は特定のパターンを表す行列をニューラルネットワークで学習する方法と考えることもできる．

DNN では層間の全ユニットが結合していた一方で，CNN では決まった少数のユニットとのみ結合をもつ．CNN の順伝搬にはプーリングやバッチ正則化 [87] などの処理が加えられるものの行列演算で記述できるため，その学習は DNN と同様に誤差逆伝搬を用いて実行できる．音源情報推定において CNN は，音声認識 [13, 43] や楽器音の発音認識 [88] に用いられており，DNN と比べ高い精度で音源情報推定ができることが示されている．

(b) 再帰型ニューラルネットワーク

RNN は時系列データの時間方向の関係性をモデル化するネットワークである [41]．音は時系列データであるため，入力は時間フレーム間で方向に強い依存関係がある．また発声のイントネーション（例えば，方言の違い）や音楽のリズムなど，音の時間変化に音源情報が埋め込まれていることもあるため，ネットワーク構造に時間方向の依存関係を表現する仕組みを組み込みたい．RNN は中間層に自分自身への帰還路を持たせ，情報を一時的に記憶させることで，時間方向への依存性を表現するニューラルネットワークである．RNN を用いた非線形写像は以下のように記述できる．

$$\mathcal{M}(\boldsymbol{x}|\Theta_{\mathcal{M}}) \leftarrow \sigma_{\mathcal{M}}\left\{\mathbf{W}^{\mathrm{out}}\boldsymbol{z}_{\tau} + \mathbf{b}^{\mathrm{out}}\right\} \tag{2.48}$$

$$\boldsymbol{z}_{\tau} = \mathbf{W}^{\mathrm{in}}\boldsymbol{x}_{\tau} + \mathbf{W}^{\mathrm{r}}\boldsymbol{z}_{\tau-1} + \mathbf{b}^{\mathrm{in}} \tag{2.49}$$

式 (2.49) の第二項が帰還路を表現しており，$\mathbf{W}^{\mathrm{r}}$ が過去の情報の帰還重みである．

RNN は帰還路を持つネットワークであるため，その学習には DNN と同様の誤差逆伝搬法を用いることができない．そのため RNN の学習には，帰還路を別の中間層とみなし，RNN を時間方向に展開して大規模なネットワークとみなす BPTT（backpropagation through time）法 [89] が用いられる．そのため，順伝搬の際には層数の少ないニューラルネットワーク構造であるが，誤差逆伝搬の際には非常に層の深いニューラルネットワークを学習することになる．ゆえに勾配消失問題が生じるため，RNN が記憶できる時間フ

レームは 10 フレーム程度であるといわれている [42]．そこでより長期の記憶を扱うために，外部記憶装置を持たせた長短記憶（LSTM: long short-term memory）ネットワーク [42] や，微分可ニューラルコンピュータ（DNC: differentiable neural computers）[90] なども提案されている．音源情報推定においては，音声認識 [44] や音源強調 [91, 32] などで用いられている．

2.4.3 深層学習における目的関数の設計の研究

本節では，目的関数の高度化に関する研究を概説する．まず観測信号と所望の音源情報のラベルデータのペアがある教師あり学習で広く用いられる，決定論的な目的関数を紹介する．次いで，ラベルデータの収集が困難な場合や，ラベルデータが一意に定まらない場合のニューラルネットワークの目的関数の例として，生成モデル学習と強化学習の目的関数を紹介する．

ニューラルネットワークの学習は誤差逆伝搬法で行われる．式 (2.41) から明らかなように，誤差逆伝搬を行うための目的関数設計の制約は，音源情報の推定精度が高いほど大きな値を返す評価関数 $\mathcal{R}\left(\boldsymbol{y}, \hat{\boldsymbol{y}}\right)$ を，ニューラルネットワークのパラメータ $\Theta_{\mathcal{M}}$ で微分可能な形で記述することである．この基準を満たす，教師あり学習のための代表的な目的関数に，二乗誤差最小化（MMSE: minimum mean square error）と交差エントロピーがある．

MMSE は回帰の問題に広く用いられる目的関数であり，評価関数 $\mathcal{R}\left(\boldsymbol{y}, \hat{\boldsymbol{y}}\right)$ を $\boldsymbol{y}$ と $\hat{\boldsymbol{y}}$ の二乗誤差としたものである．音源情報推定の中では，音源強調などの源信号を推定する問題などで用いられる．MMSE に基づく $\mathcal{J}\left(\Theta_{\mathcal{M}}\right)$ は，$\boldsymbol{y}_\tau$ は源信号のスペクトルなどを表すベクトルとして，

$$\mathcal{J}\left(\Theta_{\mathcal{M}}\right) = -\frac{1}{T}\sum_{\tau=1}^{T}\left|\boldsymbol{y}_\tau - \mathcal{M}\left(\boldsymbol{x}_\tau|\Theta_{\mathcal{M}}\right)\right|^2 \tag{2.50}$$

のように記述できる．源信号のスペクトルを推定する問題では，MMSE の代わりに一般化カルバックライブラーダイバージェンス（KLD: Kullback-Leibler divergence）[92] や板倉斎藤距離（ISD: Itakura-saito divergence）[93] などの目的関数を用いる研究もある．

一方，交差エントロピー関数は識別の問題に広く用いられる目的関数であり，評価関数 $\mathcal{R}\left(\boldsymbol{y}, \hat{\boldsymbol{y}}\right)$ を多項分布の対数尤度にしたものである．音源情報推定の中では，音声認識などの音源の種類を推定する問題などで用いられる．目的関数は $\boldsymbol{y}_\tau$ を音源の状態を 1-of-K 表現で表したベクトルとして

$$\mathcal{J}\left(\Theta_{\mathcal{M}}\right) = \frac{1}{T}\sum_{\tau=1}^{T}\sum_{k=1}^{K} z_{\tau,k} \ln \mathcal{M}\left(\boldsymbol{x}_\tau|\Theta_{\mathcal{M}}\right)_k \tag{2.51}$$

のように記述できる．

このように，観測信号と所望の音源情報のラベルデータのペアがある教師あり学習においては，目的関数がニューラルネットワークの出力値と学習データの値の誤差で求まる解析的に扱いやすい決定論的な形式で記述することができる．しかし十分な量のラベルデータがない場合や，ラベルを一意に定めることができない場合は出力の目標値を一意に定めることができないため，決定論的な形で目的関数を記述することができない．以下では，この問題を解決するために，確率的要素を含んだ形で目的関数を設計する例として生成モデル学習と強化学習を紹介する．

(a) 生成モデル学習

観測信号 $\boldsymbol{x}$ は得られるが，所望の音源情報のラベルデータ $\boldsymbol{y}$ が得られない場合の学習方法の一つに生成モデル学習がある．生成モデル学習とは，教師あり学習のようにラベルデータを推定するのではなく，観測信号を生成した確率分布（生成モデル）$p(\boldsymbol{x})$ を推定する学習法である．ゆえに，目的関数は MMSE や交差エントロピーのように決定論的な形ではなく，確率的な要素を含んだ形で記述される．近年，深層学習を用いた生成モデル学習法として変分オートエンコーダ（VAE: variational auto-encoder）[94] と敵対的生成ネットワーク（GAN: generative-adversarial networks）[95] が提案された．この 2 つの研究に共通する点は，誤差逆伝搬に用いる目的関数に観測信号や潜在変数の確率分布が満たすべき性質を記述したものを利用している点にある．

VAE は $p(\boldsymbol{x})$ の変分推論から導かれる生成モデルの学習法である．VAE では，観測信号を生成する原因や背後にある状態などを表現する潜在変数 $\boldsymbol{z}$ を導入する．音源情報推定において $\boldsymbol{z}$ は，音の突発性や周波数特性を表す音響特徴量と考えることもできる．変分推論の枠組みでは，観測信号を得た元での潜在変数の事後分布 $p(\boldsymbol{z}|\boldsymbol{x})$ を近似した分布 $q_{\Theta_{\mathcal{M}}}(\boldsymbol{z}|\boldsymbol{x})$ を考えることで，生成モデルの対数尤度 $\ln p(\boldsymbol{x})$ を以下のように書き換える．

$$\ln p(\boldsymbol{x}) = KL\left[q_{\Theta_{\mathcal{M}}}(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z}|\boldsymbol{x})\right] + \mathcal{L}(\Theta_{\mathcal{M}}, \Theta_{\mathcal{G}}|\boldsymbol{x}) \tag{2.52}$$

第一項は KLD，第二項は変分下界と呼ばれる値であり，KLD は以下のように計算できる．

$$KL\left[q(\boldsymbol{x})||p(\boldsymbol{x})\right] = -\int q(\boldsymbol{x}) \ln \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x} \tag{2.53}$$

KLD は必ず非負値であるため，VAE では変分下界を最大化することで生成モデルの対数尤度を最大化する．

$$\mathcal{L}(\Theta_{\mathcal{M}}, \Theta_{\mathcal{G}}|\boldsymbol{x}) = \int q_{\Theta_{\mathcal{M}}}(\boldsymbol{z}|\boldsymbol{x}) \ln \frac{p_{\Theta_{\mathcal{G}}}(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{q_{\Theta_{\mathcal{M}}}(\boldsymbol{z}|\boldsymbol{x})} d\boldsymbol{z} \tag{2.54}$$

$$= -KL\left[q_{\Theta_{\mathcal{M}}}(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})\right] + \int q_{\Theta_{\mathcal{M}}}(\boldsymbol{z}|\boldsymbol{x}) \ln p_{\Theta_{\mathcal{G}}}(\boldsymbol{x}|\boldsymbol{z}) d\boldsymbol{z} \tag{2.55}$$

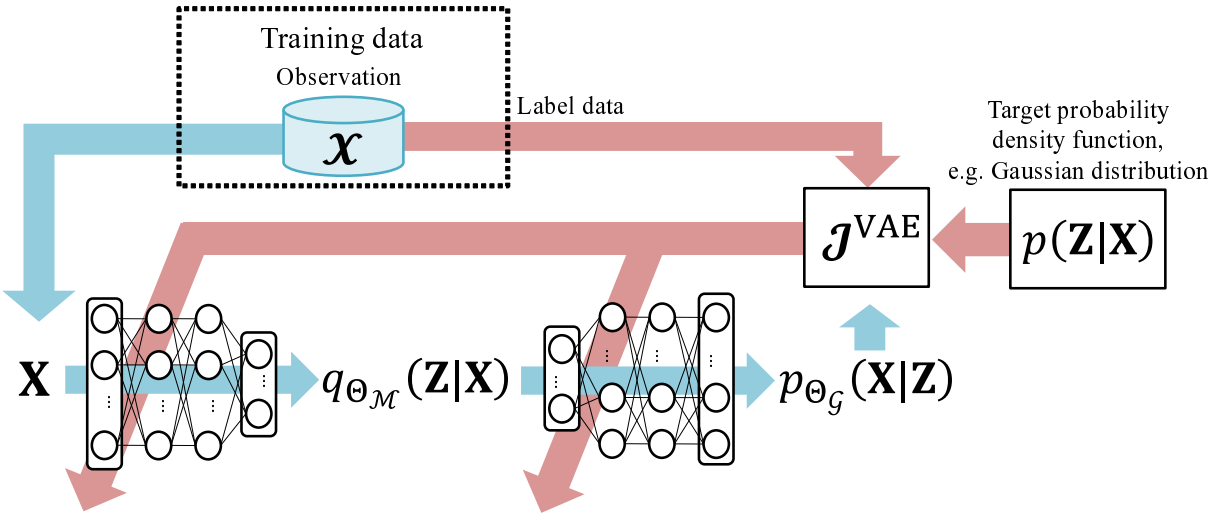


図 2.8: 変分オートエンコーダ．各ニューラルネットワークは観測信号や潜在変数の従う確率分布を出力する．出力された確率分布と目標とする確率分布の違いの最小化を目的関数とする．

式 (2.56) で具体化すべき項は観測信号を得た下での潜在変数の事後分布 $q_{\Theta_{\mathcal{M}}}(\boldsymbol{z}|\boldsymbol{x})$ と，潜在変数を得た下での観測信号の事後分布 $p_{\Theta_{\mathcal{G}}}(\boldsymbol{x}|\boldsymbol{z})$（生成モデル）である．VAE では図 2.8 のように，この 2 つの確率分布をニューラルネットワークで推定する[1]．具体的には，$q_{\Theta_{\mathcal{M}}}(\boldsymbol{z}|\boldsymbol{x})$ と $p_{\Theta_{\mathcal{G}}}(\boldsymbol{x}|\boldsymbol{z})$ をガウス分布などの確率分布で表現し，分布のパラメータをニューラルネットワークで推定する．以上より，VAE における目的関数は以下のようになる．

$$\mathcal{J}^{\mathrm{VAE}}(\Theta_{\mathcal{M}}, \Theta_{\mathcal{G}}) = \mathcal{L}(\Theta_{\mathcal{M}}, \Theta_{\mathcal{G}}|\boldsymbol{x}) \tag{2.56}$$

$$\Theta_{\mathcal{M}}, \Theta_{\mathcal{G}} \leftarrow \underset{\Theta_{\mathcal{M}}, \Theta_{\mathcal{G}}}{\arg\max}\, \mathcal{J}^{\mathrm{VAE}}(\Theta_{\mathcal{M}}, \Theta_{\mathcal{G}}) \tag{2.57}$$

VAE は，誤差逆伝搬の過程で生成モデルの勾配を計算する必要があるため，生成モデルを微分可能な確率分布として陽に表現していた．一方 GAN は，生成モデルを確率分布として陽に表現せずに生成モデルを学習する方法である．今，観測信号を生成した真の分布を $p^*(\boldsymbol{x})$ と考える．推定した生成モデル $\hat{p}(\boldsymbol{x})$ が $p^*(\boldsymbol{x})$ と完全に一致するならば，$\hat{p}(\boldsymbol{x})$ から生成されたデータは学習データと見分けがつかなくなり，その密度比は $p^*(\boldsymbol{x})/\hat{p}(\boldsymbol{x}) = 1$ を満たす．つまり，生成モデルを確率分布として陽に表現しなくとも，生成モデルから生成されたデータと学習データの見分けがつかなくなるように $\hat{p}(\boldsymbol{x})$ を学習すれば，生成モデルは学習できる．GAN ではこの考え方に基づき，潜在変数から $\boldsymbol{x}$ を生成するニューラルネットワーク $\mathcal{G}(\boldsymbol{z}|\Theta_{\mathcal{G}})$ と，$\boldsymbol{x}$ が真の分布から生成されたものか疑

[1]なお，2 つのニューラルネットワークを用いて $q_{\Theta_{\mathcal{M}}}(\boldsymbol{z}|\boldsymbol{x})$ と $p_{\Theta_{\mathcal{G}}}(\boldsymbol{x}|\boldsymbol{z})$ を表現し，変分下界を最大化するようにニューラルネットワークを学習する考え方は，Dayan らが Helmholtz Machine [96, 97] として 1995 年に発表している．

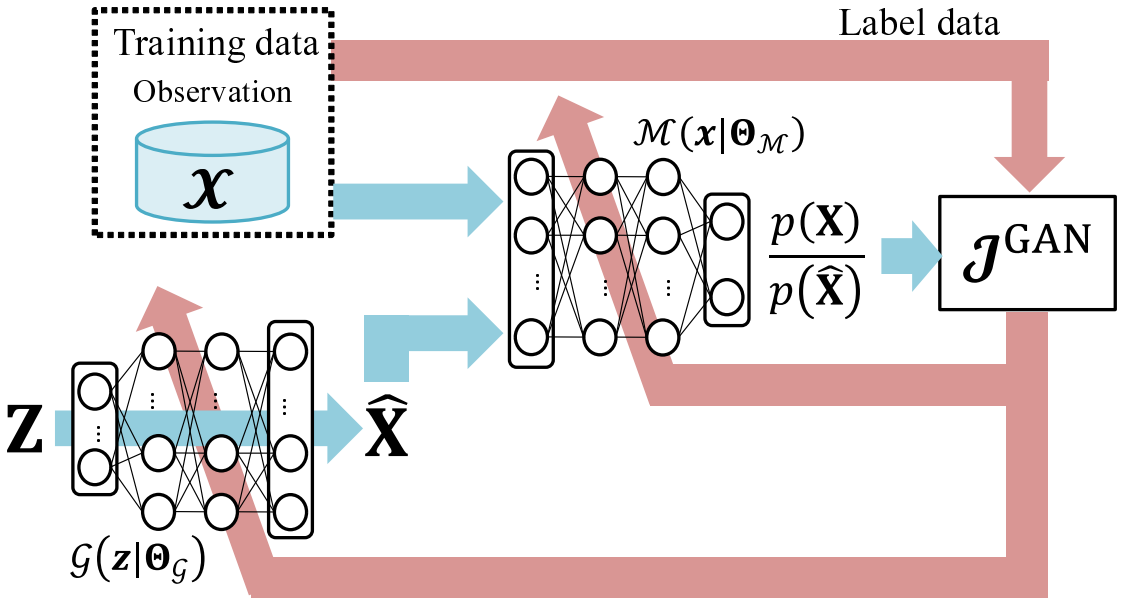


図 2.9: 敵対的生成ネットワーク．観測信号を疑似生成するニューラルネットワークと，入力が観測信号の学習データである確率と疑似生成されたデータである確率の密度比を出力するネットワークを用いる．

似生成されたものかを識別するニューラルネットワーク $\mathcal{M}(\boldsymbol{x}|\Theta_{\mathcal{M}})$ の 2 つのニューラルネットワークを利用する（図 2.9）．$\mathcal{G}(\boldsymbol{z}|\Theta_{\mathcal{G}})$ は $\mathcal{M}(\boldsymbol{x}|\Theta_{\mathcal{M}})$ をだます（誤識別）ことが目的であり，$\mathcal{M}(\boldsymbol{x}|\Theta_{\mathcal{M}})$ は $\mathcal{G}(\boldsymbol{z}|\Theta_{\mathcal{G}})$ から生成されたものを見破る（正識別）ことが目的である．ゆえに 2 つのニューラルネットワークは以下の目的関数を用いて “敵対的” に学習される．

$$\mathcal{J}^{\mathrm{GAN}}(\Theta_M, \Theta_{\mathcal{G}}) = \int p^*(\boldsymbol{x}) \ln \mathcal{M}(\boldsymbol{x}|\Theta_{\mathcal{M}}) d\boldsymbol{x} + \int \hat{p}(\boldsymbol{x}) \ln \left(1 - \mathcal{M}(\boldsymbol{x}|\Theta_{\mathcal{M}})\right) d\boldsymbol{x} \tag{2.58}$$

$$\Theta_{\mathcal{M}} \leftarrow \arg\min_{\Theta_{\mathcal{M}}} \mathcal{J}^{\mathrm{GAN}}(\Theta_{\mathcal{M}}, \Theta_{\mathcal{G}}) \tag{2.59}$$

$$\Theta_{\mathcal{G}} \leftarrow \arg\max_{\Theta_{\mathcal{G}}} \mathcal{J}^{\mathrm{GAN}}(\Theta_{\mathcal{M}}, \Theta_{\mathcal{G}}) \tag{2.60}$$

$\Theta_{\mathcal{M}}$ と $\Theta_{\mathcal{G}}$ の目的関数は同じ形をしているにもかかわらず，識別のパラメータ $\Theta_{\mathcal{M}}$ は識別率を最大化するよう，生成のパラメータ $\Theta_{\mathcal{G}}$ は識別率を最小化するように学習される．

ここまで，生成モデル学習に基づくニューラルネットワーク学習のための目的関数設計を説明してきた．これらの研究における着目すべき考え方は，観測信号や潜在変数の確率分布が満たすべき性質を評価関数として目的関数を設計し，その性質を満たすようにニューラルネットワークを学習する点にある．

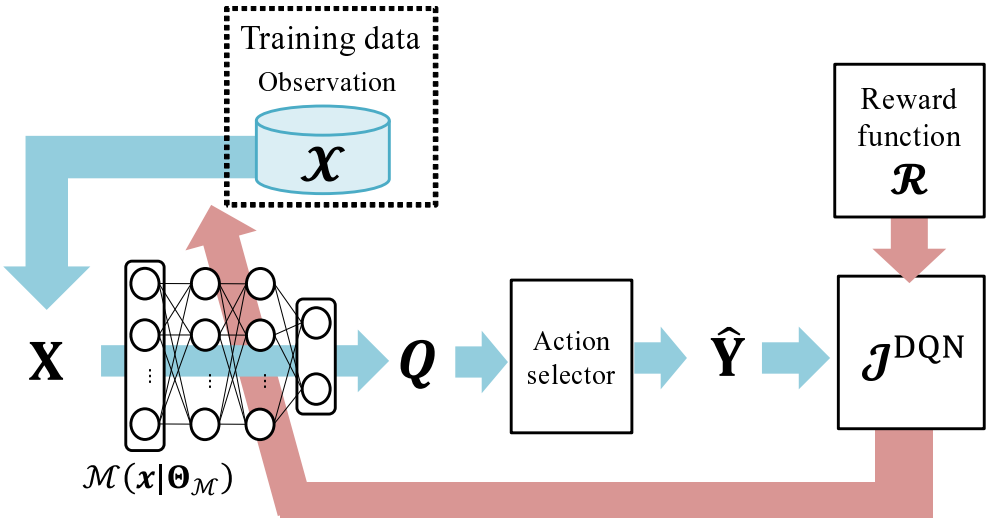


図 2.10: Deep Q-network．ニューラルネットワークは行動を選択/生成する．明示的なラベルデータの代わりに行動に関する報酬を出力する報酬関数を導入し，その報酬を最大化する行動を確率的に探索する．

(b) 強化学習

ラベルデータが得られないが，ニューラルネットワークの出力を評価できる場合の学習方法の一つに強化学習がある．強化学習 [98] は，明示的にラベルデータを用意する代わりに出力の良悪を判定する報酬関数を導入して学習を行う枠組みであり，ゲームにおける行動の最適化などで用いられている [99, 100, 101]．例えばゲームの行動を学習するときは，より高いゲームスコアを得るように行動選択することが最終目標となる．しかし，ゲームスコアを最大化するための明示的な教師データ（行動）を用意することはできない．そこで強化学習では，自身の行動方針に従いゲームを行ない，その結果，良いゲームスコアを得られれば現在の行動方針は正しい，悪いゲームスコアを得たならば現在の行動方針は誤っていると解釈しながら，試行錯誤を通じて行動方針を学習していく．

強化学習の代表的なアルゴリズムに “Q 学習” がある．Q 学習では，観測 $\boldsymbol{x}_\tau$（ゲームの画面）を得た下で，有限個の行動の候補 $\mathcal{A} = \{a_1, a_2, ..., a_A\}$（ゲームのボタン操作）の中からどの行動を選択するべきかを判断する関数 $Q(\boldsymbol{x}_\tau, a)$ を学習する[2]．ここで $Q(\boldsymbol{x}_\tau, a)$ は Q 関数と呼ばれ，“$\boldsymbol{x}_\tau$ を観測したときに行動 a を選択し，その後も現在の方策にしたがい行動を続けた際に得られる報酬の和の期待値を返す関数 [98]” を表し，簡潔に言えば $\boldsymbol{x}_\tau$ の下で行動 a を取るメリットを返す関数である．DQN（Deep Q-network）[100, 101] では，Q 関数をニューラルネットワークを用いて $\mathcal{M}(\boldsymbol{x}_\tau|\Theta_{\mathcal{M}}) = (Q(\boldsymbol{x}_\tau, a_1), ..., Q(\boldsymbol{x}_\tau, a_A))$

[2]行動の選択問題から，行動の生成問題へ拡張した研究もある [102].

で表現する（図 2.10）．時刻 τ における行動は

$$a_\tau \leftarrow \arg\max_a \mathcal{M}(\boldsymbol{x}_\tau|\Theta_\mathcal{M}) \tag{2.61}$$

で選択され，選択した行動に応じて報酬 r_τ（ゲームスコア）を得る．強化学習において最適な行動は，報酬の和（最終的なゲームスコア）を最大化する行動であるため，目的関数は以下のように記述できる．

$$\mathcal{J}^{\mathrm{DQN}}(\Theta_\mathcal{M}) = \sum_{\tau=1}^{T} \gamma^{t-1} r_\tau \tag{2.62}$$

$$\Theta_\mathcal{M} \leftarrow \arg\max_{\Theta_\mathcal{M}} \mathcal{J}^{\mathrm{DQN}}(\Theta_\mathcal{M}) \tag{2.63}$$

ここで γ は報酬の輪が無限大に発散しないようにするための定数である．このように強化学習の目的関数は自身が確率的に選択した行動に依存してその値が決定する．ゆえに目的関数は教師あり学習のように決定論的には定まらず確率的な形式でしか記述できない．その代わりに，強化学習に用いる報酬関数は誤差逆伝搬で要求されるような微分可能な関数である必要がなく，また観測信号やニューラルネットワークに対する最適な出力に関する完全な理解が無くても設計出来るため，ラベルデータが一意に定まらない問題においてもニューラルネットワークを学習できる．この研究における着目すべき考え方は，非線形写像の出力を評価する関数を別に用意し，その評価値を最大化する非線形写像の出力を確率的に探索することでニューラルネットワークを学習する点にある．

2.5 深層学習に基づく音源強調

2.1 節では，時間周波数マスクを用いた音源強調は，観測信号から時間周波数マスクやその設計のためのパラメータを求める必要があることを述べた．そしてその推定のための手がかりには，音源の方向などの物理的な特性だけなく，音源の種類などの潜在的な音源情報を併用する必要があることを述べた．2.3 節と 2.4 節では，潜在的な音源情報を推定する手段として深層学習を紹介した．これらの研究の流れに基づく音源強調の新たな方向性として，深層学習を音源強調に応用し，潜在的な音源情報を併用しながら時間周波数マスクを推定する研究が行われている [32, 33, 34, 45, 46, 47, 48, 91, 70, 103, 104, 105, 106, 107]．すなわち，観測信号を時間周波数マスクを計算するために必要なパラメータへ写像する非線形関数としてニューラルネットワークを利用する．時間周波数マスクを計算するためのパラメータは，振幅スペクトルやパワースペクトルなどの実数変数であるため，ニューラルネットワークは回帰関数として用いられる．以下では，図 2.11 に示した深層学習を用いた音源強調の概要に沿って，ニューラルネットワークの利用法と学習法を説明する．

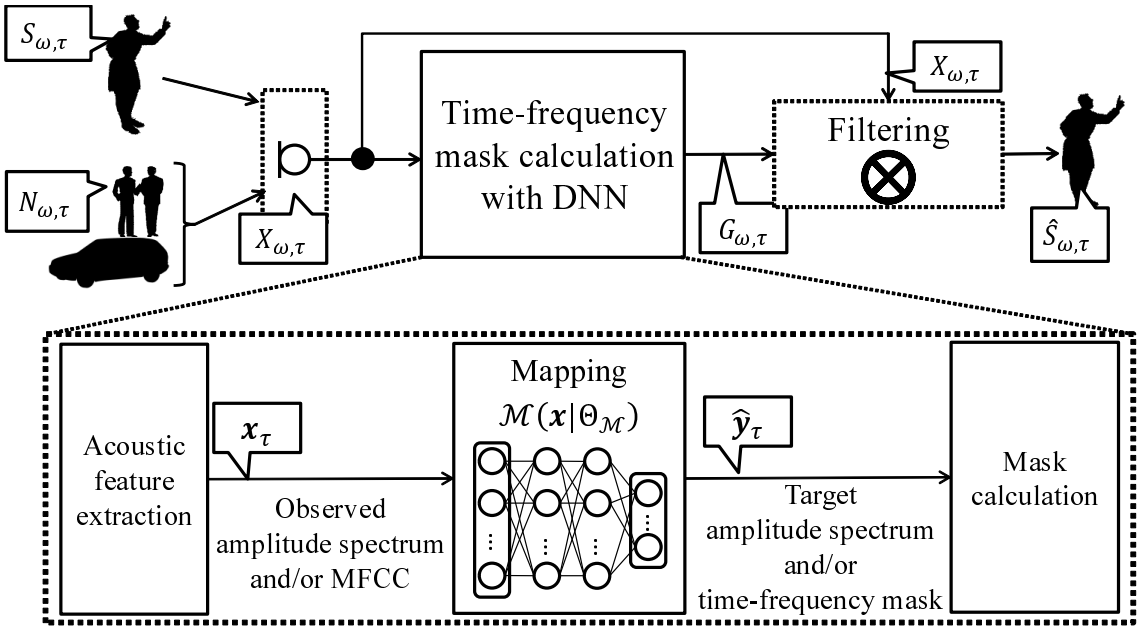


図 2.11: 深層学習に基づく音源強調.

まず，深層学習を用いた時間周波数マスクを計算するための最も基本的な方法として，全結合多層ニューラルネットワーク（DNN）を用いた時間周波数マスク計算を説明する [47, 48, 33, 34, 70]．DNN を学習するために，観測信号 $X_{\omega,\tau}$ とラベルデータ $\boldsymbol{y}_\tau \in \mathbb{R}^{\mathsf{D}}$（時間周波数マスクを計算するためのパラメータ）を生成する．式 (2.3) からわかるように，観測信号は源信号 $S_{\omega,\tau}$ と雑音 $N_{k,\omega,\tau}$ のデータから生成できるため，深層学習に基づく音源強調に必要な学習データは源信号 $\{\boldsymbol{S}_\tau|\tau = 1, ..., T\}$ と雑音 $\{\boldsymbol{N}_\tau|\tau = 1, ..., T\}$ である．ただし

$$\boldsymbol{S}_\tau = (S_{1,\tau}, S_{2,\tau}, ..., S_{\Omega,\tau}) \tag{2.64}$$

$$\boldsymbol{N}_\tau = \left(\sum_{k=1}^{K} N_{k,1,\tau}, \sum_{k=1}^{K} N_{k,2,\tau}, ... \sum_{k=1}^{K} N_{k,\Omega,\tau}\right) \tag{2.65}$$

である．

そして，式 (2.3) を用いて生成された観測信号 $X_{\omega,\tau}$ から，DNN への入力ベクトル $\boldsymbol{x}_\tau$（音響特徴量）を計算する．学習データが十分にあり，多くの隠れ層や隠れユニットを持つ DNN を利用できるならば，DNN はニューラルネットの内部で自動的に音響特徴量を抽出できるといわれている．そのため式 (2.36) で述べたように，DNN へ入力する音響特徴量は観測信号 $X_{\omega,\tau}$ の時間周波数要素を並べたベクトルとすることが多い [38, 81]．

$$\boldsymbol{x}_\tau = (\boldsymbol{X}_{\tau-P}, ..., \boldsymbol{X}_\tau, ..., \boldsymbol{X}_{\tau+P})^\top \tag{2.66}$$

$$\boldsymbol{X}_\tau = (\ln|X_{1,\tau}|, ..., \ln|X_{\Omega,\tau}|) \tag{2.67}$$

一方学習データが十分にない場合は，隠れ層や隠れユニットの数を削減するために，事前に人手で選択した音響特徴量を抽出し，$\boldsymbol{x}_\tau$ とすることもある．この音響特徴量には，メル周波数ケプストラム係数（MFCC: mel-frequency cepstrum coefficient）や線スペクトル対（LSP: line spectral pairs）などが用いるられる．

次に，DNN の出力である，時間周波数マスクを計算するためのパラメータを並べた D 次元ベクトルを $\boldsymbol{y}_\tau \in \mathbb{R}^{\mathsf{D}}$ を定義する．例えば源信号の振幅スペクトル $|S_{\omega,\tau}|$ から時間周波数マスクを計算する場合は，

$$\boldsymbol{y}_\tau = \left(|S_{1,\tau}|, |S_{2,\tau}|, ..., |S_{\Omega,\tau}|\right)^\top \tag{2.68}$$

である．

最後に DNN の順伝搬の式 (2.33)(2.34)(2.35) を利用し，以下のように $\boldsymbol{y}_\tau$ を推定する．

$$\hat{\boldsymbol{y}}_\tau \leftarrow \mathcal{M}\left(\boldsymbol{x}_\tau|\Theta_\mathcal{M}\right) = \sigma_\mathcal{M}\left\{\boldsymbol{u}_\tau^{(L)}\right\} \tag{2.69}$$

$$\boldsymbol{z}_\tau^{(l)} = \sigma_\theta\left\{\boldsymbol{u}_\tau^{(l)}\right\} \tag{2.70}$$

$$\boldsymbol{u}_\tau^{(l)} = \mathbf{W}^{(l)}\boldsymbol{z}_\tau^{(l-1)} + \mathbf{b}^{(l)} \tag{2.71}$$

ここで L，$\mathbf{W}^{(l)}$，$\mathbf{b}^{(l)}$，$\Theta_\mathcal{M}$，σ_θ，$\sigma_\mathcal{M}$ の定義は前節で説明したものと同じである．ただし DNN 音源強調は回帰問題であるため，$\sigma_\mathcal{M}$ の活性化は省略することが多く，また $\boldsymbol{z}_\tau^{(1)} = \boldsymbol{x}_\tau$ である．また，ニューラルネットワークのパラメータ $\Theta_\mathcal{M}$ は，式 (2.50) で説明した MMSE などを目的関数とし，誤差逆伝搬法で学習する．

$$\Theta_\mathcal{M} \leftarrow \arg\max_{\Theta_\mathcal{M}} \mathcal{J}\left(\Theta_\mathcal{M}\right) \tag{2.72}$$

$$\mathcal{J}\left(\Theta_\mathcal{M}\right) = -\frac{1}{T}\sum_{\tau=1}^{T}\left|\boldsymbol{y}_\tau - \mathcal{M}\left(\boldsymbol{x}_\tau|\Theta_\mathcal{M}\right)\right|^2 \tag{2.73}$$

ここまで，深層学習に基づく音源強調の最も単純な実装として，全結合多層ニューラルネットワークを利用した音源強調と，MMSE に基づくニューラルネットワークのパラメータ学習を説明した．近年では，より源信号の推定精度を向上させるために，ネットワーク構造，および目的関数の高度化の研究が行われている．以降では，それぞれの研究について説明する．

(a) ネットワーク構造に関する従来研究

近年研究が進められているネットワーク構造の高度化では，RNN や LSTM を用いた時系列のモデル化 [32, 91, 103] や，振幅スペクトルの非負性のモデル化など，音響信号の物理的性質を記述するものが多い．これは，音は物理現象であるため，物理的な性質を利用した方が観測信号と音源情報の関係性を記述するモデルを設計しやすいと考えら

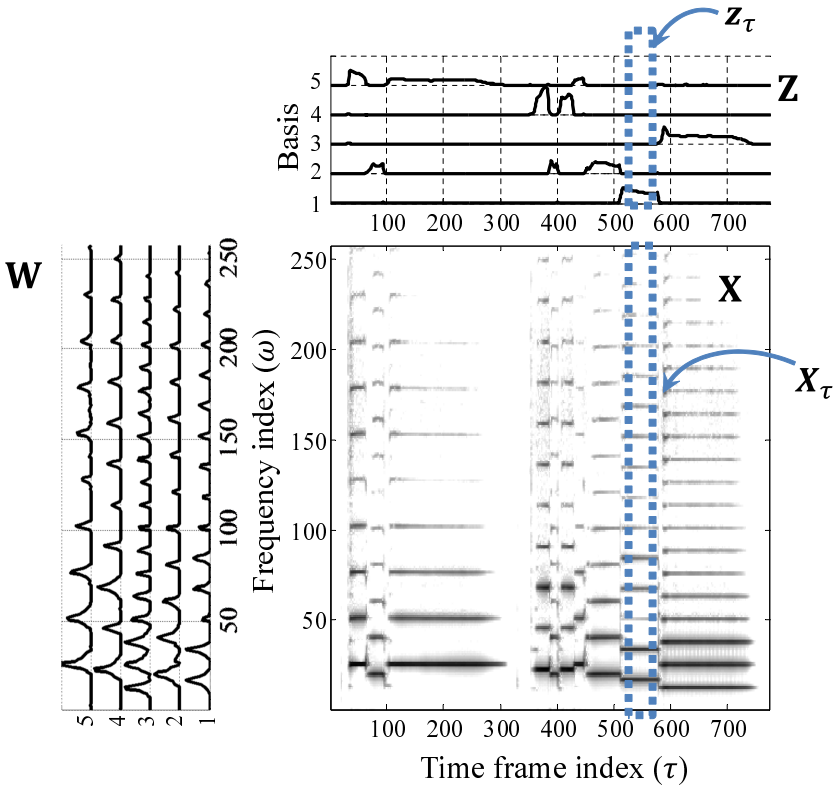


図 2.12: 非負値行列因子分解（NMF: non-negative matrix factorization）.

れるためである．時系列のモデル化については，RNN や LSTM を応用するものがほとんどである．RNN や LSTM については前節で説明したため省略し，ここでは音源強調特有のネットワーク構造である，振幅スペクトルの非負性のモデル化について説明する．

振幅スペクトルの非負性を利用して時間周波数マスクのパラメータを推定する手法として，非負値行列因子分解（NMF: non-negative matrix factorization）がある [72] の構造をニューラルネットワークの一種であるオートエンコーダを用いて表現する，非負オートエンコーダ（NAE: non-negative auto-encoder）を説明する [45, 46]．ウィナーマスクや IRM などの時間周波数マスクは，源信号と雑音の振幅スペクトルから設計されるマスクである．振幅スペクトルやパワースペクトルは音源の強度を表す値のためその値は必ず非負になるという物理的な性質がある．そのため時間周波数マスクのパラメータを推定する際には，振幅スペクトルの非負性を利用したい．NMF は，振幅スペクトルの非負性を利用して，源信号や雑音の振幅スペクトル推定する代表的な手法の一つである [72]．NMF では，振幅スペクトルの加法性やスパース性を仮定し，振幅スペクトルを時間周波数方向に並べた行列（振幅スペクトログラム）$\mathbf{X} \in \mathbb{R}_{0\leq}^{\Omega\times T}$ を，源信号と雑音の振幅スペクトルの U 個の基底を表す基底行列 $\mathbf{W} \in \mathbb{R}_{0\leq}^{\Omega\times \mathsf{U}}$ とその混合強度を表す強度行列 $\mathbf{Z} \in \mathbb{R}_{0\leq}^{\mathsf{U}\times T}$ に分解する手法である（図 2.12 ）．NMF による振幅スペクトルの分解は，以

下のように記述できる．

$$\mathbf{X} \approx \mathbf{W}\mathbf{Z} \tag{2.74}$$

$$\mathbf{X} = (\boldsymbol{X}_1, ..., \boldsymbol{X}_T) \tag{2.75}$$

$$\boldsymbol{X}_\tau = (|X_{1,\tau}|, ..., |X_{\Omega,\tau}|)^\top \tag{2.76}$$

$$\mathbf{W} = (\boldsymbol{w}_1, ..., \boldsymbol{w}_\Omega)^\top \tag{2.77}$$

$$\boldsymbol{w}_\omega = (w_{\omega,1}, ..., w_{\omega,\mathsf{U}})^\top \tag{2.78}$$

$$\mathbf{Z} = (\boldsymbol{z}_1, ..., \boldsymbol{z}_T) \tag{2.79}$$

$$\boldsymbol{z}_\tau = (z_{1,\tau}, ..., z_{\mathsf{U},\tau})^\top \tag{2.80}$$

そして，$\mathbf{W}$ と $\mathbf{Z}$ は，式 (2.75) を満たすように求められる．例えば，二乗誤差関数の最小化に基づき $\mathbf{W}$ と $\mathbf{Z}$ を求める場合は，以下となる．

$$\mathbf{W}, \mathbf{Z} \leftarrow \arg\max_{\mathbf{W},\mathbf{Z}} -\sum_{\omega=1}^{\Omega}\sum_{\tau=1}^{T}\left||X_{\omega,\tau}| - \sum_{u=1}^{\mathsf{U}} w_{\omega,u}z_{u,\tau}\right|^2 \tag{2.81}$$

最後に，源信号の振幅スペクトルに対応する基底 U_S だけを用いて信号を再構成することで，源信号の振幅スペクトルを推定する．

$$|\hat{S}_{\omega,\tau}| \leftarrow \sum_{u\in\mathsf{U}_S} w_{\omega,u}z_{u,\tau} \tag{2.82}$$

このことからわかるように，NMF において源信号の振幅スペクトルの推定は，観測信号からのアクティベーション行列 $\mathbf{Z}$ の推定と，$\mathbf{Z}$ を用いた源信号の再構成の二段階処理で行われる．NAE は，オートエンコーダの学習において基底行列 $\mathbf{W}$ に非負制約を加えて学習し，隠れ層の活性化関数をランプ関数とすることで，式 (2.74) による振幅スペクトログラムの行列積による表現を，オートエンコーダにおける入力信号の再構成とみなすものである．NAE の定義は以下のように記述できる．

$$\hat{\boldsymbol{S}}_\tau = \mathbf{W}_s\mathrm{ReLU}\left(\mathbf{W}_s^\top\boldsymbol{x}_\tau\right) \tag{2.83}$$

$$\hat{\boldsymbol{N}}_\tau = \mathbf{W}_n\mathrm{ReLU}\left(\mathbf{W}_n^\top\boldsymbol{x}_\tau\right) \tag{2.84}$$

$$\mathbf{W}_s, \mathbf{W}_n \leftarrow \arg\max_{\mathbf{W}_s,\mathbf{W}_n} -\sum_{\tau=1}^{T}\mathcal{D}(\boldsymbol{x}_\tau, (\hat{\boldsymbol{S}}_\tau + \hat{\boldsymbol{N}}_\tau)) \tag{2.85}$$

ここで $\hat{\boldsymbol{S}}_\tau$ と $\hat{\boldsymbol{N}}_\tau$ は式 (2.64)(2.65) で表現される源信号と雑音の振幅スペクトルの推定値，$\mathbf{W}_s \in \mathbb{R}_{0\leq}^{\Omega\times\mathsf{U}}$ と $\mathbf{W}_s \in \mathbb{R}_{0\leq}^{\Omega\times\mathsf{U}}$ は事前に学習された重み行列，ReLU はランプ関数，$\mathcal{D}$ は二乗誤差や一般化 KL ダイバージェンスなどの評価関数を表す．そして式 (2.85) の目的関数を用いて学習された $\mathbf{W}_s \in \mathbb{R}_{0\leq}^{\Omega\times\mathsf{U}}$ を用いて，観測信号から $\hat{\boldsymbol{S}}_\tau$ を推定し，時間周波数マ

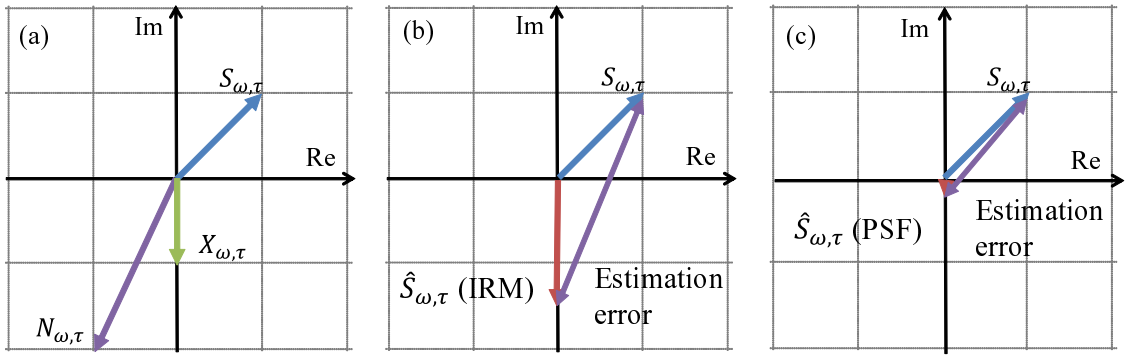


図 2.13: 位相鋭敏フィルタの目的関数を用いた音源強調．(a) 源信号，雑音および観測信号，(b) 源信号の振幅スペクトルの二乗誤差最小化による源信号の推定結果（式 (2.73)），(c) PSF の目的関数による源信号の推定結果

スクを設計する．先行研究では，ニューラルネットワークの学習に振幅スペクトルの非負性という物理制約を加えることで，同じ 3 層オートエンコーダよりも音源強調精度が向上することや，学習結果の解釈の容易化ができることが知られている．

(b) 目的関数の研究

近年進められている目的関数の高度化の研究では，ラベルデータが大量に存在する仮定の下で目的関数を拡張する研究が多い．高度化の方針には大きく分けて 2 つの方向性があり，1 つが時間周波数マスクと目的関数の複素数への拡張 [32, 91]，もう 1 つが目的関数の確率モデル化である [104, 106, 107]．

まず複素数への拡張を説明する．ウィナーマスクや IRM などの時間周波数マスクは，観測信号の振幅スペクトルを源信号に近づけるためのマスクである．ゆえに源信号の位相スペクトルは推定しておらず，逆フーリエ変換の際に出力音に歪みが生じ，音声認識率や音質低下の原因となっていた．これを解決するためには，複素スペクトルの推定誤差を評価できる目的関数を利用する必要がある．近年，位相を考慮した時間周波数マスク計算のための目的関数である，位相鋭敏フィルタ (PSF: Phase sensitive filter)[32, 91] が提案されている．

PSF と従来のニューラルネットワークを用いた時間周波数マスク計算との間には異なる点が二つある．一つは，ニューラルネットワークで時間周波数マスク $G_{\omega,\tau} \in [0,1]$ を直接求める点，もう一つが目的関数を複素領域の絶対誤差の二乗和としている点である．まず時間周波数マスク $G_{\omega,\tau} \in [0,1]$ を直接求めるために，時間周波数マスクを並べたベクトルを $\boldsymbol{G}_\tau = (G_{1,\tau}, G_{2,\tau}, ..., G_{\Omega,\tau})^\top$ とする．そして出力層の活性化関数をシグモイド関

数としたニューラルネットワークを用いて $\boldsymbol{G}_\tau$ を推定する．

$$\boldsymbol{G}_\tau \leftarrow \mathcal{M}(\boldsymbol{x}_\tau|\Theta_\mathcal{M}) = \mathrm{sigmoid}(\boldsymbol{u}_\tau^{(L)}) \tag{2.86}$$

そしてニューラルネットワークのパラメータを，以下の目的関数を最大化するように学習する．

$$\begin{aligned}\mathcal{J}_\mathcal{M}(\Theta) &= -\sum_{\tau=1}^{T}\sum_{\omega=1}^{\Omega}|S_{\omega,\tau} - G_{\omega,\tau}X_{\omega,\tau}|^2 \\ &= -\sum_{\tau=1}^{T}\sum_{\omega=1}^{\Omega}\Re(S_{\omega,\tau} - G_{\omega,\tau}X_{\omega,\tau})^2 + \Im(S_{\omega,\tau} - G_{\omega,\tau}X_{\omega,\tau})^2\end{aligned} \tag{2.87}$$

ここで $\Re(\cdot)$ と $\Im(\cdot)$ はそれぞれ，複素数の実部と虚部を表す．時間周波数マスク $G_{\omega,\tau} \in [0,1]$ を直接求めることで，式 (2.73) のような時間周波数マスクのパラメータ推定のための目的関数でなく，源信号 $S_{\omega,\tau}$ の推定精度を直接評価する目的関数を利用できるようになる．この改良を行うことで，雑音の音量が大きいときに源信号の推定精度を向上させることを図 2.13 を用いて説明する．図 2.13(a) のように雑音が大きいとき，振幅スペクトルの二乗誤差最小化で雑音を推定すると，複素領域ではその誤差が増大してしまう（図 2.13(b)）．一方 PSF の目的関数では，振幅スペクトルの推定誤差は増大するものの，複素スペクトル全体で考えたときには誤差が減少していることがわかる（図 2.13(c)）．PSF に基づく目的関数で時間周波数マスクを推定することで，音声認識率が向上することが知られている [32, 91]．PSF の他にも，源信号の推定誤差を複素領域で評価する方法として，複素理想比率マスク (cIRM: complex ideal ratio mask) なども提案されている [108]．

次に，目的関数の確率モデル化を説明する．式 (2.73) で与えられる MMSE は，$\boldsymbol{y}_\tau$ は $\mathcal{M}(\boldsymbol{x}_\tau|\Theta_\mathcal{M})$ を平均値に持つガウス分布からのサンプルであると考え，尤度最大化基準に則り $\Theta_\mathcal{M}$ を学習しているととらえることができる．これは，以下の式変形を行うことで，MMSE 推定と最尤推定が等価となることを利用している．

$$\begin{aligned}\mathcal{J}(\Theta_\mathcal{M}) &= -\frac{1}{T}\sum_{\tau=1}^{T}|\boldsymbol{y}_\tau - \mathcal{M}(\boldsymbol{x}_\tau|\Theta_\mathcal{M})|^2 \\ &\propto \frac{1}{T}\sum_{\tau=1}^{T} -\frac{1}{2}(\boldsymbol{y}_\tau - \mathcal{M}(\boldsymbol{x}_\tau|\Theta_\mathcal{M}))^\top \boldsymbol{I}_\mathsf{D}(\boldsymbol{y}_\tau - \mathcal{M}(\boldsymbol{x}_\tau|\Theta_\mathcal{M})) \\ &\propto \prod_{\tau=1}^{T}\frac{1}{(2\pi)^{\mathsf{D}/2}\sqrt{|\boldsymbol{I}_\mathsf{D}|}}\exp\left\{-\frac{1}{2}(\boldsymbol{y}_\tau - \mathcal{M}(\boldsymbol{x}_\tau|\Theta_\mathcal{M}))^\top \boldsymbol{I}_\mathsf{D}(\boldsymbol{y}_\tau - \mathcal{M}(\boldsymbol{x}_\tau|\Theta_\mathcal{M}))\right\} \\ &= \prod_{\tau=1}^{T}\mathcal{N}(\boldsymbol{y}_\tau|\mathcal{M}(\boldsymbol{x}_\tau|\Theta_\mathcal{M}, \boldsymbol{I}_\mathsf{D}))\end{aligned} \tag{2.88}$$

ここで $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ は平均 $\boldsymbol{\mu}$，共分散行列 $\boldsymbol{\Sigma}$ をパラメータにもつ多変量ガウス分布，$\boldsymbol{I}_\mathsf{D}$ は D 次元単位行列である．ここで着目すべき考え方は，ニューラルネットワークの出力

を，所望の音源情報の推定量ではなく観測信号を得た下での所望の音源情報の条件付き分布のパラメータと捉えることができる点である．ニューラルネットワークの出力を条件付き分布のパラメータととらえることで，式 (2.88) は以下のように一般化できる．

$$\begin{aligned}\mathcal{J}\left(\Theta_{\mathcal{M}}\right) &= \prod_{\tau=1}^{T} p(\boldsymbol{y}_\tau|\mathcal{M}(\boldsymbol{x}_\tau|\Theta_{\mathcal{M}})) \\ &\propto \sum_{\tau=1}^{T} \ln p(\boldsymbol{y}_\tau|\mathcal{M}(\boldsymbol{x}_\tau|\Theta_{\mathcal{M}}))\end{aligned} \tag{2.89}$$

このような拡張を行うことで，時間周波数マスクを計算するためのパラメータの推定誤差がガウス分布に従わないときの時間周波数マスクを推定できるようになる．木下らは以上の考え方に基づき，ニューラルネットワークの出力を混合ガウス分布のパラメータとして，時間周波数マスクを計算するためのパラメータを MAP 推定する手法を提案している [104]．また Hershey らはバイナリマスクの推定をクラスタリング問題に置き換える手法を提案している [106, 107]．

2.5.1 深層学習に基づく音源強調の課題

本節では，深層学習に基づく音源強調について説明してきた．音源強調の代表的な手法として，時間周波数マスクを用いた非線形フィルタリングを説明した．深層学習に基づく音源強調では，ニューラルネットワークを回帰関数として用いて源信号の振幅スペクトルを推定することを説明した．その学習は，個別に収録された源信号と雑音の学習データから，観測信号とパラメータ（振幅スペクトルなどのラベルデータ）の学習データを大量に用意し，ニューラルネットワークの出力とラベルデータの二乗誤差最小化基準などの目的関数で行われることを述べた．近年の目的関数の発展として，ニューラルネットワークの出力を観測信号を得た下での所望の音源情報の条件付き分布のパラメータと考え，ラベルデータに対する尤度の最大化を目的関数とする，確率モデル化の考え方を紹介した．この研究で着目すべき点は，音源強調のためのニューラルネットワークの学習を明示的に最尤推定と考えることで，目的関数に確率論的な考え方を導入できる点にある．しかし依然としてニューラルネットワークの学習は教師あり学習でありラベルデータを大量に必要とする．そのため 1 章で述べたような，ラベルデータが大量に集まらないケースや，ラベルデータが一意に定まらないようなケースでは，従来法をそのまま適用するだけでは源信号を推定できない．このような問題を解決するためには，確率論的目的関数を発展させる必要があると考えた．

3 章では，スポーツ競技音などのラベルデータを大量に用意することが困難な源信号の推定について考える．ラベルデータが大量にある場合は，観測信号をそのまま大規模なニューラルネットワークに入力することで，ニューラルネットワークの内部で適切な

音響特徴量が設計され，高精度に源信号を推定することができる．しかしラベルデータが少ない場合に，従来研究で用いられるような大規模なニューラルネットワークを用いると，学習データへの過適合が生じ源信号の推定精度が低下する．これを回避するためには，事前に音響特徴量を設計してネットワークのサイズを小さくすることで，ニューラルネットワークの自由度を下げる必要がある．事前に音響特徴量を設計した場合，その音響特徴量が回帰関数にとって不適切であると，源信号の推定精度が低下することが知られている．しかし回帰の精度を最大化するために入力すべき適切な音響特徴量の性質については明らかではなく，音響特徴量は技術者の経験に基づいて設計されている．3 章では，音源強調のためのニューラルネットワークの目的関数を最尤推定ととらえることで，回帰の精度を最大化するために入力すべき音響特徴量の性質が相互情報量の最大化で記述できることを示し，ニューラルネットワークに入力する音響特徴量を相互情報量の最大化に基づ選択することで音源強調の性能が向上することを示す．

4 章ではラベルデータが一意に定まらない源信号の推定について考える．高品質な音声通話のために，主観品質を最大化するように源信号を推定したい．しかし，人間が知覚する音質の劣化の大きさと源信号の推定結果の二乗誤差の大きさは必ずしも比例しないため，二乗誤差最小化などでニューラルネットワークを学習して源信号を推定しても主観品質を最大化する源信号を推定することはできない．聴覚フィルタなどを利用し，主観評価値と相関の高い音質評価値 [50, 51, 52] を計算することはできるが，評価値からそれを最大化するラベル（時間周波数マスク）は一意に定めることができない．ゆえにニューラルネットワークを教師あり学習することができなかった．4 章では，強化学習の枠組みを音源強調に応用することでこの問題を解決することを試みる．ラベルデータを明示的に用意するのではなく，音質評価値を報酬関数として用い，報酬の期待値を最大化するように時間周波数マスクの条件付き分布パラメータを出力するニューラルネットワークを学習する目的関数を提案する．

2.6 深層学習に基づく異常音検知

2.2 節では，外れ値検出に基づく異常音検知を概説し，検知精度を向上させるためには，観測信号を音響特徴量へ写像する関数を適切に設計する必要があることを述べた．2.3 節と 2.4 節では，深層学習を用いることで，柔軟な非線形変換関数を学習することができることを述べた．これらの研究の流れに基づく新たな方向性として，深層学習を用いて異常音検知のための音響特徴量を抽出する研究が行われてる．多くの場合音響特徴量は実数変数であるため，ニューラルネットワークは回帰関数として用いられる（図 2.14）．本節では，深層学習を用いた異常音検知のための音響特徴量抽出について説明する．

音源の状態を推定するための音響特徴量の抽出について研究は，観測信号の中に含ま

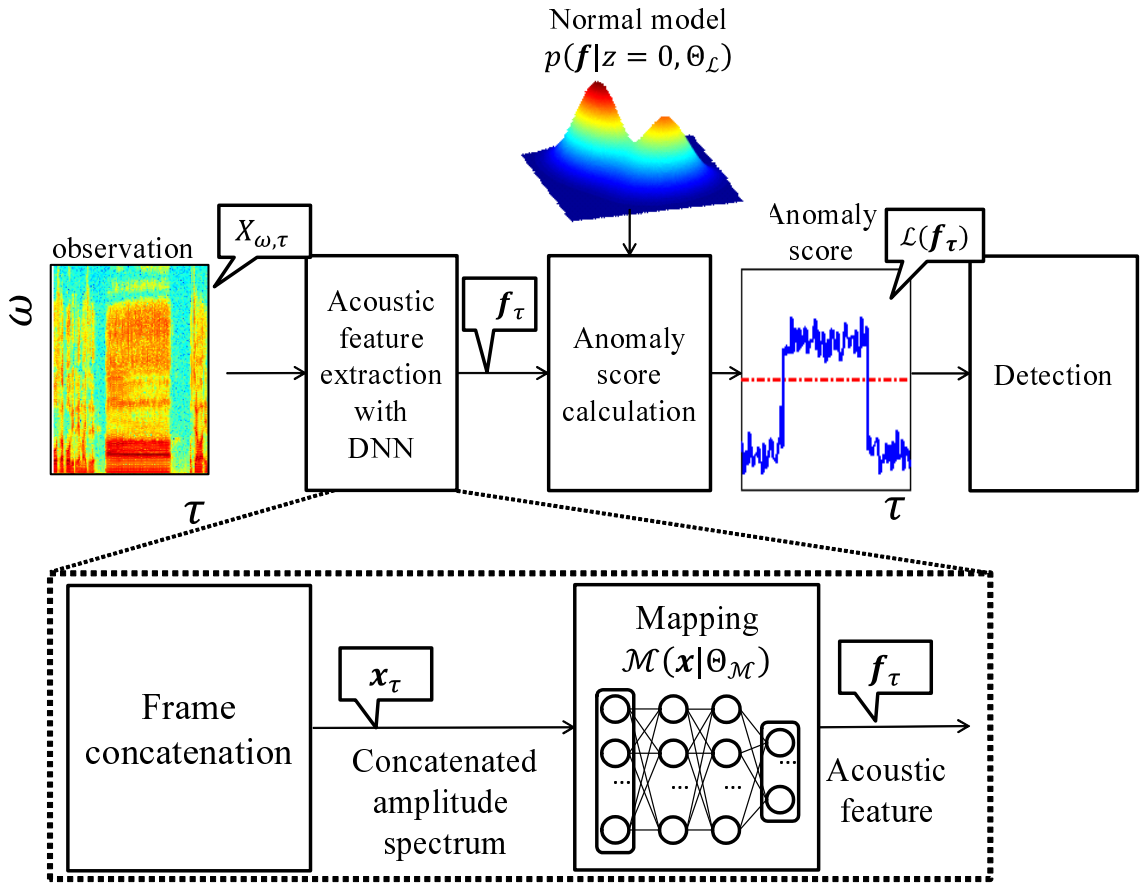


図 2.14: 深層学習を用いた外れ値検知に基づく異常音検知.

れる音源の種類を推定する音響イベント検出の分野で進められてきた [109, 110, 111]．これらの研究の中で，音源の状態を推定するためには観測信号の時間と周波数の両方の情報（時間周波数構造）を，音響特徴量として抽出する必要があることが明らかになっている．こういった音響特徴量を抽出するための手法として，時間周波数分解能の異なる複数のスペクトログラムを特徴量とする方法 [112] や，ガボールフィルタを用いた特徴量抽出法 [111] などが提案されている．近年では，より複雑な音響特徴量抽出関数を設計するための方法として，深層学習に基づく手法が研究されている [80, 113, 112]．

深層学習に基づく異常音検知のための音響特徴量抽出関数設計では，オートエンコーダや RBM などの自己符号化に基づくニューラルネットワークが広く利用される．この方法では，オートエンコーダの中間層の出力を音響特徴量とみなして異常音を検知する（図 2.15）．これは，異常音データやそのラベルデータがないため，交差エントロピー関数などを利用した，教師あり識別アプローチでニューラルネットワークを学習できないためである．そこで，音響特徴量から観測信号へ復元可能な音響特徴量であれば，少なくとも観測信号自身の時間周波数構造を保持した音響特徴量が抽出されるはずであるという考え方に基づき音響特徴量を抽出している．

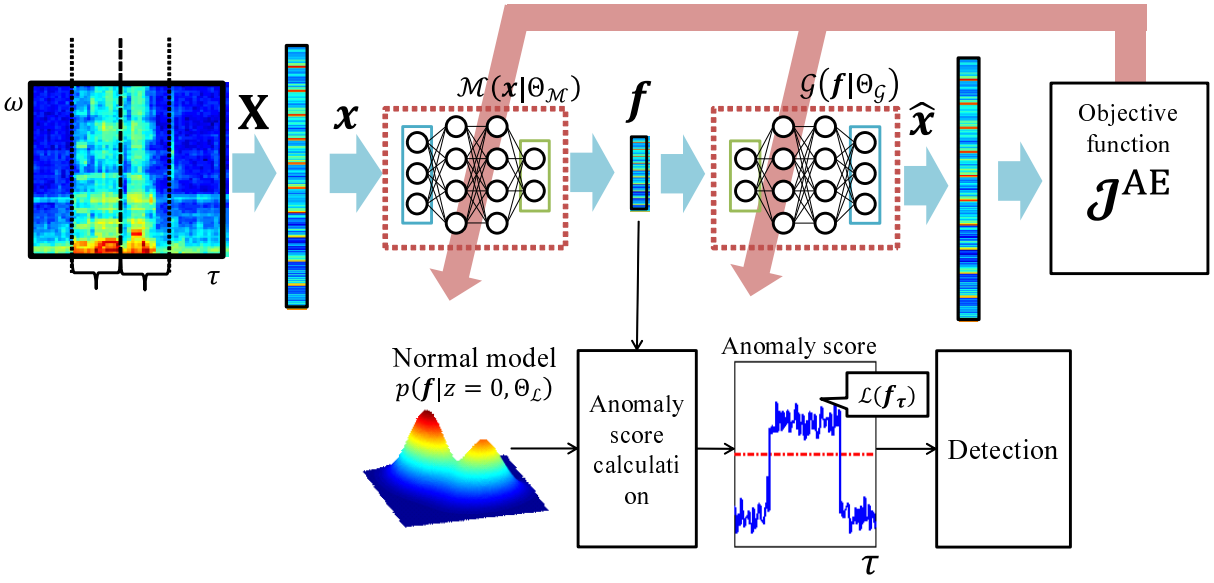


図 2.15: オートエンコーダなどの自己符号化に基づく音響特徴量を利用した異常音検知. ニューラルネットワークは再構成二乗誤差最小化基準などの目的関数で学習される

オートエンコーダは，ニューラルネットワークの入力と出力が一致するように 2 つのニューラルネトワークを学習する方法である．従来の異常音検知では，音響特徴量から観測信号へ復元可能な音響特徴量を観測信号から抽出することを目指しているため，学習に必要なデータは観測信号 $X_{\omega,\tau}$ である．そして DNN への入力ベクトル $\boldsymbol{x}_\tau$ には，観測信号自身の時間周波数構造を利用したいため，観測信号 $X_{\omega,\tau}$ の時間周波数要素を並べたベクトルとする．

$$\boldsymbol{x}_\tau = (\boldsymbol{X}_{\tau-P}, ..., \boldsymbol{X}_\tau, ..., \boldsymbol{X}_{\tau+P})^\top \tag{2.90}$$

$$\boldsymbol{X}_\tau = (\ln|X_{1,\tau}|, ..., \ln|X_{\Omega,\tau}|) \tag{2.91}$$

最後に観測信号 $\hat{\boldsymbol{x}}_\tau$ から音響特徴量 $\boldsymbol{f}_\tau$ を抽出するニューラルネットワーク $\mathcal{M}$ と，音響特徴量 $\boldsymbol{f}_\tau$ から観測信号 $\hat{\boldsymbol{x}}_\tau$ へ復元するニューラルネットワーク $\mathcal{G}$ を用いて，再構成二乗誤差を最小化するように学習する．$\mathcal{M}$ と $\mathcal{G}$ を DNN で実装すると以下のようになる．

$$\boldsymbol{f}_\tau \leftarrow \mathcal{M}(\boldsymbol{x}_\tau|\Theta_{\mathcal{M}}) = \boldsymbol{u}_\tau^{(L)} \tag{2.92}$$

$$\boldsymbol{z}_\tau^{(l)} = \sigma_\theta\left\{\boldsymbol{u}_\tau^{(l)}\right\} \tag{2.93}$$

$$\boldsymbol{u}_\tau^{(l)} = \mathbf{W}^{(l)}\boldsymbol{z}_\tau^{(l-1)} + \mathbf{b}^{(l)} \tag{2.94}$$

$$\hat{\boldsymbol{x}}_\tau \leftarrow \mathcal{G}(\boldsymbol{f}_\tau|\Theta_{\mathcal{G}}) = \boldsymbol{u}_\tau^{(L,\mathcal{G})} \tag{2.95}$$

$$\boldsymbol{z}_\tau^{(l,\mathcal{G})} = \sigma_{\theta,\mathcal{G}}\left\{\boldsymbol{u}_\tau^{(l,\mathcal{G})}\right\} \tag{2.96}$$

$$\boldsymbol{u}_\tau^{(l,\mathcal{G})} = \mathbf{W}^{(l,\mathcal{G})}\boldsymbol{z}_\tau^{(l-1,\mathcal{G})} + \mathbf{b}^{(l,\mathcal{G})} \tag{2.97}$$

ここで $\Theta_{\mathcal{M}} = \{\mathbf{W}^{(l)}, \mathbf{b}^{(l)} | l = 2, ..., L\}$，$\Theta_{\mathcal{G}} = \{\mathbf{W}^{(l,\mathcal{G})}, \mathbf{b}^{(l,\mathcal{G})} | l = 2, ..., L\}$，であり，$\sigma_\theta$ と $\sigma_{\theta,\mathcal{G}}$ は活性化関数である．また $\boldsymbol{z}_\tau^{(1)} = \boldsymbol{x}_\tau$，$\boldsymbol{z}_\tau^{(1,\mathcal{G})} = \boldsymbol{f}_\tau$ である．そして目的関数は以下のようになる．

$$\Theta_{\mathcal{M}}, \Theta_{\mathcal{G}} \leftarrow \arg\max_{\Theta_{\mathcal{M}}, \Theta_{\mathcal{G}}} \mathcal{J}^{\mathrm{AE}} \left(\Theta_{\mathcal{M}}, \Theta_{\mathcal{G}} \right) \tag{2.98}$$

$$\mathcal{J}^{\mathrm{AE}} \left(\Theta_{\mathcal{M}}, \Theta_{\mathcal{G}} \right) = -\frac{1}{T} \sum_{\tau=1}^{T} \left| \boldsymbol{x}_\tau - \mathcal{G} \left(\mathcal{M} \left(\boldsymbol{x}_\tau | \Theta_{\mathcal{M}} \right) | \Theta_{\mathcal{G}} \right) \right|^2 \tag{2.99}$$

本節ではオートエンコーダによる再構成二乗誤差最小化を目的関数としたニューラルネットワークの学習法を紹介したが，この他にも，RBM のように観測信号の尤度最大化を目的関数とするものや，2.4 節で紹介した変分オートエンコーダのように変分下界最大化 $\mathcal{J}^{\mathrm{VAE}}$ を目的関数とするものもある．

2.6.1 深層学習に基づく異常音検知の課題

本節では，深層学習に基づく異常音検知のための音響特徴量設計について説明してきた．異常音検知では異常音のラベルデータが存在しないため，その学習には正常音と異常音の識別率最大化による教師あり学習ではなく，外れ値検出に基づく手法が採用されることを説明した．外れ値検出に基づく異常音検知では，観測信号の時間周波数構造を表現する音響特徴量を抽出する必要があることを説明し [109, 110, 111]，深層学習を利用して音響特徴量を抽出する方法を説明した．これらの研究では，オートエンコーダや RBM などの自己符号化に基づく目的関数を利用して，ニューラルネットワークを学習する．これらは，音響特徴量から観測信号へ復元可能な音響特徴量であれば，少なくとも観測信号自身の時間周波数構造を保持した音響特徴量が抽出されるはずである，という考え方に基づいている．

しかし図 2.15 からわかるように，従来法では，ニューラルネットワークの学習に異常音検知の結果がフィードバックされていない．そのため，従来法でニューラルネットワークを学習しても，異常音検知の精度は最大化されない．異常音検知の精度を最大化するためには，外れ値検出に基づく異常音検知に適した目的関数で音響特徴量抽出関数であるニューラルネットワークを学習するべきである．5 章ではこの問題を解決するために，外れ値検出に基づく異常音検知のための目的関数を提案する．外れ値検出に基づく異常音検知を仮説検定とみなし，仮説検定の最適化基準であるネイマン・ピアソンの定理から，ニューラルネットワークを学習するための目的関数である “ネイマン・ピアソン指標” を導出する．

第 3 章

相互情報量最大化に基づく音響特徴量選択のための目的関数

本章では，スポーツ競技音などのラベルデータを大量に用意することが困難な源信号の推定について考える．ニューラルネットワークは，図 2.1 に示したように，音響特徴量を入力とし，時間周波数マスクを計算するために必要なパラメータを出力する回帰関数として用いられる．本章では小規模なニューラルネットワークを効率的に学習するために，ニューラルネットワークの入力に用いる音響特徴量を選択するための目的関数を提案する．そのため，ニューラルネットワーク自体は，従来法と同様に MMSE に基づく目的関数で学習する．以降では，源信号の二乗誤差に基づく推定精度を最大化するために入力すべき音響特徴量の性質が相互情報量の最大化で記述できることを示し，源信号の推定精度を最大化するための音響特徴選択の目的関数を提案する．

3.1 観測信号の定式化

高臨場メディア向けの音響再生技術として，22.2 マルチチャネル音響 [114, 115]，Dolby Atmos [116]，SAOC (spatial audio object coding) [117]，MPEG-H [118]，MPEG-4 Audio Lossless Coding (ALS) [119] などの “オブジェクトベース音響再生技術” が開発されてきた．これらの技術では，ユーザーの視点や位置に合わせて音源を定位させることにより，自由視点映像 [120, 121] のようにまるで空間に潜り込んだような音響体験を実現している．これらの技術では，定位させたい音源が事前にクリアに個別に収音できることを前提としているため，その利用用途は映画製作などに限られていた．この技術をスポーツ観戦などの実環境メディアに適用するためには，観測音から，所望の源信号だけを選択的かつクリアに収音する技術が必要である．そこで本節では，大歓声に包まれたスポーツフィールドで，サッカーボールのキック音などの特定の競技音だけを選択的に収音する，“オブジェクトベース収音技術” の実現を目指す．

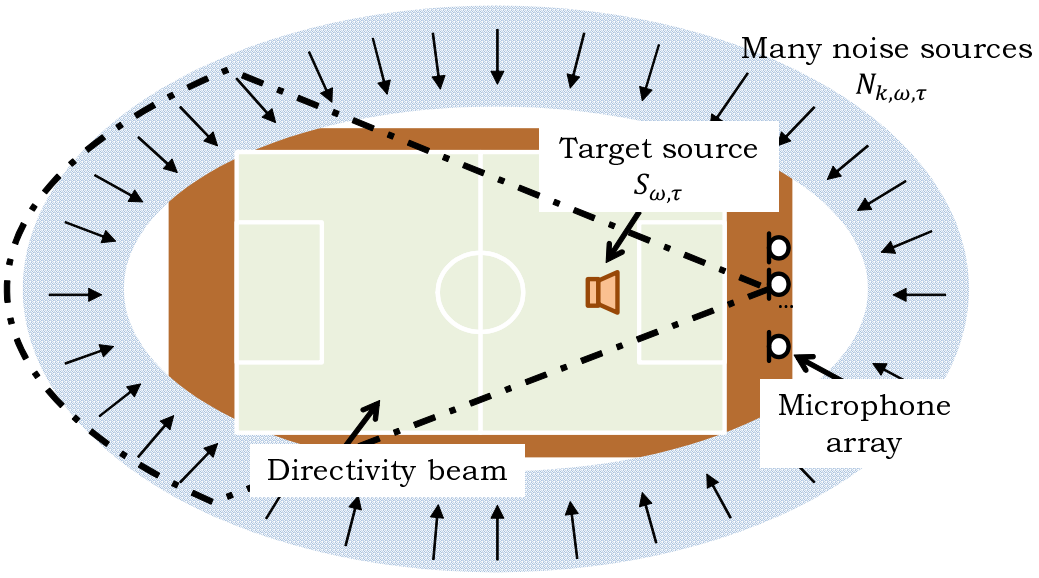


図 3.1: スポーツフィールドでの観測信号のモデル化.

まず観測信号を，スポーツフィールドでの収音を例にとりモデル化する．図 3.1 に示すように，スポーツフィールドでは，キック音などの所望の源信号 $S_{\omega,\tau} \in \mathbb{C}^{\Omega \times T}$ を取り囲むように無数の歓声雑音 $N_{k,\omega,\tau} \in \mathbb{C}^{K \times \Omega \times T}$ が存在している．ここで K は雑音源数を表す．全方向から到来する雑音の影響を低減するために，スポーツフィールドでの収音では，指向性マイクロホンやマイクロホンアレイを用いることが一般的である [4, 5, 6, 122, 123]．今，M 本の指向性マイクロホンで構成されるマイクロホンアレイでの観測信号 $X_{m,\omega,\tau} \in \mathbb{C}^{\Omega \times T}$ を以下のように記述する．

$$X_{m,\omega,\tau} = H_{m,\theta_S,\omega} V_{m,0,\omega} S_{\omega,\tau} + \sum_{k=1}^{K} H_{m,\theta_k,\omega} V_{m,k,\omega} N_{k,\omega,\tau} \tag{3.1}$$

ここで θ_S と θ_k はマイクロホンアレイから見た源信号と k 番目の雑音の方向，$V_{m,0,\omega}$ と $V_{m,k,\omega}$ は源信号と k 番目の雑音から m 番目の指向性マイクロホンまでの伝達関数，$H_{m,\theta,\omega}$ は m 番目の指向性マイクロホンの角度 θ への指向特性を表す．図 3.1 で示したように，スポーツフィールド上では全方向から雑音が到来するため，式 (3.1) のように指向性マイクロホンを用いた音源強調だけでは，源信号をクリアに収音できない．そこで観測信号から所望の源音源だけをクリアに強調するために，時間周波数マスクを用いた音源強調を行う．

$$Y_{\omega,\tau} = G^{\mathrm{WF}}_{\omega,\tau} X_{m_1,\omega,\tau} \tag{3.2}$$

ここで m_1 は，M 本のマイクロホンの中で，源信号の到来方向 $V^{(S)}_{m,\omega}$ に最も強い指向性を持つマイクロホンのインデックスである．また $G^{\mathrm{WF}}_{\omega,\tau}$ は以下の式で求まるウィナーフィル

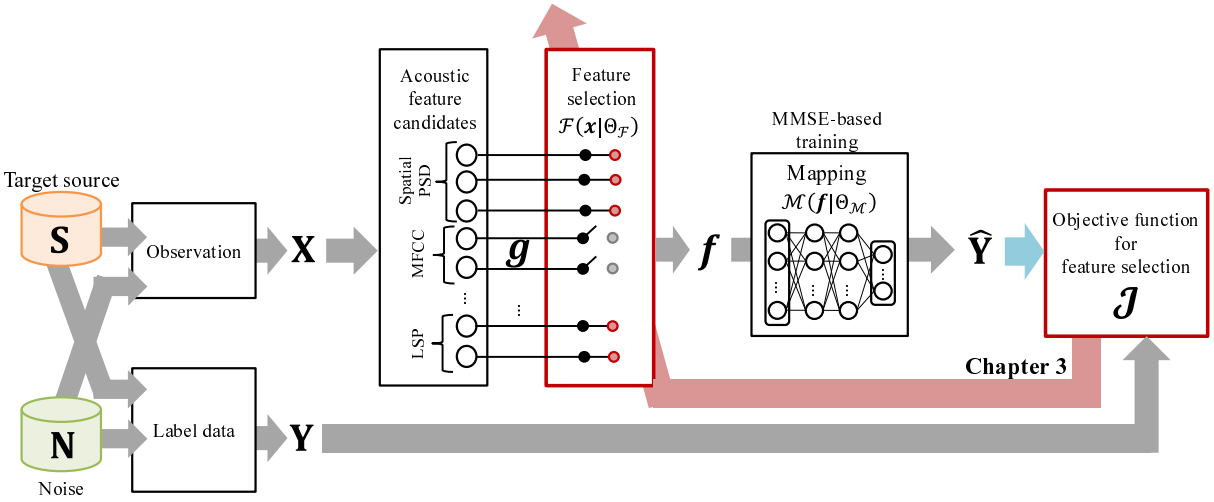


図 3.2: 本章の研究範囲．ニューラルネットワークに入力するべき適切な音響特徴量を，大量の音響特徴量候補から選択するための目的関数を提案する．

タである．

$$G_{\omega,\tau}^{\mathrm{WF}} = \frac{|H_{m_1,\theta_S,\omega} V_{m_1,0,\omega} S_{\omega,\tau}|^2}{|H_{m_1,\theta_S,\omega} V_{m_1,0,\omega} S_{\omega,\tau}|^2 + |\sum_{k=1}^{K} H_{m_1,\theta_k,\omega} V_{m_1,k,\omega} N_{k,\omega,\tau}|^2} = \frac{\exp(\xi_{\omega,k})}{1+\exp(\xi_{\omega,k})} \tag{3.3}$$

ここで $\xi_{\omega,k}$ は源信号と雑音の事前 SNR であり，以下のように計算できる．

$$\xi_{\omega,\tau} = \ln\left(\frac{|H_{m_1,\theta_S,\omega} V_{m_1,0,\omega} S_{\omega,\tau}|^2}{|\sum_{k=1}^{K} H_{m_1,\theta_k,\omega} V_{m_1,k,\omega} N_{k,\omega,\tau}|^2}\right) \tag{3.4}$$

ただしこの事前 SNR の定義は，事前 SNR に対数をとっている点で 2 章での事前 SNR の定義とは異なる．この理由は，スポーツフィールドでは雑音の音量が大きく，事前 SNR の学習データの大半の時間周波数成分が 0 に近い値をとるため，誤差計算の安定のために事前 SNR に対数をとっている．本節では，式 (3.4) で定義した事前 SNR[70, 33] $\boldsymbol{\xi}_\tau = (\xi_{1,\tau}, ..., \xi_{\Omega,\tau})^{\mathsf{T}}$ を観測信号から推定し，源音源をクリアに個別に収音する．

3.1.1 音響特徴量の選択とニューラルネットワークを用いた事前 SNR の推定

事前 SNR を推定する手段として，2 章で説明した全結合多層ニューラルネットワーク（DNN）を用いる．ラベルデータが大量にある場合は，式 (2.36) で示したように，観測信号をそのまま大規模な DNN に入力することで DNN の内部で適切な音響特徴量が抽出され，高精度に源信号を推定することができる．しかし，スポーツの競技音は無響室などの理想的な環境で収録することが難しいため，ラベルデータ（事前 SNR）を大量に収集することが困難であり，大規模な DNN を用いると学習データへの過適合が生じ源信

号の推定精度が低下する．そこで本節では過適合を回避するために，観測信号から事前に音響特徴量 $\boldsymbol{f} \in \mathbb{R}^{\mathsf{D}}$ を抽出し，小規模な DNN を用いて事前 SNR を推定する．

事前に音響特徴量を抽出して DNN で事前 SNR を推定する場合，不適切な音響特徴量を入力すると出力の推定精度が低下することが知られている．DNN に入力できる音響特徴量の候補には，メル周波数ケプストラム係数（MFCC: mel-frequency cepstrum coefficient）や線スペクトル対（LSP: line spectral pairs）など様々なものがある．これまでの研究により，音声を強調するための音響特徴量には，MFCC などのスペクトルの概形情報を選択することが有効であることがわかっている．これは音声が，音素や声色などの音色を変化させて情報を伝達するためと考えられる．しかし，サッカーボールのキック音や野球のバッティング音などのスポーツの競技音を強調するために有効な特徴量が何であるかは明らかではない．そのため，スポーツの競技音などの源信号を強調する場合には，技術者の経験や試行錯誤に基づいて音響特徴量を選択する必要があった．そこで本章では，所望の種類の源音源をクリアに強調するために DNN に入力すべき音響特徴量の性質を導出し，音響特徴量選択関数を最適化するための目的関数を提案する（図 3.2）．

本章の取り組みを定式化する．まず，MFCC や LSP などの $\boldsymbol{x}_\tau$ から抽出した音響特徴量候補から音響特徴量を選択する関数を $\mathcal{F}$，そのパラメータを $\Theta_{\mathcal{F}}$ とし，観測信号 $\boldsymbol{x}_\tau$ から音響特徴量 $\boldsymbol{f}_\tau$ を以下で抽出する．

$$\boldsymbol{f}_\tau \leftarrow \mathcal{F}\left(\boldsymbol{x}_\tau | \Theta_{\mathcal{F}}\right) \tag{3.5}$$

そして事前 SNR を以下の DNN で推定する．

$$\hat{\boldsymbol{\xi}}_\tau \leftarrow \mathcal{M}\left(\boldsymbol{f}_\tau | \Theta_{\mathcal{M}}\right) = \boldsymbol{u}_\tau^{(L)} \tag{3.6}$$

$$\boldsymbol{z}_\tau^{(l)} = \sigma_\theta\left\{\boldsymbol{u}_\tau^{(l)}\right\} \tag{3.7}$$

$$\boldsymbol{u}_\tau^{(l)} = \mathbf{W}^{(l)}\boldsymbol{z}_\tau^{(l-1)} + \mathbf{b}^{(l)} \tag{3.8}$$

ここで $\boldsymbol{z}_\tau^{(1)} = \boldsymbol{f}_\tau$ であり，L，$\mathbf{W}^{(l)}$，$\mathbf{b}^{(l)}$，$\Theta_{\mathcal{M}}$，σ_θ の定義は 1 章で説明したものと同じである．本章では，音響特徴量選択関数のパラメータを最適化するための目的関数を設計する．

$$\Theta_{\mathcal{F}} \leftarrow \arg\max_{\Theta_{\mathcal{F}}} \mathcal{J} \tag{3.9}$$

3.2 相互情報量最大化に基づく音響特徴量選択

3.2.1 二乗誤差を最小化するの音響特徴量の性質

この節では，事前 SNR を高精度に推定するために，音響特徴量の満たすべき性質について議論する．MMSE 推定は，所望の出力と DNN の出力の二乗誤差の期待値を最小化

しようとする推定方法である．期待値の計算には，事前 SNR の推定誤差の確率分布が既知である必要があるが，多くの場合その分布は未知である．そのためこの確率分布を，大量の学習データのヒストグラムで置き換えることで，期待値を学習データの平均値に置き換えている．ゆえにラベルデータが少ない場合，学習データだけから事前 SNR の推定誤差のヒストグラムを構成すると，確率分布の近似性能が低下するため，推定精度の低下や過適合が生じてしまう．そこで本節では，事前 SNR の推定誤差について確率分布を明示的に仮定し，その分布の性質から確率論的に目的関数を設計することでこの問題の解決を目指す．

まず，事前 SNR の推定誤差を以下のように定義し，その性質について議論をする．

$$e_{\omega,\tau} = \xi_{\omega,\tau} - \mathcal{M}(\mathcal{F}(\boldsymbol{x}_\tau|\Theta_\mathcal{F})|\Theta_\mathcal{M})_\omega. \tag{3.10}$$

$e_{\omega,\tau}$ の従う分布を調査するために，3.5.1 節に記述した予備実験を行った．予備実験の結果，$e_{\omega,\tau}$ は近似的に平均値に 0 を持つガウス分布に従うとみなして議論を進める．2.5 節で議論した通り，$e_{\omega,\tau}$ が独立に平均値が 0 となるガウス分布に従うとき，MMSE 推定は最尤推定と等価であることが知られている [124]．ここで MMSE に基づく目的関数における学習データの算術平均を期待値に戻すと，事前 SNR の推定誤差 $e_{\omega,\tau}$ を最小化する音響特徴量は，以下の目的関数 $\mathcal{J}$ を最大化することになる．

$$\mathcal{J} = \iint p(\mathcal{F}(\boldsymbol{x}|\Theta_\mathcal{F}), \boldsymbol{\xi}) \ln \mathcal{N}\left(\boldsymbol{\xi}|\mathcal{M}(\mathcal{F}(\boldsymbol{x}|\Theta_\mathcal{F})|\Theta_\mathcal{M}), \sigma^2\boldsymbol{I}_\Omega\right) d\boldsymbol{x}d\boldsymbol{\xi} \tag{3.11}$$

ここで $\boldsymbol{I}_\Omega$ は Ω 次元単位行列であり，σ^2 はガウス分布の分散パラメータである．ここで $\ln \mathcal{N}\left(\boldsymbol{\xi}|\mathcal{M}(\mathcal{F}(\boldsymbol{x}|\Theta_\mathcal{F})|\Theta_\mathcal{M}), \sigma^2\boldsymbol{I}_\Omega\right)$ が $\mathcal{F}(\boldsymbol{x}|\Theta_\mathcal{F})$ を得た下での $\boldsymbol{\xi}$ の条件付き分布であることに注意すれば，式 (3.11) は以下のように変形できる．

$$\mathcal{J} = \iint p(\mathcal{F}(\boldsymbol{x}|\Theta_\mathcal{F}), \boldsymbol{\xi}) \ln p(\boldsymbol{\xi}|\mathcal{F}(\boldsymbol{x}|\Theta_\mathcal{F})) d\boldsymbol{x}d\boldsymbol{\xi} \tag{3.12}$$

$$\propto \iint p(\mathcal{F}(\boldsymbol{x}|\Theta_\mathcal{F}), \boldsymbol{\xi}) \ln \frac{p(\mathcal{F}(\boldsymbol{x}|\Theta_\mathcal{F}), \boldsymbol{\xi})}{p(\mathcal{F}(\boldsymbol{x}|\Theta_\mathcal{F}))p(\boldsymbol{\xi})} d\boldsymbol{x}d\boldsymbol{\xi}. \tag{3.13}$$

ここで定数である事前 SNR のエントロピー $-\int p(\boldsymbol{\xi}) \ln p(\boldsymbol{\xi}) d\boldsymbol{\xi}$ を加えることで，式 (3.12) から式 (3.13) へ変形した．

式 (3.13) は，音響特徴量と事前 SNR の相互情報量として知られている値である．相互情報量は，音響特徴量の中に事前 SNR の情報が，潜在的にどれだけ含まれているかを示す値である [11, 59, 60, 125]．以上の議論より，事前 SNR の推定誤差が平均 0 のガウス分布に独立に従うと仮定したとき，MMSE に基づく目的関数は相互情報量の最大化と等価になる．ゆえに事前 SNR の推定精度を最大化する音響特徴量を選択するためには，相互情報量を目的関数としてそのパラメータを最適化すればよい．

$$\Theta_\mathcal{F} \leftarrow \arg\max_{\Theta_\mathcal{F}} \iint p(\mathcal{F}(\boldsymbol{x}|\Theta_\mathcal{F}), \boldsymbol{\xi}) \ln \frac{p(\mathcal{F}(\boldsymbol{x}|\Theta_\mathcal{F}), \boldsymbol{\xi})}{p(\mathcal{F}(\boldsymbol{x}|\Theta_\mathcal{F}))p(\boldsymbol{\xi})} d\boldsymbol{x}d\boldsymbol{\xi} \tag{3.14}$$

3.2.2 相互情報量最大化に基づく特徴量選択法

本節では，相互情報量を最大化する $\mathcal{F}(\boldsymbol{x}|\Theta_{\mathcal{F}})$ の設計法として，事前に定義した $\mathsf{Q}(>\mathsf{D})$ 次元の特徴量候補から，D 次元の有益な特徴量を選択する特徴量選択を採用する [126, 127]．ここで $\boldsymbol{g} \in \mathbb{R}^{\mathsf{Q}}$ を，MFCC や LSP などの $\boldsymbol{x}$ から抽出した音響特徴量候補を並べたベクトル $\boldsymbol{g}_\tau = (g_{1,\tau}, ..., g_{\mathsf{Q},\tau})^\top$ とする．特徴量選択を採用すると，特徴量抽出関数 $\mathcal{F}(\boldsymbol{x}|\Theta_{\mathcal{F}})$ は以下のように記述することができる．

$$\mathcal{F}(\boldsymbol{x}|\Theta_{\mathcal{F}}) = \boldsymbol{A}\boldsymbol{g} \tag{3.15}$$

ここで $\boldsymbol{A}$ は，各行ベクトルが 1 つの非零要素を持つ特徴量選択行列である．例えば $\mathsf{Q} = 4, \mathsf{D} = 2$ であり，$\boldsymbol{A}$ が 1 番目と 3 番目の特徴量候補を選択するとき，式 (3.15) は以下のように具体化される．

$$\begin{bmatrix} a_1 g_1 \\ a_2 g_3 \end{bmatrix} = \begin{bmatrix} a_1 & 0 & 0 & 0 \\ 0 & 0 & a_2 & 0 \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \\ g_3 \\ g_4 \end{bmatrix} \tag{3.16}$$

すると，式 (3.14) と式 (3.15) を用いれば，特徴量抽出関数 $\mathcal{F}(\boldsymbol{x}|\Theta_{\mathcal{F}})$ の最適化問題は，相互情報量を最大化する特徴量選択行列 $\boldsymbol{A}$ の最適化問題へと書き換えることができる．

$$\boldsymbol{A} \leftarrow \underset{\boldsymbol{A}\in\mathcal{A}_{\mathsf{D},\mathsf{Q}}}{\arg\max} \iint p(\boldsymbol{\xi}, \boldsymbol{A}\boldsymbol{g}) \ln \frac{p(\boldsymbol{\xi}, \boldsymbol{A}\boldsymbol{g})}{p(\boldsymbol{\xi})p(\boldsymbol{A}\boldsymbol{g})} d\boldsymbol{\xi} d\boldsymbol{g}, \tag{3.17}$$

ここで $\mathcal{A}_{\mathsf{D},\mathsf{Q}}$ は $\mathsf{D} \times \mathsf{Q}$ の特徴量選択行列がとりうる行列の集合である．

式 (3.17) で $\boldsymbol{A}$ を最適化するためには，音響特徴量と事前 SNR の相互情報量を評価する必要がある．相互情報量の計算には，音響特徴量と事前 SNR の従う確率密度関数が既知である必要があるが，これらは未知であるため学習データから推定する必要がある．しかし，例えば入出力変数の結合分布および各周辺分布をガウス分布などの解析的に扱いやすい分布で近似すると，分布間の高次相関を正確に計量することができず[1]，相互情報量が正しく評価できない．この問題を解決するために福水らは，入出力変数の各分布を仮定せずに再生核ヒルベルト空間で入出力変数の高次相関を直接計量することで，相互情報量を計量して入力の次元削減を行う “カーネル次元圧縮 [128, 129]” を提案した．カーネル次元圧縮では，$\boldsymbol{A}$ の最適化のための目的関数は以下のように記述される．

$$\boldsymbol{A} \leftarrow \underset{\boldsymbol{A}\in\mathcal{A}_{\mathsf{D},\mathsf{Q}}}{\arg\max} -\mathrm{Tr}\left[\Phi_\xi \left(\Phi_g + \epsilon \boldsymbol{I}_K\right)^{-1}\right], \tag{3.18}$$

[1] ガウス分布は 2 次モーメントまでで分布形状が決定するため，3 次以上の入出力の相関構造を計量することができない．

ここで $\epsilon > 0$ は正則化パラメータ，Φ_ξ と Φ_g は，以下で計算される中心化グラム行列である．

$$\Phi_\xi = \boldsymbol{P}\Psi_\xi\boldsymbol{P}, \tag{3.19}$$

$$\Phi_g = \boldsymbol{P}\Psi_g\boldsymbol{P}, \tag{3.20}$$

ここで $\mathbf{1} = (1, 1, ..., 1)^\mathsf{T}$，$\boldsymbol{P} = \boldsymbol{I}_K - \frac{1}{K}\mathbf{1}\mathbf{1}^\mathsf{T}$，$\Psi_\xi$ と Ψ_g は $\boldsymbol{\xi}$ と $\boldsymbol{Ag}$ のグラム行列である．

$$\Psi_\xi = \begin{bmatrix} \psi_\xi(\boldsymbol{\xi}_1, \boldsymbol{\xi}_1) & \cdots & \psi_\xi(\boldsymbol{\xi}_1, \boldsymbol{\xi}_K) \\ \vdots & \ddots & \vdots \\ \psi_\xi(\boldsymbol{\xi}_K, \boldsymbol{\xi}_1) & \cdots & \psi_\xi(\boldsymbol{\xi}_K, \boldsymbol{\xi}_K) \end{bmatrix}, \tag{3.21}$$

$$\Psi_g = \begin{bmatrix} \psi_g(\boldsymbol{Ag}_1, \boldsymbol{Ag}_1) & \cdots & \psi_g(\boldsymbol{Ag}_1, \boldsymbol{Ag}_K) \\ \vdots & \ddots & \vdots \\ \psi_g(\boldsymbol{Ag}_K, \boldsymbol{Ag}_1) & \cdots & \psi_g(\boldsymbol{Ag}_K, \boldsymbol{Ag}_K) \end{bmatrix}, \tag{3.22}$$

ここで $\psi_\xi(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j)$ と $\psi_g(\boldsymbol{Ag}_i, \boldsymbol{Ag}_j)$ は $\boldsymbol{\xi}$ と $\boldsymbol{Ag}$ の任意の半正定値カーネル関数である．式 (3.18) の右辺は条件付き共分散作用素（NTCCO: the negative-trace of the conditional covariance operator）として知られている [128, 129]．NTCCO は入出力変数間の高次相関の強さを表す変数であり，相互情報量と等価な，入力変数の中に出力変数の情報が潜在的にどれだけ含まれているかを示す値である．

福水らは，式 (3.17) の実行に組み合わせ最適化を利用していたため，入力変数の次元が大きくなると，その計算が困難となる [128]．特に音響特徴量は複数の次元の組み合わせで音の性質を表現するものが多く，例えば $\mathsf{Q} = 256$，$\mathsf{D} = 48$，$a_q \in \{0, 1\}$ であれば，$\boldsymbol{A}$ の候補数は $|\mathcal{A}_{\mathsf{D},\mathsf{Q}}| \approx 2.9 \times 10^{52}$ となり，その探索は事実上不可能である．ゆえに音響特徴量選択のような Q が大きくなる場合には，この式 (3.17) の実行が不可能であった．

3.2.3 ガウスカーネルを用いた音響特徴量選択行列の数値的最適化

本節では，組み合わせ最適化の問題を解決し，数値的に音響特徴量選択行列を最適化するために，式 (3.18) とガウスカーネルの特性に着目する．提案法による音源強調全体の概要を図 3.3 に示す．まず，相互情報量最大化に基づき音響特徴量選択行列を最適化する（図 3.3(1-1)）．次いで，選択された音響特徴量と事前 SNR を用いて，DNN を MMSE 基準で学習する（図 3.3(1-2)）．そして新たな観測信号が得られたら，学習された音響特徴量選択行列と DNN を用いて時間周波数マスクを設計し，所望の源音源を強調する（図 3.3(2)）．

式 (3.18) のグラム行列で用いるカーネルとしてガウスカーネルを採用すると，$\psi_\xi(\boldsymbol{\xi}_k, \boldsymbol{\xi}_{k'})$

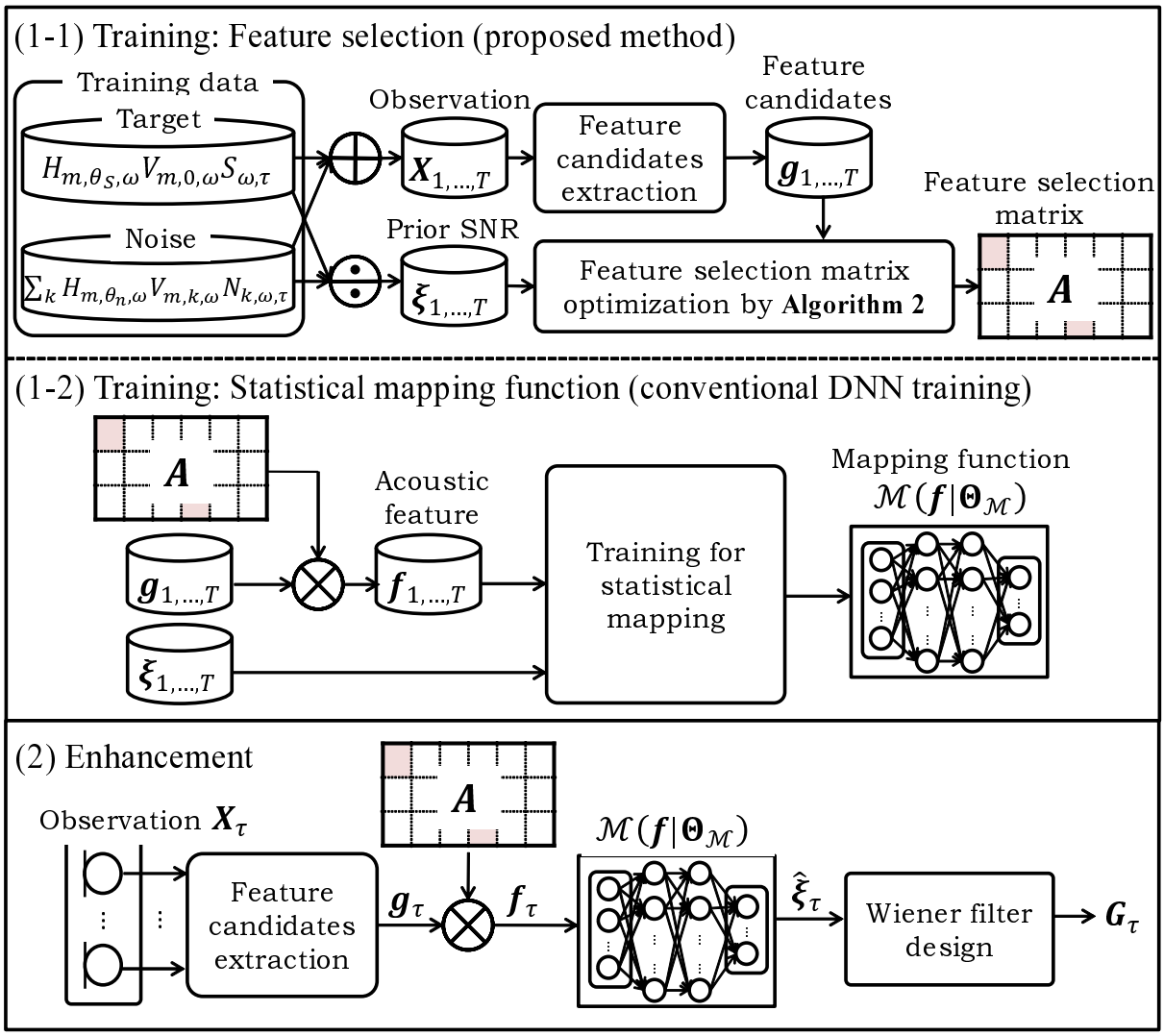


図 3.3: 提案法の実行手順.

と $\psi_g(\boldsymbol{A}\boldsymbol{g}_k, \boldsymbol{A}\boldsymbol{g}_{k'})$ は以下のように記述できる.

$$\psi_\xi(\boldsymbol{\xi}_k, \boldsymbol{\xi}_{k'}) = \exp\left\{-\frac{1}{2\varsigma_\xi^2}||\boldsymbol{\xi}_k - \boldsymbol{\xi}_{k'}||_2^2\right\} \tag{3.23}$$

$$\psi_g(\boldsymbol{A}\boldsymbol{g}_k, \boldsymbol{A}\boldsymbol{g}_{k'}) = \exp\left\{-\frac{1}{2\varsigma_g^2}||\boldsymbol{A}\boldsymbol{g}_k - \boldsymbol{A}\boldsymbol{g}_{k'}||_2^2\right\} \tag{3.24}$$

ただし ς_ξ^2 と ς_g^2 はガウスカーネルのパラメータである. ここで $\varsigma_g^2 = 1$ とすると, 式 (3.24) は

$$\psi_g(\boldsymbol{A}\boldsymbol{g}_k, \boldsymbol{A}\boldsymbol{g}_{k'}) = \exp\left\{-\frac{1}{2}(\boldsymbol{g}_k - \boldsymbol{g}_{k'})^\mathsf{T}\boldsymbol{A}^\mathsf{T}\boldsymbol{A}(\boldsymbol{g}_k - \boldsymbol{g}_{k'})\right\}, \tag{3.25}$$

と簡潔に記述することができる. ここで $\boldsymbol{A}^\mathsf{T}\boldsymbol{A} \in \mathbb{R}^{\mathsf{Q}\times\mathsf{Q}}$ は $\boldsymbol{A}$ の平方和積和行列 (SSCP matrix: sum of squares and cross-product matrix) である. 音響特徴量選択行列を構成する各行ベクトルは, 一つの非零要素のみをもつベクトルであるため, その平方和積和行列は対角行列となる. たとえば, $\mathsf{Q} = 4$ の中から, 1,3,4 番目の音響特徴量が選択され

るとき，$\boldsymbol{A}$ の平方和積和行列は以下のようになる．

$$\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A} = \begin{bmatrix} a_1 & & \\ 0 & \cdots & 0 \\ & a_3 & \\ & & a_4 \end{bmatrix} \begin{bmatrix} a_1 & 0 & \\ & \vdots & a_3 \\ & 0 & \\ & & a_4 \end{bmatrix} = \begin{bmatrix} a_1^2 & & & \\ & 0 & & \\ & & a_3^2 & \\ & & & a_4^2 \end{bmatrix}. \tag{3.26}$$

すると各対角要素 a_q^2 は，音響特徴量候補に対する重みと捉えることができるため，式 (3.25) は以下のように書き換えることができる．

$$\psi_g(\boldsymbol{A}\boldsymbol{g}_k, \boldsymbol{A}\boldsymbol{g}_{k'}) = \exp\left\{-\frac{1}{2}\sum_{q=1}^{\mathsf{Q}} a_q^2 (g_{q,k} - g_{q,k'})^2\right\} \tag{3.27}$$

式 (3.27) の構造に着目すると，式 (3.27) は $\boldsymbol{A}$ の平方和積和行列の対角要素を並べたベクトル $\boldsymbol{a} = (a_1, ..., a_{\mathsf{Q}})^{\mathsf{T}}$ で微分可能なことがわかる．また，音響特徴量選択行列の目的関数である式 (3.18) において，音響特徴量選択行列が出現するのはグラム行列中のカーネル関数のみであるため，式 (3.27) が $\boldsymbol{a}$ で微分可能ならば，目的関数である式 (3.18) も $\boldsymbol{a}$ で微分可能である．目的関数が $\boldsymbol{a}$ で微分可能ならば，勾配法などの非線形最適化法により $\boldsymbol{a}$ を最適化できるため，相互情報量を最大化する音響特徴量選択行列を数値的に求めることができる．

本章では $\boldsymbol{a}$ を求めるために勾配法を用いる．$\boldsymbol{a}$ は $\boldsymbol{A}$ の平方和積和行列の対角要素を並べたベクトルであるため，有効な音響特徴量に対応する次元だけが非零となるはスパースなベクトルとして求めたい．また，音響特徴量の中には，MFCC のように，複数の次元をまとめて一つの音源情報を表すものもある．ゆえに，こういった音響特徴量を選択するときは，全ての次元をまとめて非零にするような最適化をすべきである．これらの条件を満たすために，式 (3.18) にグループ化 L_1 正則化項を付与して $\boldsymbol{a}$ を最適化する [130]．まず，音響特徴量候補とその重みを，手動で U 個のグループに分割する．

$$\boldsymbol{g} = (\boldsymbol{g}_1, ..., \boldsymbol{g}_U) \tag{3.28}$$

$$\boldsymbol{a} = (\boldsymbol{a}_1, ..., \boldsymbol{a}_U) \tag{3.29}$$

次に，式 (3.18) を，グループ化 L_1 正則化項を付与した形に書き換える．

$$\boldsymbol{a} \leftarrow \arg\max_{\boldsymbol{a}} \left(\mathcal{J} - \lambda \sum_{u=1}^{U} |u| ||\boldsymbol{a}_u||_2\right) \tag{3.30}$$

$$\mathcal{J} = -\mathrm{Tr}\left[\Phi_\xi \left(\Phi_g + \epsilon \boldsymbol{I}\right)^{-1}\right] \tag{3.31}$$

ただし $\lambda > 0$ は正則化パラメータであり，$|u|$ は u 番目のグループの音響特徴量の次元数を表す．

式 (3.30) の最適化には，近接勾配法のうち ISTA(iterative shrinkage-thresholding algorithm) [131] と呼ばれるアルゴリズムを利用することができる．ISTA は勾配法による $\mathcal{J}$ の最大化と，閾値を用いた $\boldsymbol{a}$ の縮退を組み合わせたアルゴリズムである．まず勾配法による $\mathcal{J}$ の最大化のために，$\mathcal{J}$ の a_q による勾配を以下のように計算する．

$$\begin{aligned}\frac{\partial \mathcal{J}}{\partial a_q} &= -\mathrm{Tr}\left[\frac{\partial\left(\Phi_\xi\left(\Phi_g+\epsilon\boldsymbol{I}\right)^{-1}\right)}{\partial a_q}\right],\\ &= -\mathrm{Tr}\left[\Phi_\xi\frac{\partial\left(\Phi_g+\epsilon\boldsymbol{I}\right)^{-1}}{\partial a_q}\right],\\ &= \mathrm{Tr}\left\{\Phi_\xi\left(\Phi_g+\epsilon\boldsymbol{I}\right)^{-1}\frac{\partial\Phi_g}{\partial a_q}\left(\Phi_g+\epsilon\boldsymbol{I}\right)^{-1}\right\}, \qquad (3.32)\\ \frac{\partial\Phi_g}{\partial a_q} &= \begin{bmatrix}\phi'_{1,1,q} & \cdots & \phi'_{1,K,q}\\ \vdots & \ddots & \vdots\\ \phi'_{K,1,q} & \cdots & \phi'_{K,K,q}\end{bmatrix}. \qquad (3.33)\end{aligned}$$

ただし式 (3.32) の導出には，トレース，行列積，および逆行列の微分公式を用いた [132]．ここで $\phi'_{k,k',q}$ は

$$\begin{aligned}\phi'_{i,j,q} = &\frac{\partial\psi_g(\boldsymbol{A}\boldsymbol{g}_i,\boldsymbol{A}\boldsymbol{g}_j)}{\partial a_q} - \frac{1}{K}\sum_{i'=1}^{K}\frac{\partial\psi_g(\boldsymbol{A}\boldsymbol{g}_i,\boldsymbol{A}\boldsymbol{g}_{j'})}{\partial a_q}\\ &- \frac{1}{K}\sum_{j'=1}^{K}\frac{\partial\psi_g(\boldsymbol{A}\boldsymbol{g}_j,\boldsymbol{A}\boldsymbol{g}_{j'})}{\partial a_q} + \frac{1}{K^2}\sum_{i'=1}^{K}\sum_{j'=1}^{K}\frac{\partial\psi_g(\boldsymbol{A}\boldsymbol{g}_{i'},\boldsymbol{A}\boldsymbol{g}_{j'})}{\partial a_q},\end{aligned} \qquad (3.34)$$

のように計算可能であり，ガウスカーネルの微分は

$$\begin{aligned}\frac{\partial\psi_g(\boldsymbol{A}\boldsymbol{g}_k,\boldsymbol{A}\boldsymbol{g}_{k'})}{\partial a_q} &= \frac{\partial\exp\left\{-\frac{1}{2}(\boldsymbol{g}_k-\boldsymbol{g}_{k'})^\mathsf{T}\boldsymbol{A}^\mathsf{T}\boldsymbol{A}(\boldsymbol{g}_k-\boldsymbol{g}_{k'})\right\}}{\partial a_q},\\ &= -\frac{1}{2}\sum_{i=1}^{\mathsf{Q}}\frac{\partial\left(a_q^2(g_{i,k}-g_{i,k'})^2\right)}{\partial a_q}\exp\left\{-\frac{1}{2}\sum_{q=1}^{\mathsf{Q}}a_q^2(g_{q,k}-g_{q,k'})^2\right\},\\ &= -a_q(g_{q,k}-g_{q,k'})^2\psi_g(\boldsymbol{A}\boldsymbol{g}_k,\boldsymbol{A}\boldsymbol{g}_{k'}). \qquad (3.35)\end{aligned}$$

のように求めることができる．勾配法の収束を早めるために，提案法の最適化には，慣性項を付与した AdaDelta [133] 法を適用する．すると，勾配法による $\mathcal{J}$ の最大化は以

下のように実行できる．

$$r_q \leftarrow \gamma r_q + (1-\gamma)\left(\frac{\partial \mathcal{J}}{\partial a_q}\right)^2 \tag{3.36}$$

$$\tilde{r}_q \leftarrow \left(\frac{s_q^{1/2}+\epsilon}{r_q^{1/2}+\epsilon}\right)\frac{\partial \mathcal{J}}{\partial a_q} \tag{3.37}$$

$$s_q \leftarrow \gamma s_q + (1-\gamma)\tilde{r}_q^2 \tag{3.38}$$

$$\nu_q \leftarrow \eta \nu_q + (1-\eta)\tilde{r}_q \tag{3.39}$$

$$a_q \leftarrow a_q + \alpha \nu_q, \tag{3.40}$$

ただし $1 > \gamma > 0$ と $1 > \eta > 0$ は慣性項である．次に，式 (3.30) の閾値 λ を用いて $\boldsymbol{a}$ を縮退させる．

$$\boldsymbol{a}_u \leftarrow \begin{cases} 0 & (||\boldsymbol{a}_u||_2 < \lambda|u|) \\ (\boldsymbol{a}_u - \lambda)\frac{\boldsymbol{a}_u}{||\boldsymbol{a}_u||_2} & (\text{otherwise}) \end{cases}. \tag{3.41}$$

以上を収束条件を満たすまで繰り返すことで，$\boldsymbol{a}$ を数値的に最適化することができる．

ただし，ISTA は繰り返し最適化法であるため，式 (3.32) 中の行列 $(\Phi_g + \epsilon \boldsymbol{I})$ の逆行列を，繰り返し毎に求めなくてはならない．この行列のサイズは $\mathbb{R}^{K\times K}$ であり，K は学習データ中に含まれる全フレーム数を表す．ほとんどの場合，K は大きな値をとるため，逆行列の計算には膨大な計算量を要する．例えば学習データがサンプリング周波数 48 kHz で収録された 1 時間の音声データであり，それを窓長 30 ms，シフト長 15 ms の条件で分析した場合，$K \approx 2.4 \times 10^5$ となる．そこで計算量削減のために，本節では勾配法の実行には確率的勾配法を用いる．すなわち K フレームの学習データを，各バッチが K_z フレームを含む Z 個のミニバッチに分割し，ミニバッチごとに逆行列の計算を行う．以上をまとめた提案法の疑似コードを **Algorithm 1** にまとめた．また $\boldsymbol{a}$ から音響特徴量選択行列を求めるためには **Algorithm 2** を用いる．

3.3 評価実験

提案法の有効性を検証するために，スポーツフィールドを模擬した実験室で収集したデータを用いて，客観評価実験と主観評価実験を行った．

- 3.3.2 節では，提案法で選択した音響特徴量が従来法で選択した音響特徴量と比べて相互情報量を増加させるかを検証した．
- 3.3.3 節では，提案法が従来法の特徴量選択法と比べ音源強調性能を向上させることを客観的に示すために，信号対歪比 (SDR: signal-to-distortion ratio) [134] を用いた性能比較を行った．

Algorithm 1 ISTA による $\boldsymbol{a}$ の更新アルゴリズム

Input: Training dataset $\boldsymbol{\xi}_{1,...,K}$, $\boldsymbol{g}_{1,...,K}$, validation dataset $\boldsymbol{\xi}_{1,...,K_v}$, $\boldsymbol{g}_{1,...,K_v}$, parameters for update process $\alpha, \gamma, \eta, \lambda$ and parameters for validation process $\mathcal{J}_{\text{count-max}}$
Output: $\boldsymbol{a}$
Initialize $\boldsymbol{a}, \boldsymbol{r}, \boldsymbol{s}$ and $\boldsymbol{\nu}$
Set validation parameter: $\mathcal{J}_{\text{vmax}} \leftarrow -\infty$ and $\mathcal{J}_{\text{count}} \leftarrow 0$
while true do
  {% Update process}
  Separate the training dataset $\boldsymbol{\xi}_{1,...,K}$ and $\boldsymbol{g}_{1,...,K}$ into Z mini-batches
  for All mini-batch $z = 1, ..., \mathsf{Z}$ **do**
    Calculate Φ_s and Φ_g with z-th mini-batch with (3.19)–(3.22).
    Update $\boldsymbol{a}$ by (3.36)–(3.40) (gradient method)
    Update $\boldsymbol{a}$ by (3.41) (Soft thresholding)
  end for
  {% Validation process for early-stopping}
  Calculate $\mathcal{J}_{\text{V}}$ using $\boldsymbol{\xi}_{1,...,K_v}$, $\boldsymbol{g}_{1,...,K_v}$ by (3.31)
  if $\mathcal{J}_{\text{V}} < \mathcal{J}_{\text{vmax}}$ **then**
    $\alpha \leftarrow 0.5 \times \alpha$
    $\mathcal{J}_{\text{count}} \leftarrow \mathcal{J}_{\text{count}} + 1$
  else
    $\mathcal{J}_{\text{vmax}} \leftarrow \mathcal{J}_{\text{V}}$
  end if
  if $\mathcal{J}_{\text{count}} > \mathcal{J}_{\text{count-max}}$ **then**
    break
  end if
end while

Algorithm 2 $\boldsymbol{a}$ を用いた $\boldsymbol{A}$ の設計アルゴリズム

Input: $\boldsymbol{a}$
Output: $\boldsymbol{A}$
$\mathsf{Q} \leftarrow \text{length}(\boldsymbol{a})$
$\mathsf{D} \leftarrow$ Non-zero element count$(\boldsymbol{a} > 0)$
Initialize $\boldsymbol{A} \leftarrow \mathbf{0}_{\mathsf{D}\times\mathsf{Q}}$, $d = 1$
for $q = 1, ..., \mathsf{Q}$ **do**
  if $a_q > 0$ **then**
    $\boldsymbol{A}_{d,q} \leftarrow a_q$
    $d \leftarrow d + 1$
  end if
end for

- 3.3.4 節では，提案法が従来法の特徴量選択法と比べ音源強調性能を向上させることを主観的に示すために，比較平均オピニオン評点 (CMOS: comparison mean opinion score) を用いて，音質および明瞭度の聴取実験を行った．

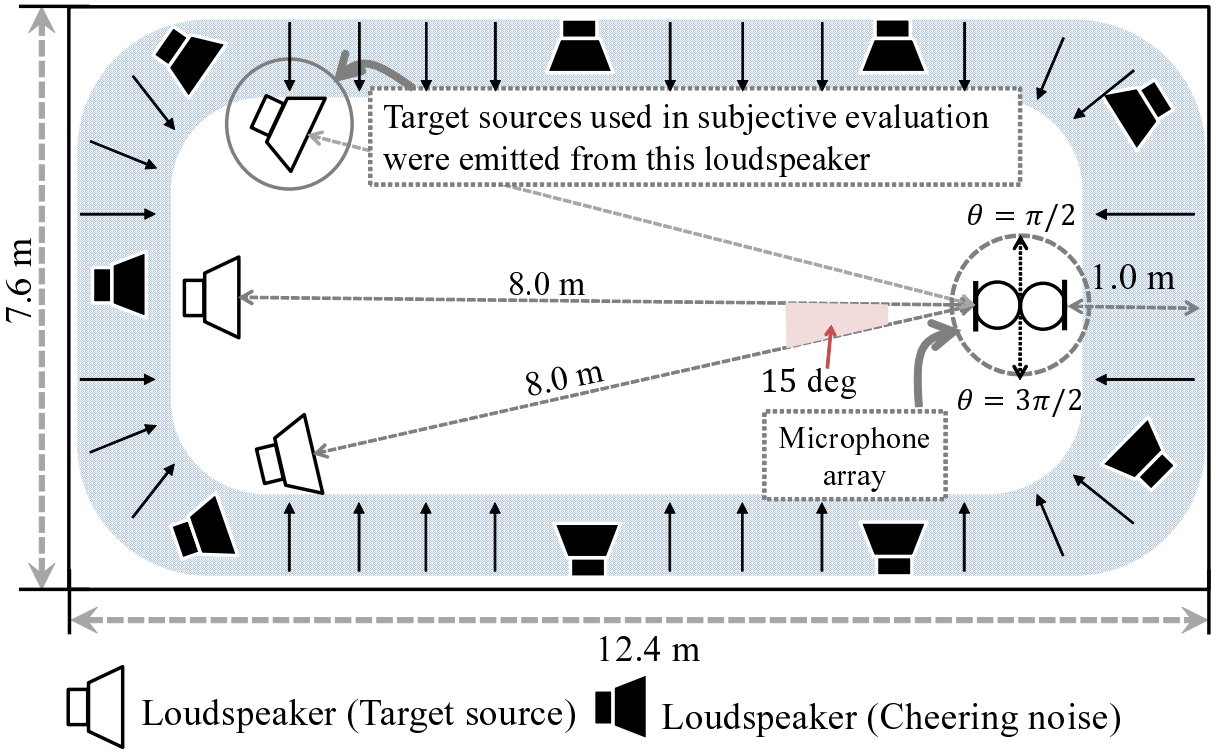
3.3.1 実験条件

歓声が全方向から到来するスポーツ競技音の強調性能を評価するために，目的音としてサッカーおよび野球の試合で発生する競技音を採用した．目的音は，サッカーボールのキック音 (FOOT) とゴールキーパーの叫び声 (SHOUT)，および野球ボールの打球音 (BASE) とした．目的音と歓声雑音を再生するスピーカは図 3.4 のように配置した．目的音を再生するスピーカは部屋の中心に配置し，それを取り囲むように複数配置したスピーカを用いて，実際のスポーツフィールドで収録した歓声雑音を再生した．マイクアレイは，目的音方向に鋭い指向性を持つショットガンマイクロホン ($m = 1$) と，雑音方向に高い指向性を持つ単一指向性マイクロホン ($m = 2$) の 2 本のマイクロホンで構成した．

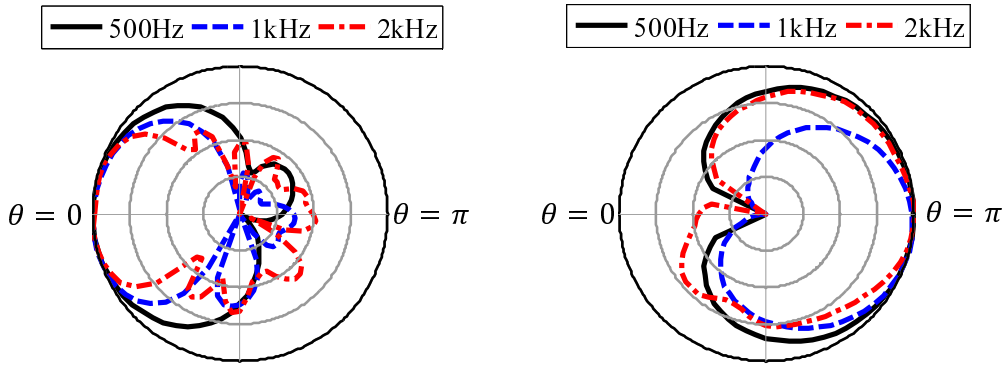
実験で比較した音響特徴量抽出法と事前 SNR の推定法を表 3.1 にまとめた．提案法では，選択した音響特徴量を事前 SNR に変換するために GMM 回帰 (3 : pGMM) と DNN 回帰 (6 : pDNN) を用いた．GMM 回帰を用いた理由は，本研究の本質が MMSE に基づく回帰関数に最適な音響特徴量を選択するための目的関数の設計であるため，手法の妥当性を示すためには DNN だけでなく，その他の射影関数でも精度を評価する必要があるためである．GMM 回帰の実行法については，3.5.2 節を参照されたい．従来法には，以下の 4 つの手法を用いた．

1 : CCA − GMM　音響特徴量候補を正準相関分析（CCA: Canonical Correlation Anal-

(a): Arrangement of microphone array and loudspeakers (top view)



(b): Polar pattern of directional microphones $H_{m,\theta,\omega}$



(c): Experimental set-up

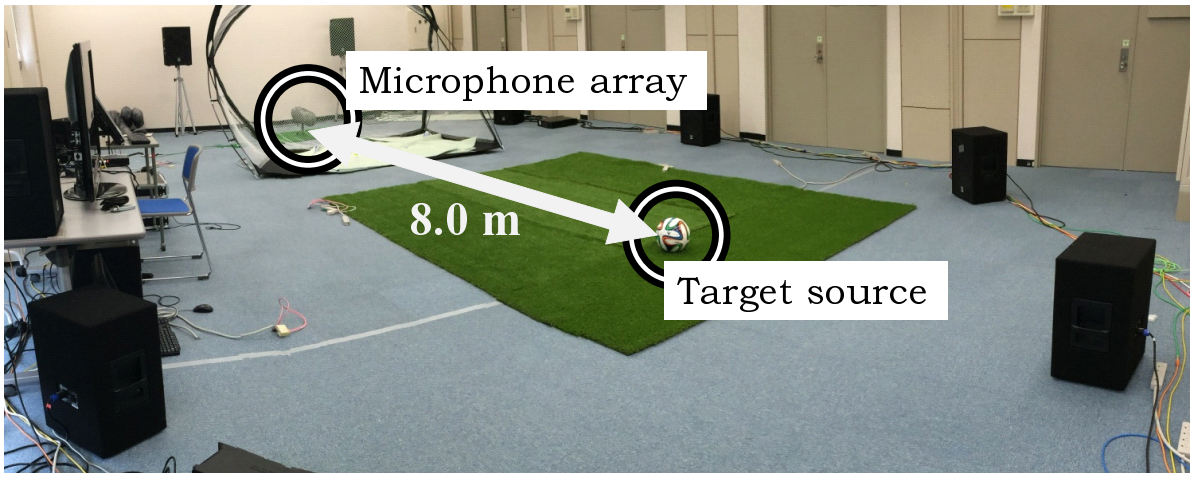


図 3.4: マイクロホンとスピーカの配置図.

表 3.1: 実験で比較した音響特徴量と回帰関数の一覧.

Name	Acoustic feature	Mapping
1 : CCA − GMM	Compressed using CCA	GMM-based mapping
2 : GMM − all	All candidates	GMM-based mapping
3 : pGMM	Selected using proposed method	GMM-based mapping
4 : CCA − DNN	Compressed using CCA	DNN-based mapping
5 : DNN − all	All candidates	DNN-based mapping
6 : pDNN	Selected using proposed method	DNN-based mapping
7 : DNN − M	32 log-mel-filterbank outputs of two microphones by concatenating five consecutive (current and four previous) frames	DNN-based mapping
8 : IRM	None	Ground truth

ysis）[135] で次元圧縮し，それを GMM 回帰で事前 SNR を推定する．圧縮後の特徴量の次元数は，3 : pGMM に用いる音響特徴量と同じ次元数とする．

2 : GMM − all　すべての音響特徴量候補を用いて，GMM 回帰で事前 SNR を推定する．

4 : CCA − DNN　音響特徴量候補を CCA[135] で次元圧縮し，それを DNN 回帰で事前 SNR を推定する．圧縮後の特徴量の次元数は，6 : pDNN に用いる音響特徴量と同じ次元数とする．

5 : DNN − all　すべての音響特徴量候補を用いて，DNN 回帰で事前 SNR を推定する．

提案法と従来法の違いは音響特徴量の選択法である．1 : CCA − GMM および 2 : GMM − all と比べ 3 : pGMM が，また 4 : CCA − DNN および 5 : DNN − all と比べ 6 : pDNN の，音源強調の性能が向上するならば，提案法による音響特徴量選択が源音源の推定に適しているといえる．また事前に音響特徴量を抽出しない方法として，メルフィルタバンク出力を音響特徴量とし DNN 回帰で事前 SNR を推定する手法 [70] も評価した 7 : DNN − M．メルフィルタバンク数は $\mathsf{B} = 32$ とし，以下のように，2 つのマイクロホンでの観測信号を過

去 5 フレーム分連結させたものを音響特徴量とした．

$$\begin{aligned}\boldsymbol{f}_\tau = \ln(&X_{1,1,\tau-4}, ..., X_{1,\mathsf{B},\tau-4}, X_{2,1,\tau-4}, \\ &..., X_{2,\mathsf{B},\tau-4}, X_{1,1,\tau-3}, ..., X_{2,\mathsf{B},\tau})^\top,\end{aligned} \tag{3.42}$$

ここで $X_{m,b,k}$ は，$X_{m,(1,...,\Omega),k}$ を b 番目のメルフィルタバンクで分析した出力である．また，音源強調性能の上限として，IRM で音源強調した結果も評価した 8 : IRM [136]．

学習データとテストデータは，目的音 $H_{m,\theta_1,\omega}V^{(S)}_{m,\omega}S_{\omega,\tau}$ と，雑音 $\sum_{k=1}^{K} H_{m,\theta_k,\omega}V^{(N_k)}_{m,\omega}N_{n,\omega,\tau}$ を重畳することで生成した．目的音と雑音は，図 3.4 (c) のようにスポーツ場を模擬した環境で個別に収録した．収録時の標本化周波数は 48 kHz とした．FOOT と BASE の学習データ数は 1200 サンプル（目的音 100 種類，再生箇所 3 箇所，雑音レベル 4 種類の組み合わせ）とした．また，SHOUT の学習データ数も 1200 サンプル（目的音 50 種類，再生箇所 3 箇所，雑音レベル 4 種類，雑音の種類はサッカー場と野球場の 2 種類の組み合わせ）とした．学習データ量が少ないため，事前 SNR は $\mathsf{B} = 32$ で圧縮したものを推定し，ウィナーフィルタの設計の際に事前 SNR をスプライン補間で線形周波数に補間した．GMM 回帰のための GMM の混合数は $\mathsf{C} = 32$ とした．また提案法で選択された音響特徴量は，GMM 回帰の入力とするには次元数がまだ大きいため，3.5.2 節で説明する GMM 回帰のための相互情報量最大化に基づく音響特徴量次元圧縮法を用いて 32 次元まで圧縮した．DNN 回帰に用いた DNN の構造は，隠れ層が 2 層，隠れ層のユニット数が 128，活性化関数はシグモイド関数とした．また過適合を防ぐため，ドロップアウトと早期終了（early-stopping）アルゴリズムを用いた．ドロップアウト確率は，入力層が 0.2，隠れ層が 0.5 とした．また，早期終了の手順は以下とした．

1. DNN パラメータをすべての学習用ミニバッチを用いて SGD で更新する．
2. 検定用データセットを用いて，平均二乗誤差を評価する
3. 更新後の平均二乗誤差が，更新前の平均二乗誤差より増大したならば，勾配法のステップサイズを半減させる．
4. 勾配法のステップサイズが一定値以下になったら，更新を終了する．

DNN は MMSE 基準の discriminative pre-training [86] を用いて初期化し，その後 fine-tuning した．また，GMM 回帰および DNN 回帰のすべての手法で，音響特徴量の各次元の平均 0，分散 1 となるよう正規化した．さらに，出力された事前 SNR は global variance equalization [47] を用いて補正を行った．その他実験条件はヒューリスティックに決定した．詳細な実験条件を表 3.2 に示す．

音響特徴量候補は，空間情報に基づくものと，スペクトル構造に基づくものを用いた．空間情報に基づく音響特徴量を計算するために，局所 PSD 推定法 [24] を用いた．この

表 3.2: 実験条件.

Parameters for signal processing	
Sampling rate	48.0 kHz
FFT length	512 pts
FFT shift length	256 pts
# of microphones	2
# of mel-filterbanks B	32
Training SNR (dB)	-25, -20, -15, -10
Parameters for ISTA	
Total dimensions of potential acoustic feature Q	238
Mini-batch size K_z	512
Kernel parameter ς_ξ^2	10^{-3}
Regularization coefficient ϵ	10^{-3}
Initial value of $\boldsymbol{a}$	$10^{-2} \times \mathbf{1}_{\mathsf{Q}}$
Initial value of $\boldsymbol{r}$	$\mathbf{0}_{\mathsf{Q}}$
Initial value of $\boldsymbol{s}$	$\mathbf{0}_{\mathsf{Q}}$
Initial value of $\boldsymbol{\nu}$	$\mathbf{0}_{\mathsf{Q}}$
Early-stopping parameter $\mathcal{J}_{\text{count-max}}$	7
Initial stepsize α	10^{-2}
AdaDelta parameter γ	0.9
Momentum parameter η	0.9
Soft threshold parameter λ	5×10^{-3}
Training parameters for DNN-based mapping	
# of hidden layers for DNNs	2
# of hidden units for DNNs	128
Initial stepsize	10^{-3}
Stepsize threshold for early stopping	10^{-6}
Dropout probability (input layer)	0.2
Dropout probability (hidden layer)	0.5
Training parameters for GMM-based mapping	
# of mixture C	32

表 3.3: 音響特徴量候補.

Group	Acoustic feature	Dim.
/1/	Log-spatial PSD (target region)	32
/2/	Log-spatial PSD (noise region)	32
/3/	Δ Spatial PSD (target region)	32
/4/	Δ Spatial PSD (noise region)	32
/5/	MFCC (target region)	23
/6/	MFCC (noise region)	23
/7/	Log-mel-filterbank outputs (shotgun mic.)	32
/8/	Spatial-information-based Wiener filter [24]	32

手法では，目的音方向に存在する音源群の PSD $\Gamma_{1,b,\tau}$ と，雑音方向に存在する音源群の PSD $\Gamma_{2,b,\tau}$ を，以下のように計算する．

$$\begin{bmatrix}\Gamma_{1,b,\tau}\\ \Gamma_{2,b,\tau}\end{bmatrix} = \left(\begin{bmatrix}|H_{1,0,b}|^2 & |H_{1,\pi,b}|^2\\ |H_{2,0,b}|^2 & |H_{2,\pi,b}|^2\end{bmatrix}\right)^{+} \begin{bmatrix}|X_{1,b,\tau}|^2\\ |X_{2,b,\tau}|^2\end{bmatrix}, \tag{3.43}$$

この式からわかるように，$\Gamma_{1,b,\tau}$ と $\Gamma_{2,b,\tau}$ は，$|X_{1,b,\tau}|^2$ と $|X_{2,b,\tau}|^2$ から，各方向に存在する音源群の PSD を強調したものともみなせる．ただし $H_{m,\theta,b}$ には，無響室で測定した各マイクロホンの指向特性を用いた．また，$\Gamma_{1,b,\tau}$ と $\Gamma_{2,b,\tau}$ の推定値が負の値をとった場合，その値は 10^{-3} で置き換えた．音響特徴量候補は以下のようになる．

- $\ln\Gamma_{1,b,\tau}$ と $\ln\Gamma_{2,b,\tau}$ のメルフィルタバンク出力，およびその時間方向への一回差分
- $\Gamma_{1,b,\tau}$ と $\Gamma_{2,b,\tau}$ から計算した MFCC
- マイクロホン m_1 の観測信号のメルフィルタバンク出力 $\ln|X_{m_1,b,\tau}|$
- $\Gamma_{1,b,\tau}$ と $\Gamma_{2,b,\tau}$ から計算したウィナーフィルタ（$\Gamma_{1,b,\tau}/(\Gamma_{1,b,\tau}+\Gamma_{2,b,\tau})$）

以上より，音響特徴量候補の次元数は $\mathsf{Q} = 238$ である．これらの音響特徴量は，表 3.3 のように，8 つのグループに分割した．

3.3.2 音響特徴量選択の動作確認実験

本節では，提案法の挙動を解析するために，2 つの実験を行う．1 つは，提案法で選択された特徴量とその重み a_q を，各競技音ごとに可視化することで，競技音の特性と音響特徴量の特性に関連があるか，定性的に論じる．また，提案法で選択した音響特徴量

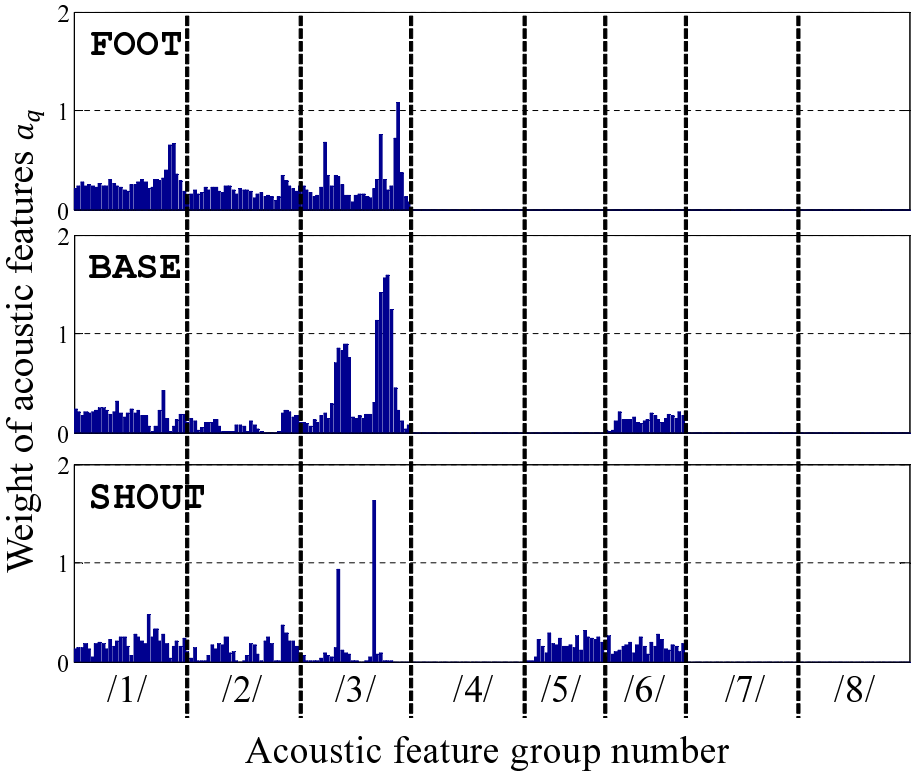


図 3.5: 音響特徴量選択の実行結果.

と事前 SNR の相互情報量，および従来法で選択した音響特徴量と事前 SNR の相互情報量を比較し，提案法が相互情報量を増加させることを確認する．

提案法による音響特徴量の選択結果を定性的に議論するために，図 3.5 に，音響特徴量候補に対する重み a_q を可視化した．重み a_q が非零となる音響特徴量が，選択された音響特徴量であることを示す．図 3.5 から，全ての競技音で，/1/ $\ln\Gamma_{1,b,\tau}$ と /2/ $\ln\Gamma_{2,b,\tau}$ が選択されており，また /7/ $\ln|X_{m_1,b,\tau}|$ が選択されていないことがわかる．/1/ $\ln\Gamma_{1,b,\tau}$ と /7/ $\ln|X_{m_1,b,\tau}|$ は，両者とも目的音方向の対数スペクトルの音響特徴量であり，その違いは局所 PSD 推定による目的音方向の事前強調処理の有無だけである．このことから，スポーツ場のように，目的音方向が事前に既知であるならば，/1/ $\ln\Gamma_{1,b,\tau}$ と /2/ $\ln\Gamma_{2,b,\tau}$ のように目的音方向のスペクトルを事前に強調した音響特徴量（音源の空間情報）を用いることが有効なことがわかる．さらに，FOOT と BASE では，/3/ $\Delta\Gamma_{1,b,\tau}$ が選択されている．FOOT や BASE は，音のパワーが時間方向に急激に変化する突発音であるため，そのパワーの時間変化を示すスペクトルの時間差分（音源の時間変化情報）を音響特徴量に用いることが有効なことがわかる．一方 SHOUT では，/5/ MFCC が選択されている．音声認識などでは，スペクトルの包絡情報が音素の識別に有効であることが知られており，この結果からも，音声を強調するにはスペクトルの包絡情報（音源の音色情報）が有効であることがわかる．

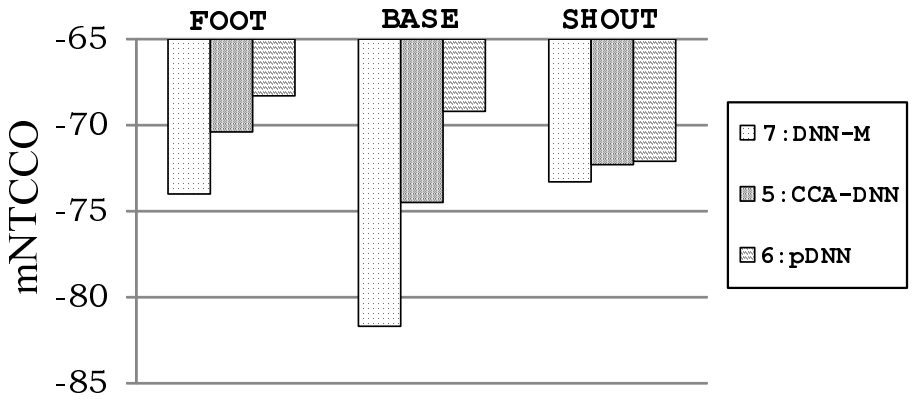


図 3.6: 条件付き共分散作用素の中央値（mNTCCO）.

提案法による音響特徴量の選択結果を定量的に議論するために，音響特徴量と事前SNRの NTCCO [128, 129] を評価する．提案法 6 : pDNN の NTCCO を，CCA による次元圧縮 4 : CCA − DNN および，観測信号のメルフィルタバンク出力 7 : DNN − M と比較した．NTCCO は相互情報量と比例するのため，提案法で選択した音響特徴量の NTCCO が従来法よりも増加するならば，提案法は従来法よりも，事前 SNR との相互情報量が大きい音響特徴量を選択できるといえる．NTCCO を簡便に計算するために，学習データの各ミニバッチから計算した NTCCO の中央値（mNTCCO: median of the NTCCO）を評価した．

$$\mathrm{mNTCCO} = \mathrm{median}\left\{-\mathrm{Tr}\left[\Phi_{\xi_z}\left(\Phi_{g_z} + \epsilon \boldsymbol{I}\right)^{-1}\right]\right\}_z, \tag{3.44}$$

ここで Φ_{ξ_z} と Φ_{g_z} は，z 番目のミニバッチから計算した，音響特徴量と事前 SNR の中心化グラム行列である．図 3.6 に mNTCCO の計算結果を示す．全ての競技音で，提案法の mNTCCO が従来法の mNTCCO よりも大きかった．この結果から，提案法は従来法よりも事前 SNR との相互情報量が大きい音響特徴量を選択できることが示唆された．

3.3.3 客観評価実験

提案法の音源強調性能を客観的に評価するために，以下の式で計算される SDR を用いた性能比較を行った．

$$\mathrm{SDR} = 10\log_{10}\left(\frac{|S_{\omega,\tau}|^2}{\left|S_{\omega,\tau} - \hat{S}_{\omega,\tau}\right|^2}\right) \tag{3.45}$$

SDR の計算には E. Vincent らが公開している “BSS EVAL Toolbox [134]” を用いた．FOOT と BASE の学習データ数は 15 サンプル（目的音 5 種類，再生箇所 3 箇所）とした．また，SHOUT の学習データ数は 30 サンプル（目的音 5 種類，再生箇所 3 箇所，雑音 2 種類）とした．また雑音レベルは信号対雑音比が -10 と -20 dB の 2 種類で実験した．

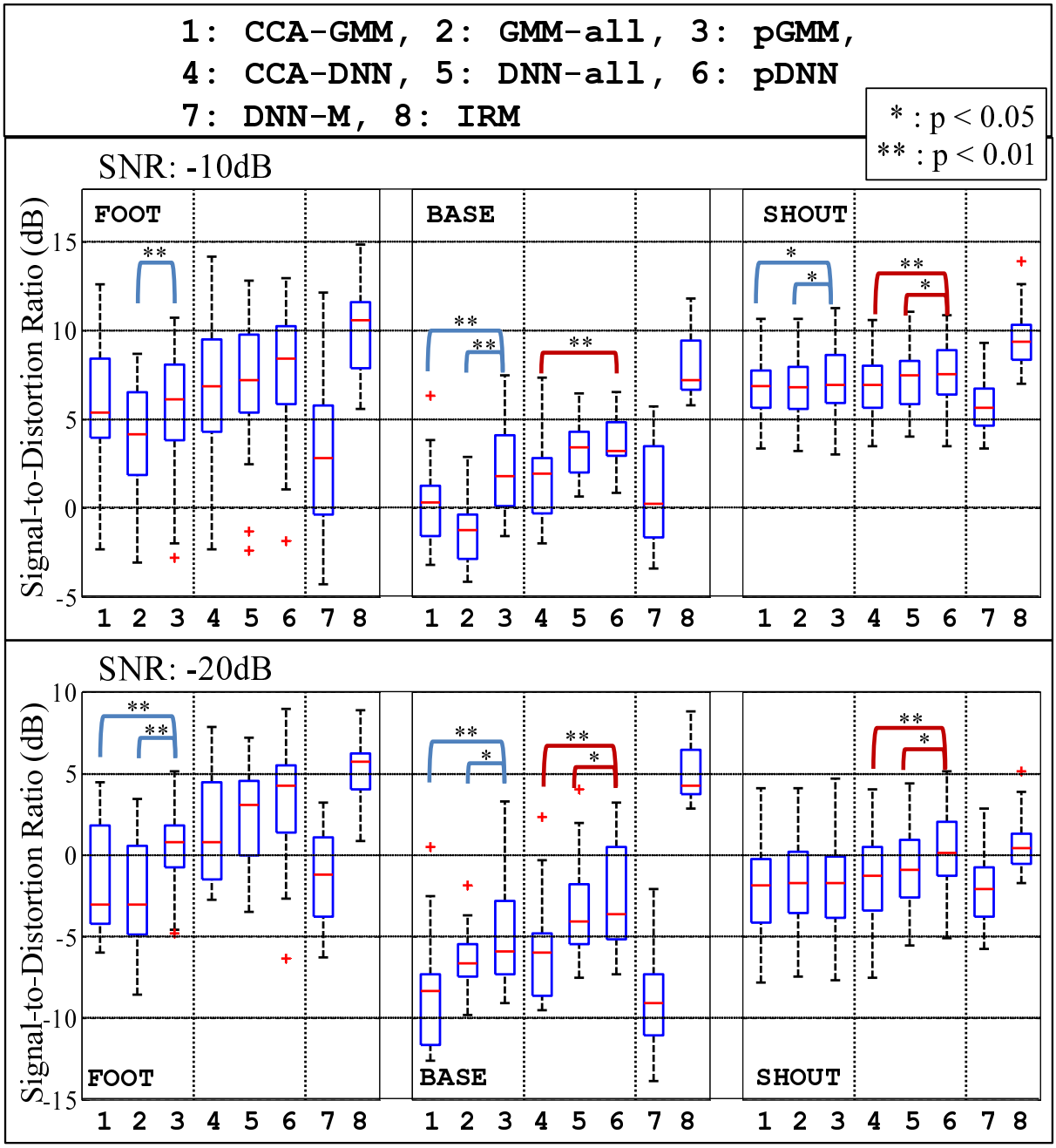


図 3.7: SDR の評価結果．アスタリスクの数は，t 検定における有意水準を表す．

図 3.7 に提案法と従来法の SDR を示す．提案法と従来法の比較では，全ての競技音および雑音レベルにおいて 6 : pDNN の SDR が最も高かった．また t 検定で各射影関数ごと，つまり 3 : pGMM は 1 : CCA − GMM および 2 : GMM − all，6 : pDNN は 4 : CCA − DNN および 5 : DNN − all と有意差を評価した結果，いくつかの競技音および雑音レベルにおいて有意差が認められた．また図 3.8 に，SNR -10 dB の雑音レベルにおける，DNN 射影の出力音の波形を示す．特に雑音区間において，提案法は雑音をより抑圧できていることがわかる．

提案法は特徴量候補から相互情報量を増加させる有益な特徴量を選択する手法のため，特徴量候補をすべて用いる場合より相互情報量が増加することはない．つまり

$$\iint p(\boldsymbol{\xi}, \boldsymbol{g}) \ln \frac{p(\boldsymbol{\xi}, \boldsymbol{g})}{p(\boldsymbol{\xi})p(\boldsymbol{g})} d\boldsymbol{\xi} d\boldsymbol{g} \geq \iint p(\boldsymbol{\xi}, \boldsymbol{A}\boldsymbol{g}) \ln \frac{p(\boldsymbol{\xi}, \boldsymbol{A}\boldsymbol{g})}{p(\boldsymbol{\xi})p(\boldsymbol{A}\boldsymbol{g})} d\boldsymbol{\xi} d\boldsymbol{g} \tag{3.46}$$

である．ゆえに学習データが十分に存在するならば，2 : GMM − all と 3 : pGMM，および 5 : DNN − all と 6 : pDNN を比較した場合，音源強調の性能はほぼ同等となる．しかし，いくつかの条件では，提案法の SDR が音響特徴量をすべて用いる従来法よりも有意に上回っていることが確認できる．この理由は，スポーツ競技音などの学習データ量が十分に集まらない源信号の強調では，回帰関数のパラメータ数が多い従来法は学習が十分に進まなかったためと考えられる．提案法は，事前 SNR と関係のない無益な特徴量を無視し，有益な音響特徴量のみを用いて DNN や GMM を学習するため，従来法と比べてパラメータ数を削減して学習を効率化できている．この結果から，特にスポーツ競技音など，所望の源信号の学習データを大量に集めることが困難な場合に，提案法による音響特徴量の選択は有効であるといえる．

3.3.4 主観評価実験

提案法の音源強調性能を主観的に評価するために，主観評価実験を行った．試験は比較平均オピニオン値（CMOS: comparison mean opinion score）試験とし，8 人の被験者が，観測音と出力音の音質と明瞭性を比較評価したただし，本試験は発話音声の通話向けの聞き取り試験ではないため，明瞭性の定義を「目的音の聞き取りやすさとし」，被験者は「観測音と比べ出力音は，キックやバッティングが行われていることを知覚しやすくなったか」という基準で比較試験を行った．CMOS 試験には，-3 – 非常に悪い，0 – ほぼ同じ，+3 – 非常に良い，の 7 段階の評点を用いた．

被験者の身体的な負担を考慮し，主観評価試験の対象は提案法 6 : pDNN と，CCA による音響特徴量選択の従来法 4 : CCA − DNN，およびベースライン 7 : DNN − M とウィナーフィルタリングの性能限界である 8 : IRM の 4 種類とした．比較音は各手法で 22 サンプルとした．また目的音の変化を考慮するために，目的音は指向性マイクの焦点方向から

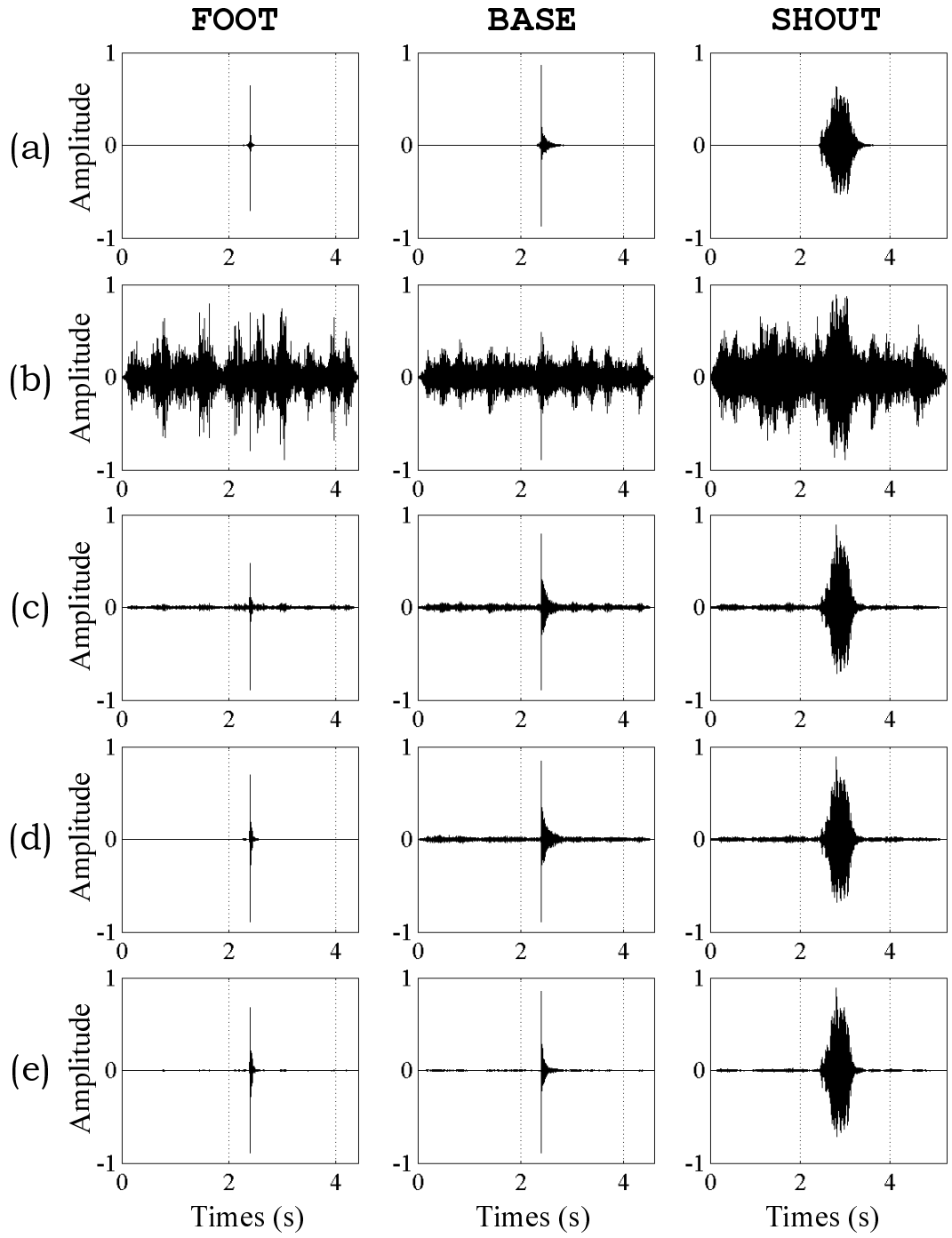


図 3.8: 音源強調結果の波形例．(a) 源信号，(b) 観測信号，(c) 4 : CCA − DNN の出力信号，(d) 5 : DNN − all の出力信号，(e) 6 : pDNN の出力信号．

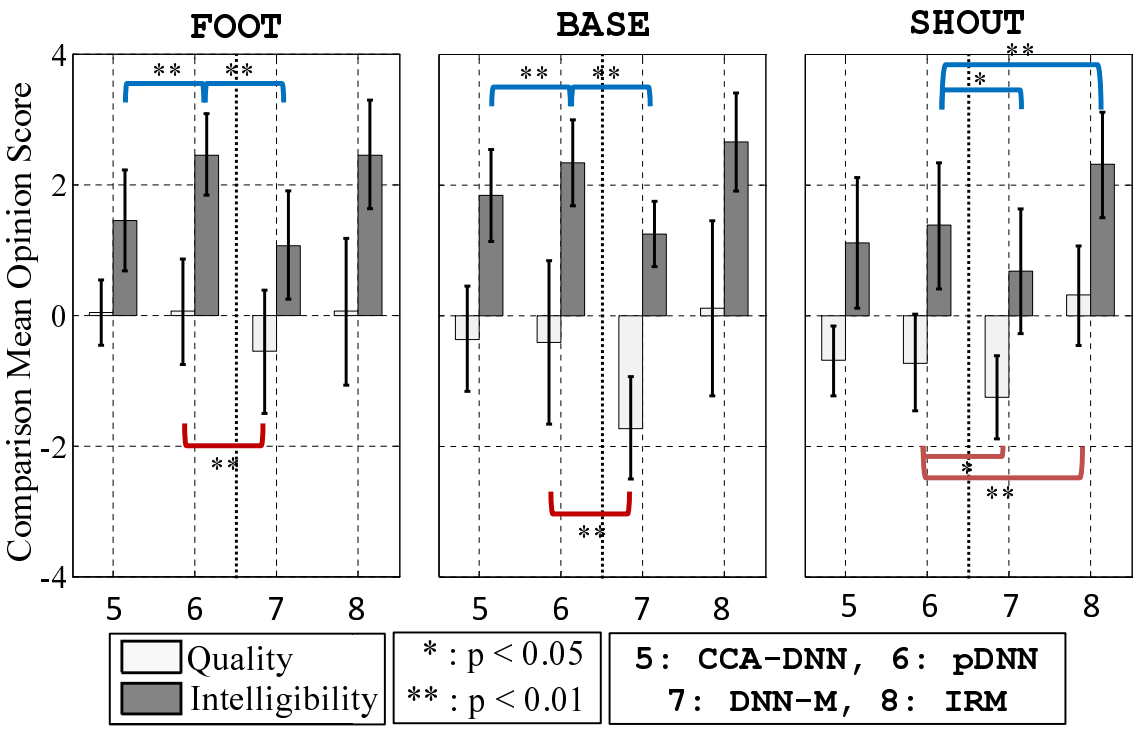


図 3.9: 主観評価結果．アスタリスクの数は，t 検定における有意水準を表す．

ずらしたスピーカから再生した（図 3.4(a)）雑音レベルは，SNR が -10 dB となるように設定した．

図 3.9 に CMOS 試験の結果を示す．明瞭度の比較において，提案法 6 : pDNN の評点はベースライン 7 : DNN − M および従来法 4 : CCA − DNN の評点よりも高くなった．また t 検定で有意差を評価した結果，いくつかの競技音において有意差が認められた．さらに FOOT と BASE においては，ウィナーフィルタリングの性能限界である 8 : IRM と有意差が認められなかった．このことから，目的音をクリアに抽出する MMSE 基準の音源強調において，(1) 有益な音響特徴量が満たすべき性質は相互情報量を最大化することであり，(2) 学習データが十分に存在しない場合に，提案法でこの基準を満たすよう音響特徴量を選択することで,「明瞭性」の意味で音源強調の性能が向上することをがわかる．

一方音質の評価では，提案法 6 : pDNN の評点は 0 を下回っており，従来法 4 : CCA − DNN の評点と比べほぼ同等である．さらに，有意差検定を行った結果，ウィナーフィルタリングの性能限界である 8 : IRM と有意差が認められている．これは，提案法で選択された音響特徴量は従来法と比べ,「音質」の意味で音源強調の性能を向上させることはなく，さらに観測音と比べ音質を劣化させていることを示している．これは，源信号の推定結果の二乗誤差の大きさと人間が知覚する音質の劣化の大きさは必ずしも比例しない [49] ためであり，主観的な音質評価を最大化するためには MMSE 以外の目的関数で DNN を学習する必要がある．主観的な音質評価を最大化するための音源強調法については，4

章で取り組む.

3.4 本章のまとめ

本章では，雑音下で特定の種類の源信号を強調するオブジェクトベース集音を実現するために，スポーツの競技音など，ラベルデータが十分に存在しない源信号を強調するための手法について研究した．DNN を用いてラベルデータが十分に存在しない源信号を強調するためには，MFCC や LPC などの音響特徴量の候補から人手で事前に設計/選択した音響特徴量を観測信号から抽出することでネットワークのサイズを小さくして DNN の自由度を下げる必要がある．しかし，音響特徴量の候補の次元は大きく最適な組み合わせを探索的に決定することは困難であり，また音源の種類によって適切な音響特徴量が異なるため，音響特徴量を人手で選択することは困難だった．そこで本章では DNN の推定誤差について確率分布を明示的に仮定し，その分布の性質から確率論的に目的関数を設計することで，適切な音響特徴量を自動選択する方法を目指した．源信号の推定誤差がガウス分布に従うと仮定し，源信号の推定誤差を最小化するための音響特徴量選択の目的関数として相互情報量を利用した．相互情報量を正確に計算するための手段として，福水らの提案した相互情報量を再生核ヒルベルト空間で計算する “カーネル次元圧縮 [128, 129]” を採用した．本章の新規性は，特徴量候補の次元数が大きい音響特徴量選択にカーネル次元圧縮法を適用するために，スパース正則化法に基づく微分可能な目的関数を導出し，大量な音響特徴量候補から適切な音響特徴量を勾配法により選択できる音響特徴量選択法を提案したことにある．定量評価試験では，提案法を用いて音響特徴量を選択することで従来の音響特徴量選択法と比べ SDR が向上することを示した．また主観評価試験では，提案法を用いて音響特徴量を選択することで CCA に基づく次元削減法と比べ源信号の明瞭性が向上することを示した．一方で音質については従来法と比べて改善は見られなかった．これは，源信号の推定結果の二乗誤差の大きさと人間が知覚する音質の劣化の大きさは必ずしも比例しない [49] ためであり，主観的な音質評価を最大化するためには MMSE 以外の目的関数で DNN を学習する必要がある．主観的な音質評価を最大化するための音源強調法については，4 章で取り組む.

3.4.1 本章の貢献と関連研究

本章の内容は，研究業績リスト [J-1] の内容をまとめたものである．この研究の貢献は，福水らの提案したカーネル次元圧縮法を特徴量候補の次元数が大きい音響特徴量選択に応用するために，スパース正則化法を利用した微分可能な目的関数を導出した点にある.

学習データが十分に得られないときに，音響特徴量を事前に選択することでニューラルネットワークの自由度を低減することで，過適合を防ぐアプローチは従来から存在していた．また，特徴量選択の研究分野では，相互情報量最大化を特徴量選択の目的関数として利用するアプローチや，変数間の高次相関を再生核ヒルベルト空間で計量するアプローチは存在していた．しかし，ニューラルネットワークへ入力する音響特徴量を選択するために，カーネル次元圧縮を利用した例は存在しない．これは，音響特徴量は複数の次元（変数）を組み合わせて音響的な特徴を表現するため，音響特徴量候補の次元が大きくなりやすく，その選択には組み合わせ最適化問題を解く必要があるからである．

そこで本章では，高次相関を再生核ヒルベルト空間で計量する際に用いるカーネル関数にガウスカーネルを利用し，ガウスカーネルの特性を利用したスパース正則化に基づく微分可能な目的関数を導出した．本研究で導出した目的関数を用いることで，音響特徴量選択を勾配法で最適化できるようになるため，ニューラルネットワークへ入力する音響特徴量を現実的な計算時間で選択できるようになった．この成果により，これまで推定が困難とされてた，学習データが十分に得られないような源信号や，これまで源信号の推定対象とされてこず，適切な音響特徴量が未知な源信号も推定できるようになった．

3.5 本節の付録

3.5.1 事前 SNR の誤差分布の確認

式 (3.10) で定義される事前 SNR の推定誤差分布を各メル周波数バンド $e_{(1,\ldots,\mathrm{B}),\tau}$ で調査した．調査には 3.3 章の評価実験で用いたデータを利用した．条件は $\mathtt{6:DNN-all}$ のものと同じとし，$\boldsymbol{x}_\tau$ と $\xi_{b,\tau}$ は平均 0，分散 1 となるように事前に正規化した．

図 3.10(a) に $e_{20,\tau}$ のヒストグラムと，平均のガウス分布（赤線）を示す．厳密には一致していないが，おおまかにガウス分布に従っていることが見て取れる．図 3.10(b)(c) に，各メル周波数バンドの推定誤差の平均，分散，歪度，尖度をプロットした．平均，歪度，尖度がそれぞれ，ほぼ 0，0，3 となっており，$e_{b,k}$ は近似的に平均 0 のガウス分布に従うことを確認した．ただし，分散は各バンドで異なっており，式 (3.11) のように等分散のガウス分布でモデル化することは，厳密には成立しないことも確認した．

3.5.2 GMM 回帰と GMM 回帰のための相互情報量最大化に基づく次元圧縮法

3.3 章で用いた GMM 回帰を説明する [137]．また，本論文はニューラルネットワークを用いることを前提とするため本文中では省略したが，回帰関数が GMM 回帰，$\mathcal{F}$ が次元圧縮で実装される際に式 (3.14) をより簡便に実装する方法を紹介する（研究業績リスト [C-5]）．$\mathcal{F}$ を次元圧縮で実装するため，$\mathcal{F}$ の実装方法は本文と類似した $\boldsymbol{f} = \mathcal{F}(\boldsymbol{x}|\Theta_{\mathcal{F}}) = \boldsymbol{A}^\top \boldsymbol{g}$

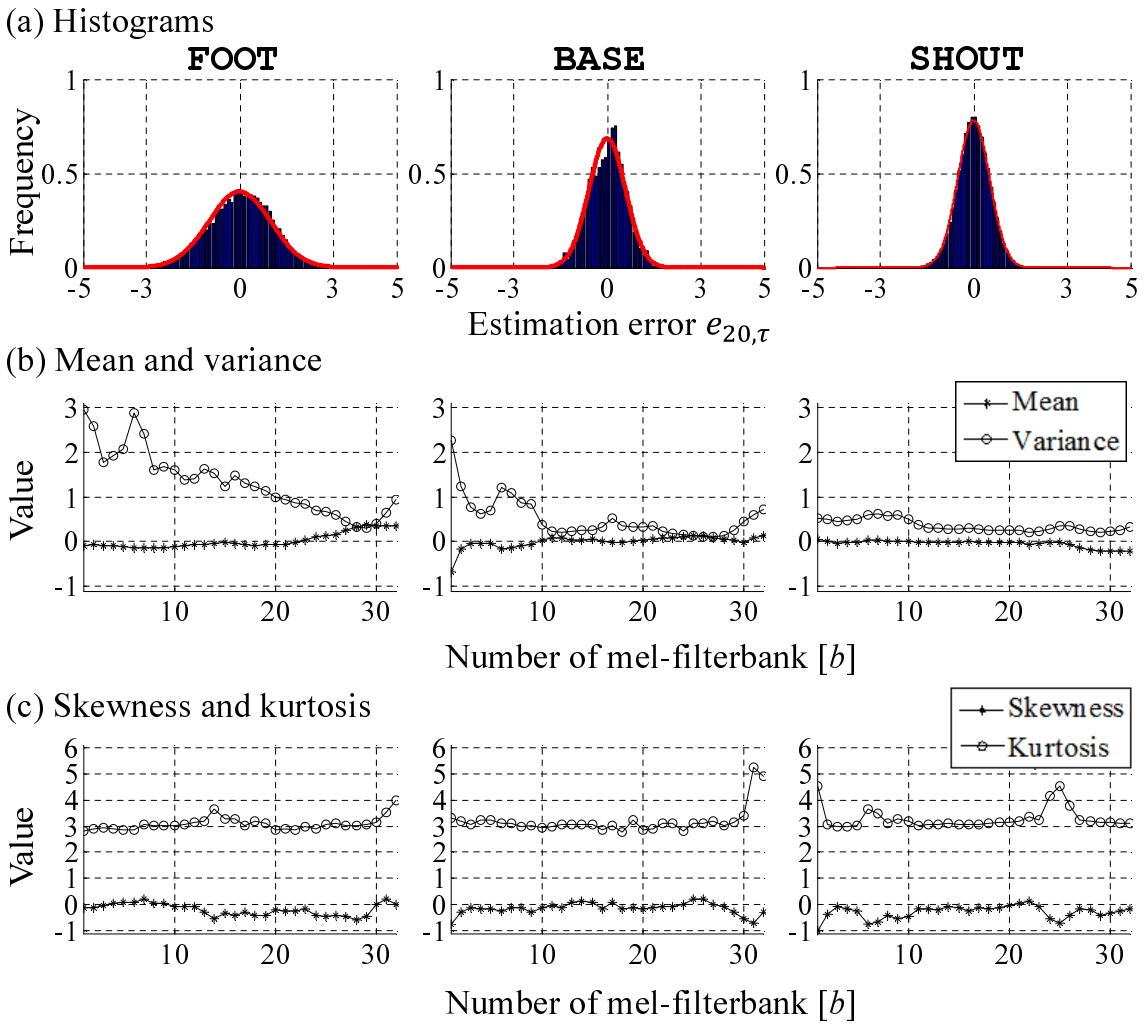


図 3.10: 事前 SNR の誤差の分布形状の調査．(a) $b = 20$ バンドにおける事前 SNR の推定誤差のヒストグラムと平均 0 のガウス分布（赤線），(b) 全バンドにおける事前 SNR の推定誤差の平均と分散，(c) 全バンドにおける事前 SNR の推定誤差の歪度と尖度．

であるが，$\boldsymbol{A}$ はすべての要素に値を持つ圧縮行列であることに注意されたい．

GMM 回帰では，$\boldsymbol{\xi}$ と $\boldsymbol{f}$ の同時確率密度関数を GMM で表現し，$\boldsymbol{f}$ を得た下での $\boldsymbol{\xi}$ の条件付き確率密度関数を求めることで回帰を実行する．まず，$\boldsymbol{\xi}$ と $\boldsymbol{f}$ の結合ベクトルを $\boldsymbol{\zeta} = (\boldsymbol{\xi}, \boldsymbol{f})^\top$ とし，$\boldsymbol{\xi}$ と $\boldsymbol{f}$ の同時確率密度関数を以下でモデル化する．

$$p(\boldsymbol{\zeta}) = \sum_{i=1}^{\mathsf{C}} w_i \mathcal{N}(\boldsymbol{\zeta}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{3.47}$$

ここで C は混合数，$w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$ はそれぞれ i 番目のガウス分布の混合係数，平均ベクトル，分散共分散行列，$\mathcal{N}(\boldsymbol{\zeta}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ は以下の多変量ガウス分布である．

$$\mathcal{N}\left(\boldsymbol{\zeta}|\boldsymbol{\mu}, \Sigma\right) = \frac{|\Sigma|^{-\frac{1}{2}}}{(2\pi)^{\mathsf{D}_\zeta/2}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\zeta} - \boldsymbol{\mu}\right)^\top \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\zeta} - \boldsymbol{\mu}\right)\right\} \tag{3.48}$$

ただし D_ζ は $\boldsymbol{\zeta}$ の次元数である．ここで $\boldsymbol{\mu}_i$ と $\boldsymbol{\Sigma}_i$ はそれぞれ以下のように記述できる．

$$\boldsymbol{\mu}_i = \begin{bmatrix} \boldsymbol{\mu}_i^{\boldsymbol{\xi}} \\ \boldsymbol{\mu}_i^{\boldsymbol{f}} \end{bmatrix}, \quad \boldsymbol{\Sigma}_i = \begin{bmatrix} \boldsymbol{\Sigma}_i^{\boldsymbol{\xi}\boldsymbol{\xi}} & \boldsymbol{\Sigma}_i^{\boldsymbol{\xi}\boldsymbol{f}} \\ \boldsymbol{\Sigma}_i^{\boldsymbol{f}\boldsymbol{\xi}} & \boldsymbol{\Sigma}_i^{\boldsymbol{f}\boldsymbol{f}} \end{bmatrix} \tag{3.49}$$

ここで $\boldsymbol{\mu}_i^{\boldsymbol{\xi}}$ と $\boldsymbol{\mu}_i^{\boldsymbol{f}}$ は i 番目の分布における $\boldsymbol{\xi}$ と $\boldsymbol{f}$ の平均ベクトル，$\boldsymbol{\Sigma}_i^{\boldsymbol{\xi}\boldsymbol{\xi}}$ と $\boldsymbol{\Sigma}_i^{\boldsymbol{f}\boldsymbol{f}}$ は i 番目の分布における $\boldsymbol{\xi}$ と $\boldsymbol{f}$ の分散共分散行列である．$\boldsymbol{\xi}$ と $\boldsymbol{f}$ の同時確率密度関数を式 (3.47) で表現したとき，$\boldsymbol{f}$ を得た下での $\boldsymbol{\xi}$ の条件付き確率密度関数 $p(\boldsymbol{\xi}|\boldsymbol{f})$ はの期待値は以下で計算できる．

$$\begin{aligned} \int \boldsymbol{\xi} p(\boldsymbol{\xi}|\boldsymbol{f}) d\boldsymbol{\xi} &= \sum_{i=1}^{\mathsf{C}} p(i|\boldsymbol{f}) \int \boldsymbol{\xi} p(\boldsymbol{\xi}|i, \boldsymbol{f}) d\boldsymbol{\xi} \\ &= \sum_{i=1}^{\mathsf{C}} \mathcal{W}_i \int \boldsymbol{\xi} \mathcal{N}\left(\boldsymbol{\xi}|\boldsymbol{\mu}_i^{\boldsymbol{\xi}|\boldsymbol{f}}, \Sigma_i^{\boldsymbol{\xi}\boldsymbol{\xi}|\boldsymbol{f}}\right) d\boldsymbol{\xi} \end{aligned} \tag{3.50}$$

$$\mathcal{W}_i = \frac{w_i \mathcal{N}\left(\boldsymbol{f}|\boldsymbol{\mu}_i^{\boldsymbol{f}}, \boldsymbol{\Sigma}_i^{\boldsymbol{f}\boldsymbol{f}}\right)}{\sum_{j=1}^{\mathsf{C}} w_j \mathcal{N}\left(\boldsymbol{f}|\boldsymbol{\mu}_j^{\boldsymbol{f}}, \boldsymbol{\Sigma}_j^{\boldsymbol{f}\boldsymbol{f}}\right)} \tag{3.51}$$

ここで $p(\boldsymbol{\xi}|i, \boldsymbol{f})$ は i 番目の条件付きガウス分布であり，その平均ベクトルと分散共分散行列は以下のように計算できる．

$$\boldsymbol{\mu}_i^{\boldsymbol{\xi}|\boldsymbol{f}} = \boldsymbol{\mu}_i^{\boldsymbol{\xi}} + \boldsymbol{\Sigma}_i^{\boldsymbol{\xi}\boldsymbol{f}} \left(\boldsymbol{\Sigma}_i^{\boldsymbol{f}\boldsymbol{f}}\right)^{-1} \left(\boldsymbol{f} - \boldsymbol{\mu}_i^{\boldsymbol{f}}\right) \tag{3.52}$$

$$\Sigma_i^{\boldsymbol{\xi}\boldsymbol{\xi}|\boldsymbol{f}} = \boldsymbol{\Sigma}_i^{\boldsymbol{\xi}\boldsymbol{\xi}} - \boldsymbol{\Sigma}_i^{\boldsymbol{\xi}\boldsymbol{f}} (\boldsymbol{\Sigma}_i^{\boldsymbol{f}\boldsymbol{f}})^{-1} \boldsymbol{\Sigma}_i^{\boldsymbol{f}\boldsymbol{\xi}} \tag{3.53}$$

MMSE を最小化するのは平均値であるため，GMM 回帰は以下のように実装できる．

$$\mathcal{M}(\boldsymbol{f}|\Theta_{\mathcal{M}}) = \sum_{i=1}^{\mathsf{C}} \mathcal{W}_i \left(\boldsymbol{\mu}_i^{\boldsymbol{\xi}} + \boldsymbol{\Sigma}_i^{\boldsymbol{\xi}\boldsymbol{f}} \left(\boldsymbol{\Sigma}_i^{\boldsymbol{f}\boldsymbol{f}}\right)^{-1} \left(\boldsymbol{f} - \boldsymbol{\mu}_i^{\boldsymbol{f}}\right)\right) \tag{3.54}$$

同時確率密度関数や条件付き確率密度関数をガウス分布で明示的に記述したことで，式 (3.14) を明示的な形で記述することができるようになる．まず相互情報量における期待値演算を，学習データの算術平均に置きかえる．すると目的関数は

$$\mathcal{J} = \sum_{\tau=1}^{T} \ln \frac{p(\mathcal{F}(\boldsymbol{x}_\tau|\Theta_{\mathcal{F}}), \boldsymbol{\xi}_\tau)}{p(\mathcal{F}(\boldsymbol{x}_\tau|\Theta_{\mathcal{F}}))p(\boldsymbol{\xi}_\tau)} \tag{3.55}$$

$$= \sum_{\tau=1}^{T} \ln p(\boldsymbol{\xi}_\tau|\boldsymbol{A}^\top \boldsymbol{g}_\tau) - \ln p(\boldsymbol{A}^\top \boldsymbol{g}_\tau) \tag{3.56}$$

と記述できる．$p(\boldsymbol{\xi}_\tau|\boldsymbol{A}^\top \boldsymbol{g}_\tau)$ と $p(\boldsymbol{A}^\top \boldsymbol{g}_\tau)$ は GMM で明示的に表現できるため，式 (3.56) は勾配法で数値的に最大化することができる．ここで $\boldsymbol{A}^\mathsf{T}\boldsymbol{g}_\tau$ が $\boldsymbol{A}^\mathsf{T}\boldsymbol{g}_\tau = \sum_{q=1}^{\mathsf{Q}} \boldsymbol{a}_q g_{q,\tau}$ で

記述できることに着目すれば，$\boldsymbol{A}$ の各行ベクトル $(\boldsymbol{a}_{1,\ldots,\mathsf{Q}})$ は以下のように勾配法で最適化できる．

$$\boldsymbol{a}_q \leftarrow \boldsymbol{a}_q - \epsilon \nabla \boldsymbol{a}_q, \tag{3.57}$$

ここで ϵ はステップサイズであり，勾配ベクトル $\nabla \boldsymbol{a}_q$ は以下のように計算できる．

$$\nabla \boldsymbol{a}_q = \frac{1}{T} \sum_{\tau=1}^{T} g_{q,\tau} \sum_{i=1}^{\mathsf{C}} \gamma_{i,\tau} \left\{ \boldsymbol{d}_{\xi,i,\tau}^{\mathsf{T}} \boldsymbol{\Lambda}_k^{\boldsymbol{\xi f}} + \boldsymbol{d}_{f,i,\tau}^{\mathsf{T}} \boldsymbol{\Lambda}_k^{\boldsymbol{ff}} \right\} - \eta_{i,\tau} \boldsymbol{d}_{f,i,\tau}^{\mathsf{T}} \left(\boldsymbol{\Sigma}_k^{\boldsymbol{ff}} \right)^{-1} \tag{3.58}$$

$$\boldsymbol{d}_{\xi,i,\tau} = \boldsymbol{\xi}_\tau - \boldsymbol{\mu}_i^{\boldsymbol{\xi}} \tag{3.59}$$

$$\boldsymbol{d}_{f,i,\tau} = \boldsymbol{A}^{\mathsf{T}} \boldsymbol{g}_\tau - \boldsymbol{\mu}_i^{\boldsymbol{f}} \tag{3.60}$$

$$\gamma_{i,\tau} = \frac{w_i \mathcal{N}\left(\boldsymbol{\zeta}_\tau | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)}{\sum_{j=1}^{\mathcal{C}} w_j \mathcal{N}\left(\boldsymbol{\zeta}_\tau | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)} \tag{3.61}$$

$$\eta_{i,\tau} = \frac{w_i \mathcal{N}\left(\boldsymbol{A}^{\mathsf{T}} \boldsymbol{g}_\tau | \boldsymbol{\mu}_i^{\boldsymbol{f}}, \boldsymbol{\Sigma}_i^{\boldsymbol{ff}}\right)}{\sum_{j=1}^{\mathcal{C}} w_j \mathcal{N}\left(\boldsymbol{A}^{\mathsf{T}} \boldsymbol{g}_\tau | \boldsymbol{\mu}_j^{\boldsymbol{f}}, \boldsymbol{\Sigma}_j^{\boldsymbol{ff}}\right)} \tag{3.62}$$

$$\left(\boldsymbol{\Sigma}_i\right)^{-1} = \begin{bmatrix} \boldsymbol{\Lambda}_i^{\boldsymbol{\xi\xi}} & \boldsymbol{\Lambda}_i^{\boldsymbol{\xi f}} \\ \boldsymbol{\Lambda}_i^{\boldsymbol{f\xi}} & \boldsymbol{\Lambda}_i^{\boldsymbol{ff}} \end{bmatrix} \tag{3.63}$$

式 (3.57) で圧縮行列を更新した場合，同時確率密度関数である GMM も更新しなくてはならない．そこで，式 (3.56) の最大化は，GMM の更新と圧縮行列の更新を交互に行う一般化 EM（GEM: generalized expectation-maximization）アルゴリズムで行う．また更新の安定のために，式 (3.57) の実行毎に $\boldsymbol{A}$ を準直交化する [138]．GEM アルゴリズムの流れは以下である．

E-step：

1. 式 (3.61) で $\gamma_{i,\tau}$ を更新する．

M-step：

1. 同時確率密度関数を EM アルゴリズムで更新する

2. $\boldsymbol{A}$ を更新する
   2-1. 各行ベクトルを式 (3.57) で更新する
   2-2. $\boldsymbol{A}$ を準直交化する
   2-3. アルゴリズムが収束しなければ 2-1 に戻る

第 4 章

聴感評点を最大化する音源強調ための目的関数

本章では，主観品質を最大化する時間周波数マスクのような，ラベルデータを一意に定めることのできない源信号の推定について考える．ニューラルネットワークは，フレーム結合した観測信号を入力とし，時間周波数マスクを選択/生成するための確率もしくは確率分布のパラメータを出力するための関数として用いる．以降では，主観評価値と相関の高い音質評価値（聴感評点）[50, 51, 52] を報酬ととらえ，強化学習のフレームワークを応用し，聴感評点を最大化するようにニューラルネットワークを学習するための目的関数を提案する．

4.1 強化学習に基づく音源強調

3 章では，高臨場音響系の実現に向け，無数の源信号が含まれた観測信号から，所望の源信号を選択的に強調するための “オブジェクトベース収音技術” について研究を行ってきた．本章では，高品質な音声通信や聴覚補助の実現に向け，所望の源信号を高い主観品質で強調するための音源強調技術の実現を目指す．3 章の実験結果からも明らかなように，源信号の推定結果の二乗誤差の大きさと人間が知覚する音質の劣化の大きさは必ずしも比例しない [49]．そのため二乗誤差最小化などでニューラルネットワークを学習して源信号を推定しても，主観品質を最大化する源信号を推定することはできなかった．聴覚フィルタなどを利用し，PESQ (perceptual evaluation of speech quality) [50] や PEASS (perceptual evaluation methods for audio source separation) [52] などの主観評価値と相関の高い音質評価値を計算することはできるが，評価値からそれを最大化するラベルデータは一意に定めることができない．ゆえに教師あり学習とは別の枠組みでニューラルネットワークを学習する必要があると考えた．本章では，強化学習のフレームワークを応用した，源信号を推定するための新たな目的関数とその最適化法を提案する．

2.4.3 節で説明してきたように，強化学習は明示的なラベルデータを用いずにニューラルネットワークを学習できる枠組みである [98, 99, 100, 101, 102]．強化学習では，明示

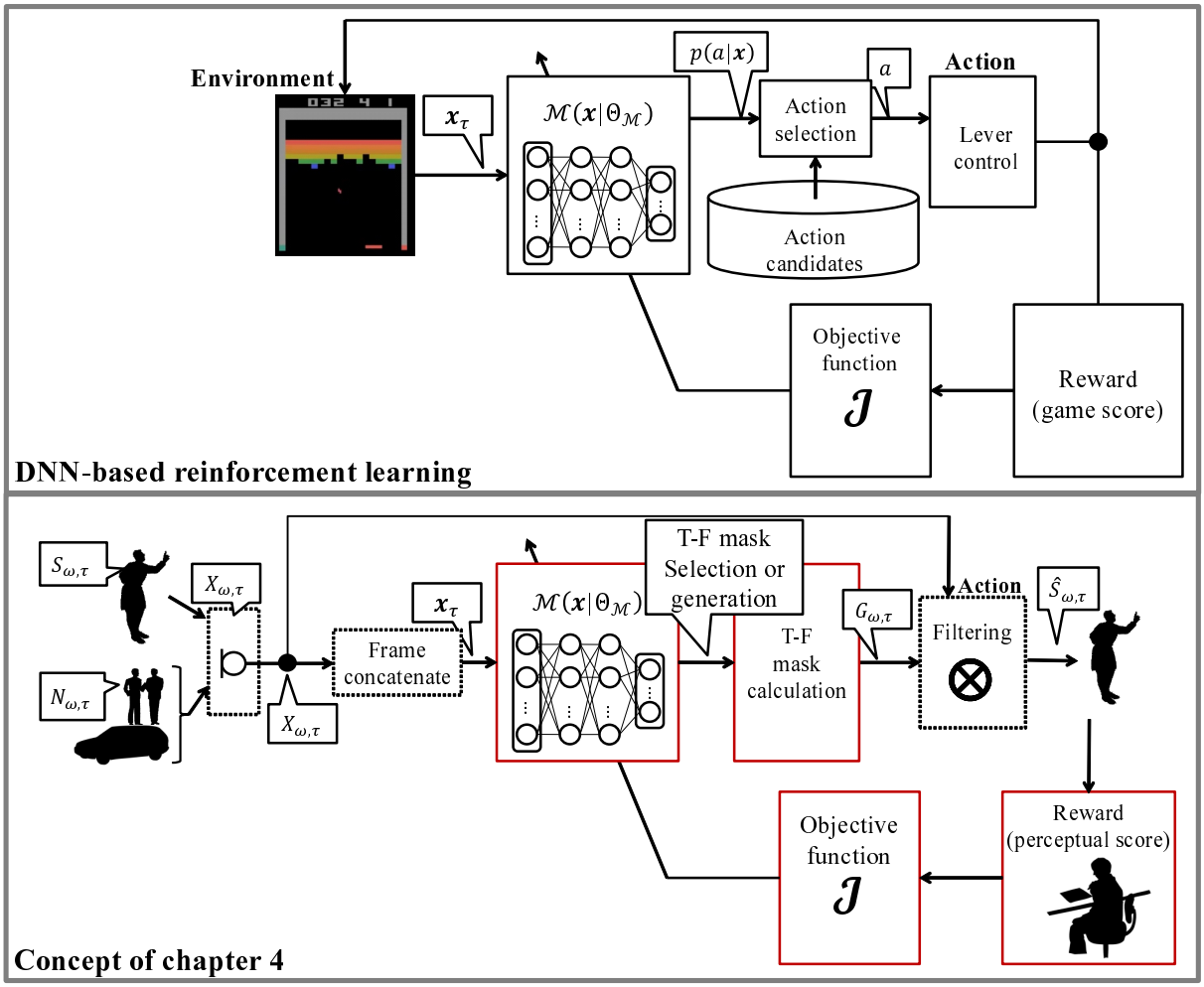


図 4.1: ゲーム操作のための強化学習（上）と 強化学習のフレームワークを利用した音源強調（下）．環境を観測信号，行動を時間周波数マスク処理，報酬関数を聴感評点に置き換え，方策関数を学習する．

的なラベルデータを与える代わりに，採用した行動の有効性を返す “報酬関数 (reward function)” を定義する．そして現在の環境や自身の行動方針（方策関数）に従い行動し，得られた報酬をもとにより高い報酬を得られるよう，試行錯誤をしながら方策関数を学習していく[1]．近年成功を収めているゲーム操作のための深層学習を用いた強化学習では，ゲームの画面（環境）$\boldsymbol{x}_\tau$ を入力とし，有限個のゲームのレバー操作（行動）の候補 $\mathcal{A} = \{a_1, a_2, ..., a_A\}$ のうち a 番目の行動が最良の行動である確率 $p(a|\boldsymbol{x}_\tau)$ をニューラルネットワークで出力する（図 4.1 上）．そして，ゲームスコアやゲームの勝敗を報酬関数として強化学習を行うことで，明示的なラベルデータを用いずにニューラルネットワー

[1]価値反復に基づく強化学習において方策は，報酬の和の期待値である行動価値関数（Q 関数）を通して表現する．一方，方策勾配に基づく強化学習において方策は，観測 $\boldsymbol{x}$ を得た下で行動 a が最適である確率的方策 $p(a|\boldsymbol{x})$ で表現する．本章では，行動を決定するための方策を求める関数をを総称して「方策関数」と呼ぶ．

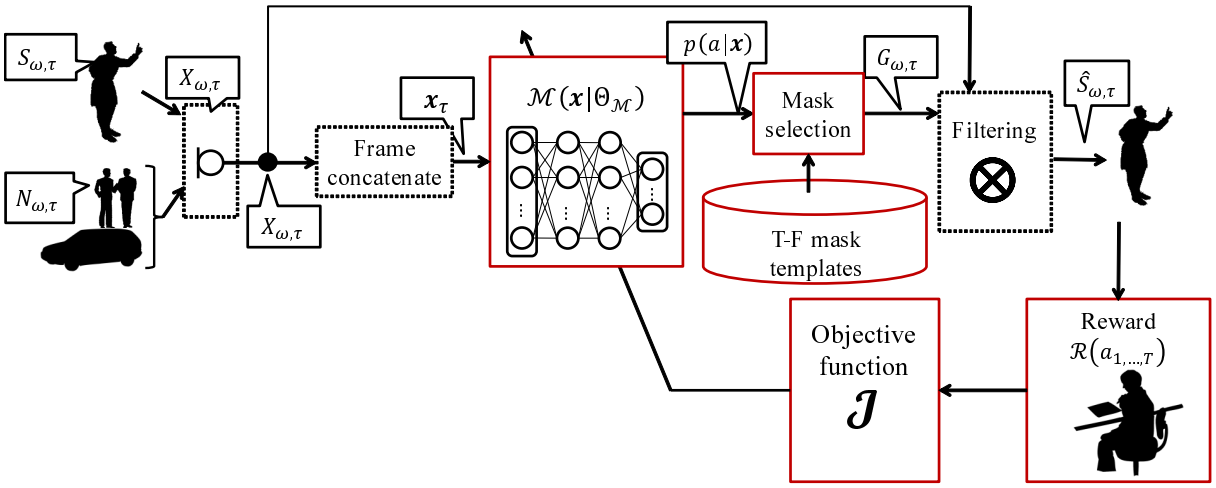


図 4.2: 時間周波数マスクの選択に基づく音源強調.

クを学習できるようになり，人間の能力を超えるゲーム操作が可能になることがわかっている [99, 100, 101].

そこで，音質評価に関する適切な報酬関数さえ設計できれば，音源強調のためのニューラルネットワークも，強化学習のフレームワークで学習できると考えた（図 4.1 下）．本章では，時間周波数マスクをニューラルネットワークで表現される方策関数に基づき計算する．そして，環境を観測信号，行動を時間周波数マスク処理，報酬関数を，報酬を主観評価値と相関の高い音質評価値（聴感評点）[50, 51, 52] に置き換え，聴感評点を最大化するようにニューラルネットワークを学習するための目的関数を提案する．

まず 4.2 節では，従来の行動の選択に基づく強化学習の枠組みを利用した音源強調を提案する．観測信号 $\boldsymbol{x}_\tau$ を入力とし，有限個の各時間周波数マスクテンプレートが聴感評点を最大化する確率をニューラルネットワークで推定する．そしてその確率に基づき時間周波数マスクを “選択” する．この枠組みの中で，聴感評点を最大化するようにニューラルネットワークを学習するための目的関数を提案する．次に 4.3 節では，従来の強化学習の枠組みを拡張し，行動の生成に基づく強化学習の枠組みを利用した音源強調を提案する．観測信号 $\boldsymbol{x}_\tau$ を入力とし，聴感評点を最大化する時間周波数マスクの確率密度関数を出力するニューラルネットワークを用いて，時間周波数マスクを直接 “生成” する．この枠組みの中で，聴感評点を最大化するようにニューラルネットワークを学習するための目的関数を提案する．なお本章では，ラベルデータを一意に定めることのできない源信号の推定について考えるため，音源強調の学習データである源信号と雑音のデータは十分に用意できるものとして議論を進める．

4.2 時間周波数マスクの選択に基づく音源強調

図 4.2 に，提案する時間周波数マスクの選択に基づく音源強調の枠組みを示す．一般的な強化学習の枠組みでは，事前に有限個の “行動” $\mathcal{A} = \{a_1, ..., a_A\}$ を定義する必要がある [98]．時間周波数マスキングを有限個の行動として表現するために，本節では時間周波数マスクテンプレート $\boldsymbol{\mathcal{G}}_{1,...,A}$ を事前に用意し，行動を a 番目の時間周波数マスク $\boldsymbol{\mathcal{G}}_a = (G_{1,a}, ..., G_{\Omega,a})^\top$ の選択として定義する．時間周波数マスクは，観測信号 $\boldsymbol{x}_\tau$ と時間周波数マスクの選択方針（方策関数）である $\mathcal{M}(\boldsymbol{x}_\tau, a|\Theta_\mathcal{M})$ に従って選択される．

$$\hat{\boldsymbol{G}}_\tau \leftarrow \boldsymbol{\mathcal{G}}_{a_\tau} \tag{4.1}$$

$$a_\tau \leftarrow \arg\max_{a \in \mathcal{A}} \mathcal{M}(\boldsymbol{x}_\tau, a|\Theta_\mathcal{M}) \tag{4.2}$$

その後選択された時間周波数マスクを用いて源信号を強調し，音質評価関数 $\mathcal{R}$ を用いて主観評価値と相関の高い音質評価値（聴感評点）を計算する．そして，聴感評点を最大化するように方策関数を更新する．以降では表記の簡単のために，式 (4.1)(4.2) で得られた時間周波数マスクを用いて音源強調を行うことを “DNN-RL” と呼ぶ．

式 (4.1)(4.2) では，$\mathcal{M}(\boldsymbol{x}_\tau, a|\Theta_\mathcal{M})$ は聴感評点を最大化する時間周波数マスクテンプレート $\boldsymbol{\mathcal{G}}_{a_\tau}$ を識別しているとみなすこともできる．そこで，$\mathcal{M}(\boldsymbol{x}_\tau, a|\Theta_\mathcal{M})$ は，観測信号 $\boldsymbol{x}_\tau$ を得た下での a 番目の時間周波数マスクテンプレートが聴感評点を最大化する事後確率を返しているとみなす．

$$p(a|\boldsymbol{x}) = \mathcal{M}(\boldsymbol{x}, a|\Theta_\mathcal{M}) \tag{4.3}$$

式 (4.3) を利用すると，式 (4.2) は聴感評点を最大化する事後確率を最大化する行動を選択している，MAP (maximum a posteriori) 推定とみなすことができる．本節では事後確率を正確に求める方法として，出力層の活性化関数を softmax 関数としたニューラルネットワークを利用する．全結合多層ニューラルネットワーク（DNN）を利用した場合，$\mathcal{M}(\boldsymbol{x}_\tau, a|\Theta_\mathcal{M})$ は以下のようになる．

$$\mathcal{M}(\boldsymbol{x}_\tau, a|\Theta_\mathcal{M}) = \frac{\exp(u^{(L)}_{\tau,a})}{\sum_{i=1}^{A} \exp(u^{(L)}_{\tau,i})} \tag{4.4}$$

$$\boldsymbol{z}^{(l)}_\tau = \sigma_{\Theta_\mathcal{M}}\left\{\boldsymbol{u}^{(l)}_\tau\right\} \tag{4.5}$$

$$\boldsymbol{u}^{(l)}_\tau = \mathbf{W}^{(l)}\boldsymbol{z}^{(l-1)}_\tau + \mathbf{b}^{(l)} \tag{4.6}$$

ここで $\boldsymbol{z}^{(1)}_\tau = \boldsymbol{x}_\tau$，$\boldsymbol{u}^{(L)} = (u^{(L)}_{\tau,1}, ..., u^{(L)}_{\tau,A})^\top$ であり，L，$\mathbf{W}^{(l)}$，$\mathbf{b}^{(l)}$，$\Theta_\mathcal{M}$，$\sigma_{\Theta_\mathcal{M}}$ の定義は 1 章で説明したものと同じである．つまり $\Theta_\mathcal{M} = \{\mathbf{W}^{(l)}, \mathbf{b}^{(l)}\}$ であり，聴感評点を最大化するようにこれらのパラメータを学習する．次節では，方策学習に用いる報酬関数および目的関数の設計と学習手順を説明する．

4.2.1 方策学習のための報酬関数と目的関数の設計

代表的な聴感評点である PESQ や PEASS は，音源強調の性能だけでなく観測信号の SNR や雑音の種類によっても値が変動してしまう．そのため，聴感評点をそのまま報酬として用いることは困難である．そこで本研究では DNN-RL の出力音の音質評価と，2 章で説明した MMSE 基準で学習したニューラルネットワークから求めた時間周波数マスクを用いて音源強調した出力を比較することで報酬を計算する．以降表記の簡単のために，MMSE 基準で学習したニューラルネットワークから求めた時間周波数マスクを用いた音源強調を “DNN-MMSE” と呼ぶ．いま，DNN-RL で得られた強調信号から求めた聴感評点 $\mathcal{Z}^{\mathrm{RL}}$，DNN-MMSE で得られた強調信号から求めた聴感評点 $\mathcal{Z}^{\mathrm{MMSE}}$ とする．そしてこの 2 つの聴感評点を比較した報酬を以下のように求める．

$$\mathcal{R}(a_{1,\dots,T}) = \tanh\left\{\alpha\left(\mathcal{Z}^{\mathrm{RL}} - \mathcal{Z}^{\mathrm{MMSE}}\right)\right\}, \tag{4.7}$$

ここで $\alpha > 0$ は報酬のスケーリング係数であり，tanh は報酬のクリッピングのための双曲線正接関数である．この比較報酬 $\mathcal{R}(a_{1,\dots,T})$ は，ゲームの勝敗から着想を得た値である [99, 100, 101]．もし $\mathcal{Z}^{\mathrm{RL}}$ が $\mathcal{Z}^{\mathrm{MMSE}}$ より大きいということは，MMSE に基づく音源強調よりも強化学習で学習した現在の方策の方が音質が高いということであり，$\mathcal{Z}^{\mathrm{RL}}$ を求めるために行った音源強調（行動）は正しかったと判断することができる（$\mathcal{R}(a_{1,\dots,T}) > 0$）．一方，$\mathcal{Z}^{\mathrm{RL}}$ が $\mathcal{Z}^{\mathrm{MMSE}}$ より小さいということは，MMSE に基づく音源強調よりも，強化学習で学習した現在の方策の方が音質が低いということであり，$\mathcal{Z}^{\mathrm{RL}}$ を求めるために行った音源強調（行動）は誤っていたと判断することができる（$\mathcal{R}(a_{1,\dots,T}) < 0$）．このように提案法では，DNN-MMSE という DNN-RL と敵対する音源強調を設けることで，音源強調の性能以外からの聴感評点への影響を低減し，また MMSE に基づく音源強調よりも高い音質で音源強調できるよう DNN を学習することを狙う．

また，時間周波数マスクは時間的に変化するため，報酬も数十 ms 単位の各時間フレーム τ ごとに変化すべきである．しかし，PESQ をはじめとする多くの聴感評点は，8 秒程度の一発話全体の音源強調結果から 1 つの値しか求まらない．報酬を時間変化させるために，以下で求める時変係数 E_τ を用いて比較報酬 $\mathcal{R}(a_{1,\dots,T})$ を時間変化させ，各時間フレームごとの報酬 r_τ を計算する．

$$r_\tau = \begin{cases} (1 - E_\tau)\mathcal{R}(a_{1,\dots,T}) & (\mathcal{R}(a_{1,\dots,T}) > 0) \\ E_\tau \mathcal{R}(a_{1,\dots,T}) & (\text{other}) \end{cases}, \tag{4.8}$$

$$E_\tau = \frac{\tilde{E}_\tau}{\max_{\tau \in T}(\tilde{E}_\tau)} \tag{4.9}$$

$$\tilde{E}_\tau = \sum_{\omega=1}^{\Omega} \left|\ln|\mathcal{G}_{\omega,a_\tau} X_{\omega,\tau}| - \ln|S_{\omega,\tau}|\right|^2. \tag{4.10}$$

式 (4.9)(4.10) から見て取れるように，時変係数 $0 < E_k < 1$ は DNN-RL の出力と所望の源信号の正規化二乗誤差である．

そして，報酬 r_τ を用いて方策関数の目標値 $\tilde{\mathcal{Q}}(\boldsymbol{x}_\tau, a_\tau)$ を計算する．

$$\tilde{\mathcal{Q}}(\boldsymbol{x}_\tau, a_\tau) = \begin{cases} r_\tau + \max_{a\in\mathcal{A}} \mathcal{M}(\boldsymbol{x}_\tau, a_\tau|\Theta_\mathcal{M}) & (\mathcal{R}(a_{1,\ldots,T}) > 0) \\ \mathcal{M}(\boldsymbol{x}_\tau, a_\tau|\Theta_\mathcal{M}) & (\text{other}) \end{cases} \tag{4.11}$$

ここで，もし $a_\tau \neq a_k^{\mathrm{MMSE}}$ の場合，$\tilde{\mathcal{Q}}(\boldsymbol{x}_k, a_k^{\mathrm{MMSE}})$ は以下のように再計算する．

$$\tilde{\mathcal{Q}}(\boldsymbol{x}_\tau, a_\tau^{\mathrm{MMSE}}) = \begin{cases} \mathcal{M}(\boldsymbol{x}_\tau, a_\tau^{\mathrm{MMSE}}|\Theta_\mathcal{M}) & (\mathcal{R}(a_{1,\ldots,T}) > 0) \\ \mathcal{M}(\boldsymbol{x}_\tau, a_\tau^{\mathrm{MMSE}}|\Theta_\mathcal{M}) - r_\tau & (\text{other}) \end{cases}, \tag{4.12}$$

ここで a_τ^{MMSE} は MMSE 基準で選択されるマスクであり，以下のように求める．

$$a_\tau^{\mathrm{MMSE}} \leftarrow \arg\min_{a\in\mathcal{A}} \sum_{\omega=1}^{\Omega} \left| |S_{\omega,\tau}| - |\mathcal{G}_{\omega,a} X_{\omega,\tau}| \right|^2. \tag{4.13}$$

ただし $\mathcal{M}(\boldsymbol{x}_\tau, a_\tau|\Theta_\mathcal{M})$ はソフトマックス関数の出力である事後確率を表しているため，方策関数の出力の目標値 $\tilde{\mathcal{Q}}(\boldsymbol{x}_\tau, a)$ は $\tilde{\mathcal{Q}}(\boldsymbol{x}_\tau, a) \geq 0$ を満たすように切り上げを行った後に $\sum_{i=1}^{A} \tilde{\mathcal{Q}}(\boldsymbol{x}_\tau, i) = 1$ となるように正規化される．また式 (4.12) による報酬の計算は，一般的な方策学習では用いられない計算である．この式の意図は，もし DNN-MMSE の方が音質が高い場合（$\mathcal{R}(a_{1,\ldots,T}) < 0$），現在の方策に基づく行動 a_τ よりも MMSE ベースの行動 a_τ^{MMSE} の方が音質が高くなると考えられる点にある．そこで，負の報酬 r_τ を $\mathcal{Q}(\boldsymbol{x}_\tau, a_\tau^{\mathrm{MMSE}})$ から減算することで，MMSE ベースの行動 a_τ^{MMSE} の行動確率を高めることを目的としている．

最後にニューラルネットワークの出力 $\mathcal{M}(\boldsymbol{x}_\tau, a_\tau|\Theta_\mathcal{M})$ が方策関数の目標値 $\tilde{\mathcal{Q}}(\boldsymbol{x}_\tau, a_\tau)$ となるようにニューラルネットワークのパラメータを更新する．ゆえに時間周波数マスクの選択に基づく音源強調の目的関数は以下のようになる．

$$\Theta_\mathcal{M} \leftarrow \arg\max_{\Theta_\mathcal{M}} -\frac{1}{T} \sum_{\tau=1}^{T} \sum_{i=1}^{A} \left| \tilde{\mathcal{Q}}(\boldsymbol{x}_\tau, i) - \mathcal{M}(\boldsymbol{x}_\tau, i|\Theta_\mathcal{M}) \right|^2 \tag{4.14}$$

4.2.2 方策関数の学習

本節では，提案法による方策関数の学習手順を述べる．提案法は，初期化フェーズと学習フェーズの 2 段階処理で学習を行う．また提案法で用いる学習データは所望の源信号 $\{S_{\omega,\tau}|\omega = 1, ..., \Omega, \tau = 1, ..., T\}$ と雑音 $\{N_{\omega,\tau}|\omega = 1, ..., \Omega, \tau = 1, ..., T\}$ [2]である．図 4.3 に提案法の学習手順を示す．以下では，この図に沿って学習手順を説明する．

[2]表記の簡単のために $N_{\omega,\tau} = \sum_{k=1}^{K} N_{k,\omega,\tau}$ とした．

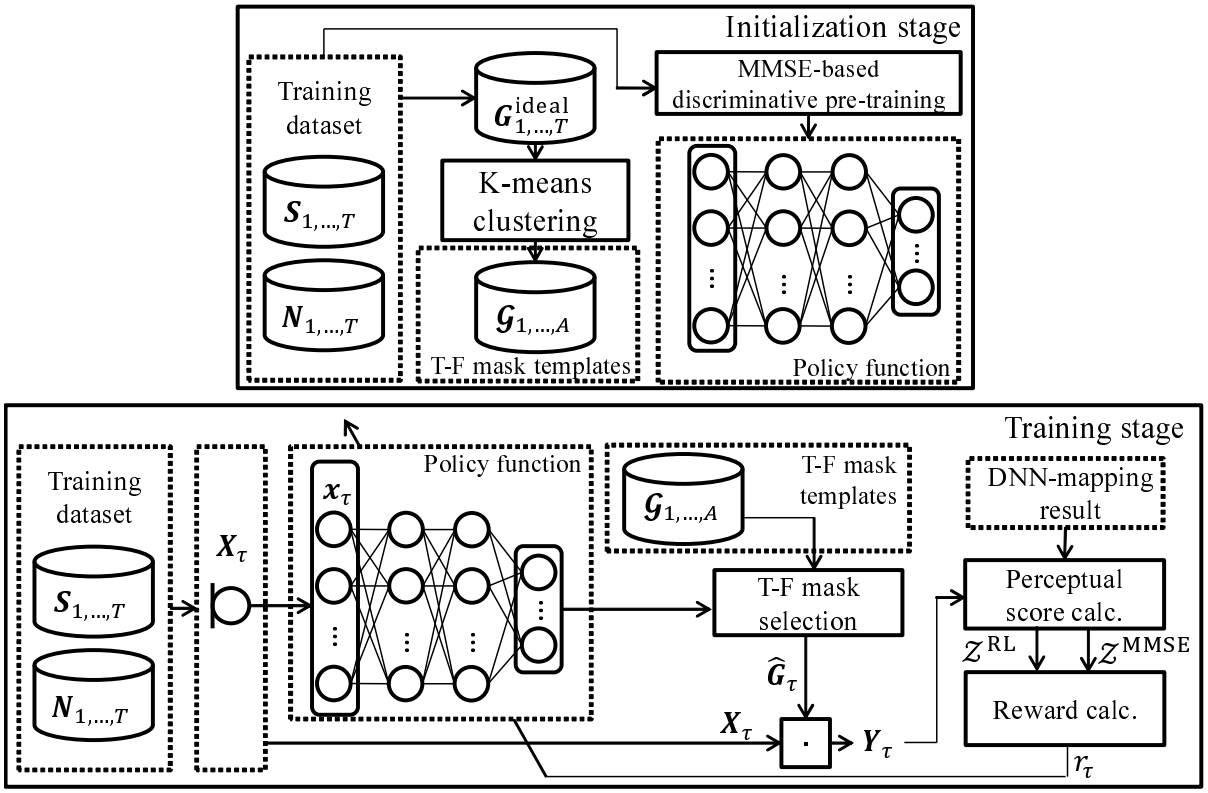


図 4.3: 時間周波数マスクの選択に基づく音源強調の学習手順．(上) 初期化フェーズ，(下) 学習フェーズ．

初期化フェーズでは，時間周波数マスクテンプレート $\boldsymbol{\mathcal{G}}_{1,...,A}$ の準備と DNN の事前学習を行う．まず，学習データ $S_{\omega,\tau}, N_{\omega,\tau}$ から，以下の式で理想ウィナーフィルタを計算する．

$$G^{\text{ideal}}_{\omega,\tau} = \frac{|S_{\omega,\tau}|^2}{|S_{\omega,\tau}|^2 + |N_{\omega,\tau}|^2} \tag{4.15}$$

そして，$\boldsymbol{G}^{\text{ideal}}_{\tau} = (G^{\text{ideal}}_{1,\tau}, ..., G^{\text{ideal}}_{\Omega,\tau})^\top$ を k-means アルゴリズムでクラスタリングし，そのクラスタ中心を時間周波数マスクテンプレート $\boldsymbol{\mathcal{G}}_{1,...,A}$ とする．次いで，DNN パラメータ $\Theta_{\mathcal{M}}$ を MMSE ベースの時間周波数マスクテンプレートを正解とした識別基準で事前学習する．つまり，$\boldsymbol{x}_\tau$ を入力すると a^{MMSE}_τ を選択するように DNN を初期化する．なお，discriminative pre-training [86] を利用する．また最終層の DNN パラメータ $\mathbf{W}^{(L)}, \mathbf{b}^{(L)}$ は，強化学習の時間周波数マスク選択のランダム性を高め，収束を早めるために乱数初期化する．

学習フェーズでは，報酬を最大化するように DNN パラメータ $\Theta_{\mathcal{M}}$ を学習する．まず源信号の学習データから 1 発話をランダムに選択し，雑音データセットからそれと同じ長さのノイズをランダムに取得し，源信号を雑音をランダムな SNR で重畳することで観測信号を確率的に生成する．次いで，式 (4.1)(4.2) および ϵ-greedy アルゴリズムを用いて時間周波数マスクを選択し，出力信号を得る．ここで ϵ-greedy アルゴリズムは，確率

ϵ でランダムに行動を選択することで，学習の収束を早めるアルゴリズムである．その後報酬と方策関数の目標値を計算し，式 (4.14) の目的関数を最大化するように DNN パラメータを更新する．なお提案法では，式 (4.14) の実行のために，RMSProp アルゴリズムを用いる [140].

4.2.3 評価実験

(a) 実験条件

強化学習に基づく時間周波数マスク選択の有効性を確かめるために，客観評価試験および主観評価試験を行った．聴感評点には PESQ と PEASS より OPS (overall perceptual score) を用いた．所望の源信号は日本語の発話音声とし，学習データには ATR 音声データベース [141] から男性 11 名，女性 11 名による全 3316 発話を利用した．雑音は CHiME-3 の雑音データセット [142] より，“cafes”，“street junctions”，“public transport (buses)”，“pedestrian areas” の 4 種類を用いた．観測信号は，源信号と雑音を SNR が 0，3，6 dB のどれかとなるようにランダムに重畳することで生成した．これらのデータは 16kHz でサンプリングした．

DNN-RL のニューラルネットワークには全結合多層ニューラルネットワーク（DNN）を用いた．DNN の自由度を下げ過適合を防ぐために，構造は隠れ層 2 層，隠れユニット数 64 とした．また時間周波数マスクのテンプレート数は $A = 32$ とした．報酬係数 α は，PESQ が 20.0，PEASS が 1.0 とした．ϵ-greedy アルゴリズムの確率 ϵ は 0.01 とした．学習は，各聴感評点で 50,000 エピソード実行した．ここでエピソードとは，1 発話の観測信号をシミュレートし音源強調を実行して，報酬を計算したのちに DNN を更新する，1 発話毎の学習の一連の流れを意味する．DNN-MMSE のニューラルネットワークにも全結合多層ニューラルネットワーク（DNN）を用いた．DNN-RL を小さなネットワーク構造としたため，DNN-MMSE のネットワーク構造にも隠れ層 2 層，隠れユニット数 128 の小さな DNN を用いた．DNN-MMSE では時間周波数マスク推定の安定性を高めるために，従来研究にならって，時間周波数マスクを直接推定するのではなく所望の源信号の対数振幅スペクトル $\ln|S_{\omega,k}|$ を推定した [47]．また過適合を防ぐためにドロップアウト [47] を利用し，初期化には discriminative pre-training [86] を利用した．各手法において，入力 $\boldsymbol{x}$ のコンテキスト窓の大きさは $P_b = P_f = 5$ とした．DNN の入出力の次元数を抑えるために，$\boldsymbol{X}_\tau$ と $\boldsymbol{G}_\tau$ は $\mathsf{B} = 64$ のメルフィルタバンクで圧縮し，時間周波数マスク設計の際にスプライン補間で線形周波数に補間した．短時間フーリエ変換のフレームサイズは 512 サンプルとし，シフト幅は 256 サンプルとした．その他詳細な実験条件を表 4.1 にまとめた．

表 4.1: 実験条件.

Parameters for signal processing	
Sampling rate	16.0 kHz
FFT length	512 pts
FFT shift length	256 pts
# of mel-filterbanks B	64
context window size P_f, P_b	5, 5
Training SNR (dB)	0, 3, 6
Training parameters for DNN-RL	
# of hidden layers for DNNs	2
# of hidden units for DNNs	64
activation function	sigmoid
# of T-F mask templates A	32
ϵ-greedy parameter ϵ	0.01
reward coefficient α (PESQ, PEASS)	20.0, 1.0
Training parameters for DNN-MMSE	
# of hidden layers for DNNs	2
# of hidden units for DNNs	128
activation function	sigmoid
Dropout probability (input layer)	0.2
Dropout probability (hidden layer)	0.5

(b) 客観評価実験

提案法によって聴感評点が向上するかを確認するために，エピソードの増加と聴感評点の変化を評価した．もし提案法により方策関数が正しく学習できているならば，エピソードが増加するに従い聴感評点も向上するはずである．実験には，源信号の学習データとは異なる日本語の男性 50 発話および女性 50 発話を利用した．雑音は学習データと同じデータを用い，SNR は 0 dB と 6 dB でそれぞれ評価した．図 4.4 に実験結果を箱ひげ図で評価したものを示す．PESQ および PEASS-OPS で学習した両方の場合で，エピソードの増加に応じて聴感評点が向上していることがわかる．また全ての SNR および聴感評点で，最終的に DNN-MMSE よりも評点が向上しており，聴感評点: PESQ，SNR: 6 dB，および聴感評点: PEASS-OPS，SNR: 0 dB の条件では，評価結果の上下 5% の外れ値を除外した対応のない片側 t 検定において有意差が認められた（有意水準 5%）.

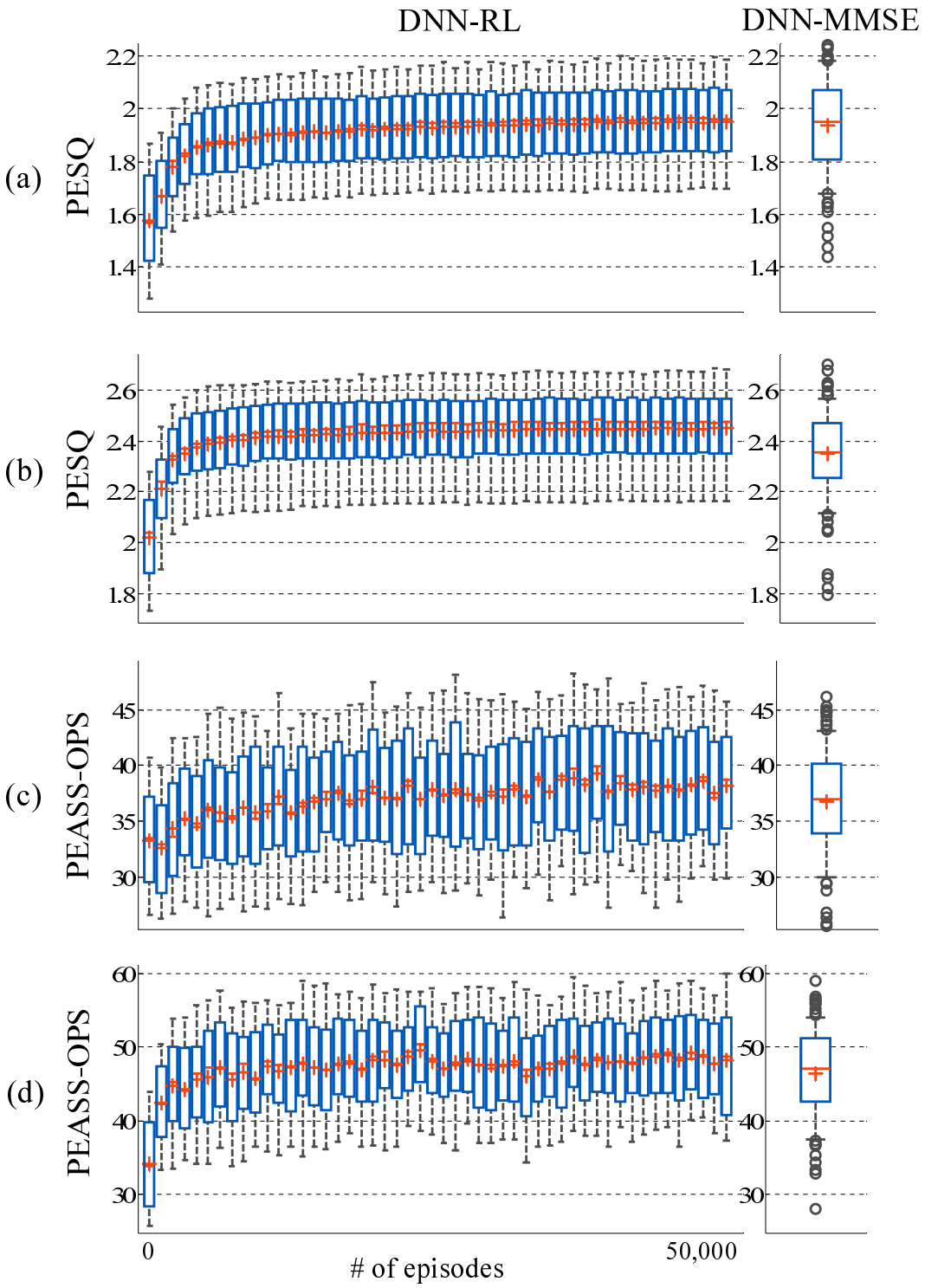


図 4.4: 聴感評点とエピソード数の関係性．x 軸がエピソード数，y 軸が聴感評点の値を表し，左図が DNN-RL，右図が DNN-MMSE の結果を表す．(a) 聴感評点: PESQ，SNR: 0 dB，(b) 聴感評点: PESQ，SNR: 6 dB，(c) 聴感評点: PEASS-OPS，SNR: 0 dB，(d) 聴感評点: PEASS-OPS，SNR: 6 dB．

表 4.2: MOS の一覧.

Method	PESQ			PEASS-OPS			DNN-MMSE
# of episodes	500	5,000	50,000	500	5,000	50,000	-
MOS	2.7	2.9	**3.0**	2.5	2.7	**3.0**	2.8

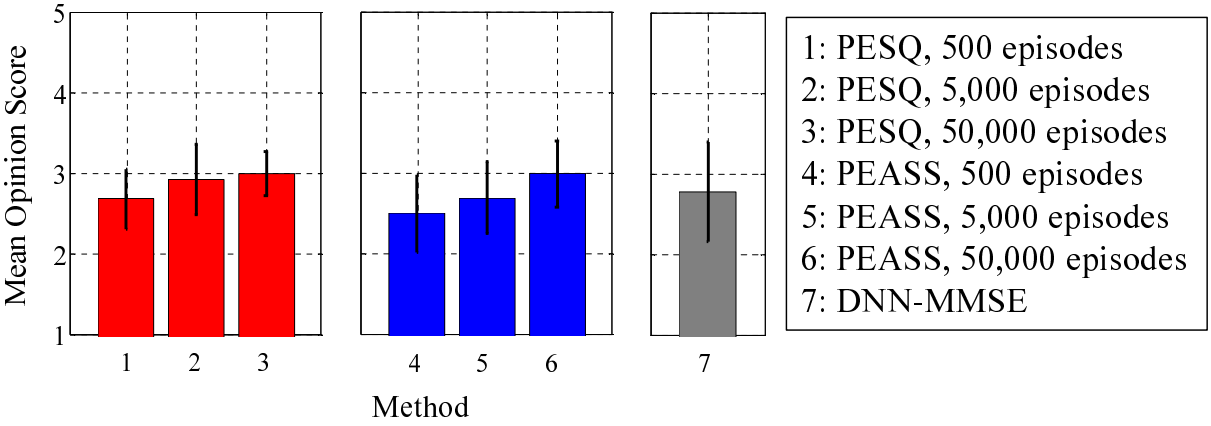


図 4.5: 主観評価試験の結果．エラーバーは標準偏差を表す.

(c) 主観評価実験

提案法により音源強調の出力の主観品質が向上するかを調べるために，主観評価試験を行った．主観評価には，5 段階の平均オピニオン評点 (MOS: mean-opinion-score) を利用した（1 – 非常に悪い，3 – 普通，5 – 非常に良い）．7 名の正常な聴力を持つ被験者が，DNN-RL と DNN-MMSE の出力音を評価した．エピソードの増加による音質の向上を確かめるために，DNN-RL は 500，5,000，50,000 エピソード目の，3 つのエピソードの学習結果を評価した．源信号と雑音の SNR は 3 dB になるように設定した．また外れ値を取り除くために，評価結果の上下 5% を除外して MOS を計算した．

図 4.5 および表 4.2 に主観評価試験の結果を示す．DNN-RL の MOS はエピソードの増加とともに向上していることがわかり，また両方の聴感評点において，500 エピソードと 50,000 エピソードの間には，対応のない片側 t 検定で有意差が認められた（p 値 $= 0.05$）．また，両方の聴感評点において 50,000 エピソードの MOS は DNN-MMSE の MOS を上回っており，対応のない片側 t 検定で有意差が認められた（有意水準 5%）．これらの結果から，提案法する目的関数で音源強調のためのニューラルネットワークを学習することで，明示的にラベルデータを設計できない主観品質も向上するといえる．

4.2.4 時間周波数マスクの選択に基づく音源強調のまとめ

本節では，強化学習に基づく音源強調関数の学習方法の単純な実装方法として，時間周波数マスクの選択に基づく音源強調を提案した．時間周波数マスキングを有限個の行動として表現するために，時間周波数マスクテンプレートを利用し，DNN を時間周波数マスクテンプレートを MAP 推定により選択する関数として利用した．聴感評点をそのまま報酬関数とするのではなく，DNN-RL の出力音の音質評価と DNN-MMSE の出力音の音質評価の比較報酬を利用して報酬関数と目的関数を設計することで，MMSE を目的関数として DNN を学習する音源強調法よりも高い主観品質で音源強調ができることがわかった．

本節では DNN の自由度を下げるために，時間周波数マスクテンプレートの数を 32 個に制限した．しかし本来，時間周波数マスクは所望の源信号と雑音の音圧比により決定する連続変数であり，有限個のテンプレートで表現することは妥当ではない．MMSE に基づく目的関数を利用した従来研究のように，連続変化する時間周波数マスクを推定するためには，従来の強化学習のような離散的な行動の “選択問題” ではなく，連続的な行動の “生成問題” として問題を再定義する必要がある．次節では，強化学習において連続的な行動のための方策関数を勾配法で学習する “方策勾配法”[139] を利用し，時間周波数マスクを “生成” するための目的関数と学習法を提案する．

4.3 時間周波数マスクの生成に基づく音源強調

前節では，事前に定義した有限個の時間周波数マスクテンプレートから，聴感評点を最大化する時間周波数マスクを選択する関数として DNN を利用した．本節では，時間周波数マスク推定の柔軟性を高めるために，聴感評点を最大化する時間周波数マスク処理の出力音の従う連続確率分布のパラメータを推定する関数として DNN を利用する．聴感評点の多くは DNN パラメータに関する勾配が解析的に求まらないブラックボックス関数であるため，聴感評点を利用して設計した目的関数もまた，DNN パラメータに関する勾配が解析的に求まらない．本節では目的関数の勾配を求めるために，ブラックボックスな目的関数の勾配をサンプリング法に基づき数値的に推定する “方策勾配法”[139] を利用する．すなわち，DNN で推定した時間周波数マスク処理の出力音の従う連続確率分布から出力音を複数サンプリングし，その中から聴感評点を向上させた出力音の生成確率を高めるように，DNN パラメータの勾配を求める．また，サンプリング法に基づく勾配計算の安定性を高めるために，二つの追加処理を行う．

- サンプリングした出力音が，時間周波数マスク処理音としての制約を満たすためのサンプリングアルゴリズム（4.3.3 節）．

- 勾配の分散を低減させるための聴感評点の正規化処理と，時間周波数マスクのクリッピング処理（4.3.4 節）．

以降では，まず 4.3.1 節で，DNN の出力を連続確率密度関数のパラメータとみなして音源強調を行う枠組みの先行研究として，最尤推定法に基づく DNN 音源強調関数の学習を概説する．次いで 4.3.2 節は，提案する方策勾配法に基づく DNN 音源強調関数の学習を説明する．その後，4.3.3 節および 4.3.4 節で，サンプリング法に基づく勾配計算の安定性を高めるための追加処理を説明し，4.3.5 節で提案法の学習アルゴリズムの詳細を述べる．

4.3.1 最尤推定法に基づく DNN 音源強調関数の学習

最尤推定法に基づく DNN 音源強調関数の学習では，DNN は観測信号を得た下での目的音の条件付き確率密度関数 $p(\boldsymbol{S}_\tau|\boldsymbol{X}_\tau, \Theta_{\mathcal{M}})$ のパラメータを推定するための関数として用いられる．まず，目的音 $S_{\omega,\tau}$ と観測信号 $X_{\omega,\tau}$ を，各時間フレーム毎に周波数方向にまとめてベクトル化したものを以下のように定義する．

$$\boldsymbol{S}_\tau := (S_{1,\tau}, ..., S_{\Omega,\tau})^\top \tag{4.16}$$

$$\boldsymbol{X}_\tau := (X_{1,\tau}, ..., X_{\Omega,\tau})^\top \tag{4.17}$$

ここで $\top$ は転置を表す．DNN パラメータ $\Theta_{\mathcal{M}}$ は，以下の対数尤度の期待値を最大化するように学習される．

$$\Theta_{\mathcal{M}} \leftarrow \arg\max_{\Theta_{\mathcal{M}}} \mathcal{J}^{\mathrm{ML}}(\Theta_{\mathcal{M}}) \tag{4.18}$$

ここで最尤推定法の目的関数 $\mathcal{J}^{\mathrm{ML}}(\Theta_{\mathcal{M}})$ は以下である．

$$\mathcal{J}^{\mathrm{ML}}(\Theta_{\mathcal{M}}) = \mathbb{E}_{\boldsymbol{S},\boldsymbol{X}}\left[\ln p(\boldsymbol{S}|\boldsymbol{X}, \Theta_{\mathcal{M}})\right] \tag{4.19}$$

ただし $\mathbb{E}_x[\cdot]$ は x に関する期待値演算を表す．ここで式 (4.19) 内の期待値は解析的に求めることがでいないため，$S_{\omega,\tau}$ と $X_{\omega,\tau}$ の学習データに関する平均で近似される．

$$\mathcal{J}^{\mathrm{ML}}(\Theta_{\mathcal{M}}) \approx \frac{1}{T}\sum_{\tau=1}^{T} \ln p(\boldsymbol{S}_\tau|\boldsymbol{X}_\tau, \Theta_{\mathcal{M}}). \tag{4.20}$$

すると，$p(\boldsymbol{S}_\tau|\boldsymbol{X}_\tau, \Theta_{\mathcal{M}})$ が $\Theta_{\mathcal{M}}$ に関する勾配が解析的に求まる関数の合成関数で設計されているならば，目的関数 $\mathcal{J}^{\mathrm{ML}}(\Theta_{\mathcal{M}})$ の $\Theta_{\mathcal{M}}$ に関する勾配は誤差逆伝搬法 [82] で計算できる．

$$\nabla_{\Theta_{\mathcal{M}}}\mathcal{J}^{\mathrm{ML}}(\Theta_{\mathcal{M}}) \approx \frac{1}{T}\sum_{\tau=1}^{T} \nabla_{\Theta_{\mathcal{M}}} \ln p(\boldsymbol{S}_\tau|\boldsymbol{X}_\tau, \Theta_{\mathcal{M}}), \tag{4.21}$$

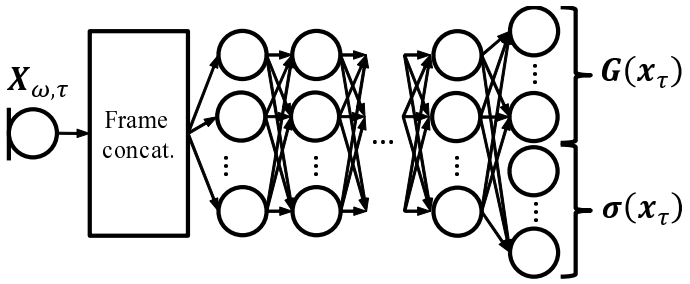


図 4.6: 複素ガウス分布として $p(\boldsymbol{S}_\tau|\boldsymbol{X}_\tau,\Theta_{\mathcal{M}})$ を設計するための DNN$\mathcal{M}(\boldsymbol{x}_\tau|\Theta_{\mathcal{M}})$ の例.

ここで ∇_x は x に関する偏微分を表す.

式 (4.21) を解析的に計算可能な $p(\boldsymbol{S}_\tau|\boldsymbol{X}_\tau,\Theta_{\mathcal{M}})$ のモデル化法として，$S_{\omega,\tau}$ の推定誤差分布を平均 0，分散 $\sigma^2_{\omega,\tau}$ の複素ガウス分布でモデル化する方法がある．これは，目的音 $S_{\omega,\tau}$ と出力音 $\hat{S}_{\omega,\tau}$ の二乗誤差を複素平面上で最小化する時間周波数マスクを求める，位相鋭敏スペクトル近似マスク (PSA: phase sensitive approximation) の確率的モデルへの拡張とみなすことができる．ここで時間周波数マスクを用いた音源強調では，出力音 $\hat{S}_{\omega,\tau}$ が $\hat{G}_{\omega,\tau}X_{\omega,\tau}$ で求められることに着目すると，$p(\boldsymbol{S}_\tau|\boldsymbol{X}_\tau,\Theta_{\mathcal{M}})$ は複素ガウス分布を用いて以下のように記述できる.

$$p(\boldsymbol{S}_\tau|\boldsymbol{X}_\tau,\Theta_{\mathcal{M}})=\prod_{\omega=1}^{\Omega}\frac{1}{2\pi\sigma^2_{\omega,\tau}}\exp\left\{-\frac{\left|S_{\omega,\tau}-\hat{G}_{\omega,\tau}X_{\omega,\tau}\right|^2}{2\sigma^2_{\omega,\tau}}\right\}. \tag{4.22}$$

式 (4.22) において，未知パラメータは時間周波数マスク $\hat{G}_{\omega,\tau}$ と分散 $\sigma^2_{\omega,\tau}$ であるため，DNN は，図 4.6 のように，$\hat{G}_{\omega,\tau}$ と $\sigma^2_{\omega,\tau}$ 推定する関数として設計する．まず，$G_{\omega,\tau}$ と $\sigma^2_{\omega,\tau}$ を，各時間フレーム毎に周波数方向にまとめてベクトル化したものを以下のように定義する.

$$\boldsymbol{G}(\boldsymbol{x}_\tau):=\left(\hat{G}_{1,\tau},...,\hat{G}_{\Omega,\tau}\right)^\top, \tag{4.23}$$

$$\boldsymbol{\sigma}(\boldsymbol{x}_\tau):=\left(\sigma^2_{1,\tau},...,\sigma^2_{\Omega,\tau}\right)^\top, \tag{4.24}$$

そしてこれらのパラメータを以下のように推定する.

$$\boldsymbol{G}(\boldsymbol{x}_\tau)\leftarrow\phi_g\left\{\mathbf{W}^{(\mu)}\boldsymbol{z}^{(L-1)}_\tau+\mathbf{b}^{(\mu)}\right\}, \tag{4.25}$$

$$\boldsymbol{\sigma}(\boldsymbol{x}_\tau)\leftarrow\phi_\sigma\left\{\mathbf{W}^{(\sigma)}\boldsymbol{z}^{(L-1)}_\tau+\mathbf{b}^{(\sigma)}\right\}+C_\sigma, \tag{4.26}$$

$$\boldsymbol{z}^{(l)}_\tau=\phi_h\left\{\mathbf{W}^{(l)}\boldsymbol{z}^{(l-1)}_\tau+\mathbf{b}^{(l)}\right\}, \tag{4.27}$$

ここで C_σ は小さな分散値を避けるための正の定数である．また，l は層のインデックス，L は層数，$\mathbf{W}^{(\cdot)}$ と $\mathbf{b}^{(\cdot)}$ はそれぞれ重み行列とバイアスベクトルである．ゆえに DNN のパラメータは $\Theta_{\mathcal{M}}=\{\mathbf{W}^{(\mu)},\mathbf{W}^{(\sigma)},\mathbf{b}^{(\mu)},\mathbf{b}^{(\sigma)},\mathbf{W}^{(l)},\mathbf{b}^{(l)}|l\in(2,...,L-1)\}$ となる．また，関

数 ϕ_g，ϕ_σ，ϕ_h は活性化関数であり，従来法では，ϕ_g はシグモイド関数 [32, 91]，ϕ_σ は指数関数 [104] で実装される．また入力ベクトルは $\boldsymbol{z}_\tau^{(1)} = \boldsymbol{x}_\tau = (\boldsymbol{X}_{\tau-P}, ..., \boldsymbol{X}_\tau, ..., \boldsymbol{X}_{\tau+P})^\top$ である．

4.3.2 方策勾配法に基づく DNN 音源強調関数の学習

本節では，聴感評点を最大化するための目的関数，および方策勾配法を用いた DNN 音源強調関数の学習を説明する．今，$\mathcal{B}(\hat{\boldsymbol{S}}, \boldsymbol{X})$ を出力音 $\hat{\boldsymbol{S}}$ から音質を計算する評価関数として定義する．ここで $\mathcal{B}(\hat{\boldsymbol{S}}, \boldsymbol{X})$ は PESQ などの聴感評点から計算される関数であるため，DNN パラメータに関する勾配が解析的に求まらないブラックボックス関数であるとする．この設計の詳細は 4.3.4 節で説明する．

聴感評点を向上させるための目的関数として，評価関数 $\mathcal{B}(\hat{\boldsymbol{S}}, \boldsymbol{X})$ の期待値最大化を考える．

$$\mathbb{E}_{\hat{\boldsymbol{S}},\boldsymbol{X}}\left[\mathcal{B}(\hat{\boldsymbol{S}}, \boldsymbol{X})\right] = \iint \mathcal{B}(\hat{\boldsymbol{S}}, \boldsymbol{X}) p(\hat{\boldsymbol{S}}, \boldsymbol{X}) d\hat{\boldsymbol{S}} d\boldsymbol{X}, \tag{4.28}$$

ここで出力音 $\hat{\boldsymbol{S}}$ は観測信号 $\boldsymbol{X}$ より計算されるものである．そこで同時分布 $p(\hat{\boldsymbol{S}}, \boldsymbol{X})$ を観測信号の周辺分布 $p(\boldsymbol{X})$ と，観測信号を得た下での出力音の条件付き分布 $p(\hat{\boldsymbol{S}}|\boldsymbol{X})$ へと分化する．すると式 (4.28) は以下のように書き換えることができる．

$$\mathbb{E}_{\hat{\boldsymbol{S}},\boldsymbol{X}}\left[\mathcal{B}(\hat{\boldsymbol{S}}, \boldsymbol{X})\right] = \int p(\boldsymbol{X}) \int \mathcal{B}(\hat{\boldsymbol{S}}, \boldsymbol{X}) p(\hat{\boldsymbol{S}}|\boldsymbol{X}) d\hat{\boldsymbol{S}} d\boldsymbol{X}. \tag{4.29}$$

我々の目的は，聴感評点を最大化するように出力音を求める DNN を学習することである．そこで，最尤推定に基づく DNN 学習と同様に，条件付き分布 $p(\hat{\boldsymbol{S}}|\boldsymbol{X})$ を，方策関数 $q(\hat{\boldsymbol{S}}|\boldsymbol{X}, \Theta_\mathcal{M})$ として DNN で求めることにする．すると目的関数は以下のように記述することができる．

$$\mathcal{J}(\Theta_\mathcal{M}) = \mathbb{E}_{\hat{\boldsymbol{S}},\boldsymbol{X}}\left[\mathcal{B}(\hat{\boldsymbol{S}}, \boldsymbol{X})\right], \tag{4.30}$$

$$= \int p(\boldsymbol{X}) \int \mathcal{B}(\hat{\boldsymbol{S}}, \boldsymbol{X}) q(\hat{\boldsymbol{S}}|\boldsymbol{X}, \Theta_\mathcal{M}) d\hat{\boldsymbol{S}} d\boldsymbol{X}. \tag{4.31}$$

しかし，前述のとおり評価関数 $\mathcal{B}(\hat{\boldsymbol{S}}, \boldsymbol{X})$ はブラックボックス関数であるため，この目的関数の $\Theta_\mathcal{M}$ に関する勾配は解析的に求めることができない．そこで，この勾配を数値的に推定する “方策勾配法” [139] を利用する．方策関数である条件付き分布 $q(\hat{\boldsymbol{S}}|\boldsymbol{X}, \Theta_\mathcal{M})$ が $\Theta_\mathcal{M}$ に関して微分可能な関数の合成関数で設計されているとすれば，式 (4.31) の勾配は対数微分則 $\nabla_x p(\boldsymbol{x}) = p(\boldsymbol{x}) \nabla_x \ln p(\boldsymbol{x})$ を利用することで以下のように求めることができる．

$$\nabla_{\Theta_\mathcal{M}} \mathcal{J}(\Theta_\mathcal{M}) = \int p(\boldsymbol{X}) \int \mathcal{B}(\hat{\boldsymbol{S}}, \boldsymbol{X}) \nabla_{\Theta_\mathcal{M}} q(\hat{\boldsymbol{S}}|\boldsymbol{X}, \Theta_\mathcal{M}) d\hat{\boldsymbol{S}} d\boldsymbol{X}, \tag{4.32}$$

$$= \mathbb{E}_{\boldsymbol{X}}\left[\mathbb{E}_{\hat{\boldsymbol{S}}|\boldsymbol{X}}\left[\mathcal{B}(\hat{\boldsymbol{S}}, \boldsymbol{X}) \nabla_{\Theta_\mathcal{M}} \ln q(\hat{\boldsymbol{S}}|\boldsymbol{X}, \Theta_\mathcal{M})\right]\right]. \tag{4.33}$$

ここで式 (4.33) 内の期待値は解析的に求まらず，また聴感評点を最大化する $\hat{\boldsymbol{S}}$ は未知であるため，$\boldsymbol{X}$ に関する期待値を学習データの平均値で，また $\hat{\boldsymbol{S}}$ に関する期待値をサンプリング法で求める．

$$\nabla_{\Theta_{\mathcal{M}}}\mathcal{J}(\Theta_{\mathcal{M}}) \approx \frac{1}{T}\sum_{\tau=1}^{T}\frac{1}{K}\sum_{k=1}^{K}\mathcal{B}(\hat{\boldsymbol{S}}_{\tau}^{(k)},\boldsymbol{X}_{\tau})\nabla_{\Theta_{\mathcal{M}}}\ln q(\hat{\boldsymbol{S}}_{\tau}^{(k)}|\boldsymbol{X}_{\tau},\Theta_{\mathcal{M}}), \tag{4.34}$$

$$\hat{\boldsymbol{S}}_{\tau}^{(k)} \sim q(\hat{\boldsymbol{S}}|\boldsymbol{X}_{\tau},\Theta_{\mathcal{M}}), \tag{4.35}$$

ここで $\hat{\boldsymbol{S}}_{\tau}^{(k)}$ は方策関数からサンプリングで求めた出力音，K は期待値演算を近似するのに十分なサンプリング回数，上付き文字の (k) は k 回目のサンプリングに関する変数を表すインデックス，また $\sim$ は右辺の確率分布からのサンプリングを表す演算子である．なお，式 (4.35) のサンプリングの実装法については 4.3.3 節で詳しく述べる．

PESQ をはじめとする多くの聴感評点は，8 秒程度の一発話全体の音源強調結果から 1 つの値しか求まらない．そのため，聴感評点から値の求まる評価関数 $\mathcal{B}(\hat{\boldsymbol{S}}_{\tau}^{(k)},\boldsymbol{X}_{\tau})$ や，式 (4.34) の勾配はは各時間フレーム τ ごとに求まらない．そこで期待値を近似するための平均値を，時間フレーム τ ではなく，$\mathcal{I}$ 個の発話データに関して求める．まず，i 番目の観測音を $\mathbf{X}^{(i)} := (\boldsymbol{X}_1^{(i)},...,\boldsymbol{X}_{T^{(i)}}^{(i)})$，$i$ 番目の観測音に基づきサンプリングした k 番目の出力音を $\hat{\mathbf{S}}^{(i,k)} := (\hat{\boldsymbol{S}}_1^{(i,k)},...,\hat{\boldsymbol{S}}_{T^{(i)}}^{(i,k)})$ と定義する．そして，目的関数の勾配を以下のように求める．

$$\nabla_{\Theta_{\mathcal{M}}}\mathcal{J}(\Theta_{\mathcal{M}}) \approx \frac{1}{\mathcal{I}}\sum_{i=1}^{\mathcal{I}}\nabla_{\Theta_{\mathcal{M}}}\mathcal{J}^{(i)}(\Theta_{\mathcal{M}}) \tag{4.36}$$

$$\nabla_{\Theta_{\mathcal{M}}}\mathcal{J}^{(i)}(\Theta_{\mathcal{M}}) \approx \sum_{k=1}^{K}\frac{\mathcal{B}\left(\hat{\mathbf{S}}^{(i,k)},\mathbf{X}^{(i)}\right)}{KT^{(i)}}\sum_{\tau=1}^{T^{(i)}}\nabla_{\Theta_{\mathcal{M}}}\ln q(\hat{\boldsymbol{S}}_{\tau}^{(i,k)}|\boldsymbol{X}_{\tau}^{(i)},\Theta_{\mathcal{M}}) \tag{4.37}$$

ここで $T^{(i)}$ は i 番目の発話の時間フレーム数を表す．なお式 (4.36) の導出は，本章の付録である 4.5.1 節で述べる．

4.3.3 時間周波数マスク処理の制約に基づくサンプリングアルゴリズム

式 (4.35) におけるサンプリング演算は，メルセンヌ・ツイスター法 [143] などの一般的な疑似乱数生成アルゴリズムで実装すると，時間周波数マスク処理である式 (2.4) を満たさないような出力音が生成されてしまう．図 4.7 に，この問題と提案する解決法を図示した．提案法において時間周波数マスクは，$0 \leq G_{\omega,\tau} \leq 1$ を満たすと仮定されている．すなわち，時間周波数マスク処理は位相スペクトルを変化させないため，その出力音は図 4.7 の点線上に存在しなくてはならない．しかし，$q(\hat{\boldsymbol{S}}|\boldsymbol{X},\Theta_{\mathcal{M}})$ を式 (4.22) の複素ガウス分布などで実装した場合，一般的な乱数生成器から生成される出力音は，必ずしもこの制約を満たすとは限らない．この解決策の一つとして，提案法では，PSA に基づく時

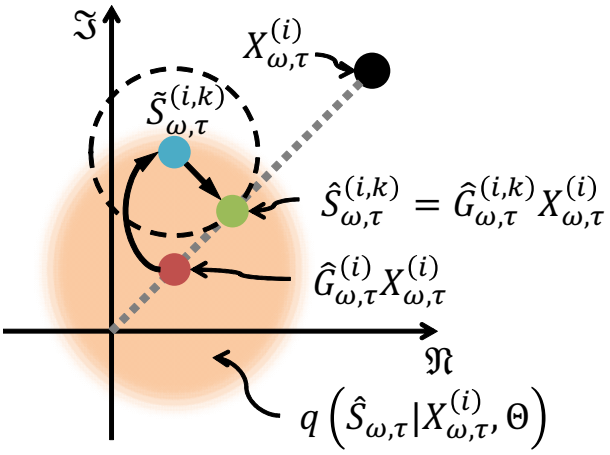


図 4.7: $q(\hat{\boldsymbol{S}}|\boldsymbol{X}, \Theta_{\mathcal{M}})$ を複素ガウス分布とした際の提案サンプリングアルゴリズム．

間周波数マスクの再設計を行う．まず，一般的な乱数生成アルゴリズムを用いて，仮の出力音 $\tilde{S}^{(i,k)}_{\omega,\tau}$ を生成する．次いで，時間周波数マスク $\hat{G}^{(i,k)}_{\omega,\tau}$ を，仮の出力音と $\tilde{S}^{(i,k)}_{\omega,\tau}$ 時間周波数マスクの処理音 $\hat{G}^{(i,k)}_{\omega,\tau} X^{(i)}_{\omega,\tau}$ の二乗誤差を最小化するように，PSA の設計式 [32, 91] に基づき求める．

$$\hat{G}^{(i,k)}_{\omega,\tau} = \min\left(1, \max\left(0, \frac{|\tilde{S}^{(i,k)}_{\omega,\tau}|}{|X^{(i)}_{\omega,\tau}|} \cos\left(\phi^{(\tilde{S}^{(i,k)})}_{\omega,\tau} - \phi^{(X^{(i)})}_{\omega,\tau}\right)\right)\right) \tag{4.38}$$

ここで $\phi^{(\tilde{S}^{(i,k)})}_{\omega,\tau}$ と $\phi^{(X^{(i)})}_{\omega,\tau}$ は，それぞれ $\tilde{S}^{(i,k)}_{\omega,\tau}$ と $X^{(i)}_{\omega,\tau}$ の位相スペクトルである．すると，出力音は以下のように生成できる．

$$\hat{S}^{(i,k)}_{\omega,\tau} = \hat{G}^{(i,k)}_{\omega,\tau} X^{(i)}_{\omega,\tau}. \tag{4.39}$$

4.3.4 学習を安定させるための評価関数と時間周波数マスクの設計

本節では，式 (4.36) の勾配計算を安定させるための処理について説明する．提案法では，勾配をサンプリング法で近似的に求めるため，勾配の推定誤差に起因して学習が不安定になることがある．勾配推定を安定させる方法として，勾配の推定量 $\nabla_{\Theta_{\mathcal{M}}} \mathcal{J}(\Theta_{\mathcal{M}})$ の分散を低減させる方法がある．式 (4.37) からわかるように，提案法において勾配は，評価関数 $\mathcal{B}(\hat{\mathbf{S}}, \mathbf{X})$ と対数尤度の勾配 $\nabla_{\Theta_{\mathcal{M}}} \ln q(\hat{\boldsymbol{S}}^{(i,k)}_{\tau} | \boldsymbol{X}^{(i)}_{\tau}, \Theta_{\mathcal{M}})$ の積である．ゆえに，$\nabla_{\Theta_{\mathcal{M}}} \mathcal{J}(\Theta_{\mathcal{M}})$ の分散を低減させるためには，評価関数と対数尤度の勾配の分散を低減させればよい．そこで提案法では，以下の二つの処理を導入することで，$\nabla_{\Theta_{\mathcal{M}}} \mathcal{J}(\Theta_{\mathcal{M}})$ の分散を低減させる．

- 聴感評点を評価関数 $\mathcal{B}(\hat{\mathbf{S}}, \mathbf{X})$ として直接用いるのではなく，聴感評点を正規化したものを評価関数として用いる．

- DNN が推定する時間周波数マスク $\hat{G}^{(i)}_{\omega,\tau}$ とサンプリングされた時間周波数マスク $\hat{G}^{(i,k)}_{\omega,\tau}$ の差分である $\Delta\hat{G}^{(i,k)}_{\omega,\tau} = \hat{G}^{(i,k)}_{\omega,\tau} - \hat{G}^{(i)}_{\omega,\tau}$ が区間 $[-\lambda, \lambda]$ に収まるようにクリッピング処理を行う．

以降，評価関数 $\mathcal{B}(\hat{\mathbf{S}}, \mathbf{X})$ と聴感評点の記述を区別するために，$\mathbf{X}$ と $\hat{\mathbf{S}}$ より求まる聴感評点を $\mathcal{Z}(\hat{\mathbf{S}}, \mathbf{X})$ と記述する．

評価関数の分散を低減する方法として，$\mathcal{Z}(\tilde{\mathbf{S}}, \mathbf{X})$ から定数を減算して評価値とする，ベースライン減算法 [139] がある．提案法ではベースラインして，観測信号を得た下での聴感評点の期待値 $\mathbb{E}_{\hat{\mathbf{S}}|\mathbf{X}}\left[\mathcal{Z}(\hat{\mathbf{S}}, \mathbf{X})\right]$ を採用する [144]．

$$\mathcal{B}\left(\hat{\mathbf{S}}, \mathbf{X}\right) = \mathcal{Z}(\hat{\mathbf{S}}, \mathbf{X}) - \mathbb{E}_{\hat{\mathbf{S}}|\mathbf{X}}\left[\mathcal{Z}(\hat{\mathbf{S}}, \mathbf{X})\right] \tag{4.40}$$

ただしこの期待値は解析的に求まらないため，期待値演算を K 回のサンプリング結果の平均値として，以下のように実装する．

$$\mathcal{B}\left(\hat{\mathbf{S}}^{(i,k)}, \mathbf{X}^{(i)}\right) = \mathcal{Z}(\hat{\mathbf{S}}^{(i,k)}, \mathbf{X}^{(i)}) - \frac{1}{K}\sum_{j=1}^{K}\mathcal{Z}(\hat{\mathbf{S}}^{(i,j)}, \mathbf{X}^{(i)}) \tag{4.41}$$

また，$\nabla_{\Theta_{\mathcal{M}}} \ln q(\hat{\boldsymbol{S}}^{(i,k)}_{\tau}|\boldsymbol{X}^{(i)}_{\tau}, \Theta_{\mathcal{M}})$ の分散は，DNN が推定した聴感評点を最大化する出力音の MAP 推定量 $\hat{G}^{(i)}_{\omega,\tau}\hat{X}^{(i)}_{\omega,\tau}$ と，サンプリング法で生成した出力音 $\hat{G}^{(i,k)}_{\omega,\tau}\hat{X}^{(i)}_{\omega,\tau}$ の差分が大きくなるほど増大する．そこで，$\nabla_{\Theta_{\mathcal{M}}} \ln q(\hat{\boldsymbol{S}}^{(i,k)}_{\tau}|\boldsymbol{X}^{(i)}_{\tau}, \Theta_{\mathcal{M}})$ の分散を低減させるために，DNN が推定する時間周波数マスク $\hat{G}^{(i)}_{\omega,\tau}$ とサンプリングされた時間周波数マスク $\hat{G}^{(i,k)}_{\omega,\tau}$ の差分である $\Delta\hat{G}^{(i,k)}_{\omega,\tau} = \hat{G}^{(i,k)}_{\omega,\tau} - \hat{G}^{(i)}_{\omega,\tau}$ が区間 $[-\lambda, \lambda]$ に収まるようにクリッピング処理を行う．

$$\Delta\hat{G}^{(i,k)}_{\omega,\tau} \leftarrow \hat{G}^{(i,k)}_{\omega,\tau} - \hat{G}^{(i)}_{\omega,\tau} \tag{4.42}$$

$$\Delta\hat{G}^{(i,k)}_{\omega,\tau} \leftarrow \begin{cases} \lambda & (\Delta\hat{G}^{(i,k)}_{\omega,\tau} > \lambda) \\ \Delta\hat{G}^{(i,k)}_{\omega,\tau} & (-\lambda \leq \Delta\hat{G}^{(i,k)}_{\omega,\tau} \leq \lambda) \\ -\lambda & (\Delta\hat{G}^{(i,k)}_{\omega,\tau} < -\lambda) \end{cases} \tag{4.43}$$

$$\hat{G}^{(i,k)}_{\omega,\tau} \leftarrow \hat{G}^{(i)}_{\omega,\tau} + \Delta\hat{G}^{(i,k)}_{\omega,\tau} \tag{4.44}$$

4.3.5 提案法の学習アルゴリズム

本節では，提案する学習アルゴリズムの処理手順を，図 4.8 に沿って説明する．以降では，サンプリングアルゴリズムの処理の簡単のために，$q(\hat{\boldsymbol{S}}|\boldsymbol{X}, \Theta_{\mathcal{M}})$ は複素ガウス分布で実装されているとする．

まず,，源信号の学習データから 1 発話をランダムに選択し，雑音データセットからそれと同じ長さのノイズをランダムに取得し，源信号を雑音をランダムな SNR で重畳する

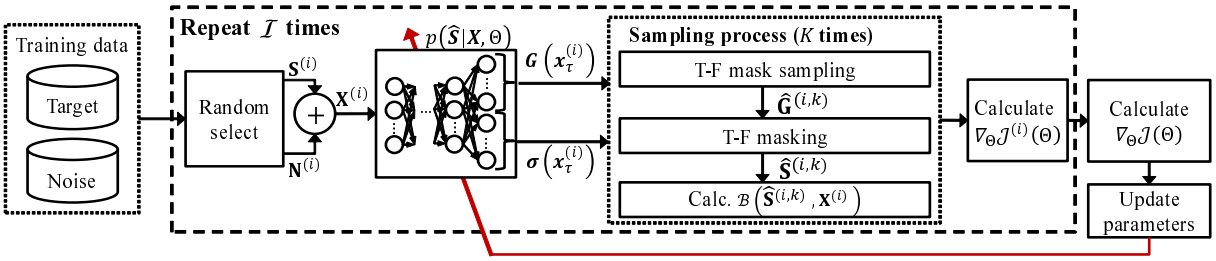


図 4.8: 提案法の学習手順

ことで i 番目の観測信号 $\mathbf{X}^{(i)}$ を生成する．次に，時間周波数マスク $\boldsymbol{G}(\boldsymbol{x}_\tau^{(i)})$ と誤差分散 $\boldsymbol{\sigma}(\boldsymbol{x}_\tau^{(i)})$ を推定する．そして，k 番目の仮の出力音 $\tilde{S}_{\omega,\tau}^{(i,k)'}$ を，以下の複素ガウス分布から，疑似乱数生成アルゴリズムで生成する．

$$\begin{bmatrix} \Re\left(\tilde{S}_{\omega,\tau}^{(i,k)}\right) \\ \Im\left(\tilde{S}_{\omega,\tau}^{(i,k)}\right) \end{bmatrix} \sim \mathcal{N}_{\mathbb{C}}\left(\hat{G}_{\omega,\tau}^{(i)} \begin{bmatrix} \Re\left(X_{\omega,\tau}^{(i)}\right) \\ \Im\left(X_{\omega,\tau}^{(i)}\right) \end{bmatrix}, \sigma_{\omega,\tau}^2 \boldsymbol{I}\right), \tag{4.45}$$

ここで $\boldsymbol{I}$ は 2×2 の単位行列であり，$\Re$ と $\Im$ は複素数の実部と虚部を表す．そして時間周波数マスク $\hat{G}_{\omega,\tau}^{(i,k)}$ を式 (4.38) および式 (4.42)–(4.44) で求める．ただし，学習の収束を早めるために，各時間周波数ビンごとに，確率 $1-\epsilon$ で，DNN が推定した時間周波数マスク $\hat{G}_{\omega,\tau}^{(i)}$ を用いる，以下の ϵ-greedy アルゴリズムを用いる．

$$\hat{G}_{\omega,\tau}^{(i,k)} \leftarrow \begin{cases} \min\left(1, \max\left(0, \frac{|\tilde{S}_{\omega,\tau}^{(i,k)}|}{|X_{\omega,\tau}^{(i)}|} \cos\left(\phi_{\omega,\tau}^{(\tilde{S}^{(i,k)})} - \phi_{\omega,\tau}^{(X^{(i)})}\right)\right)\right) & \text{(with prob. } \epsilon\text{)} \\ \hat{G}_{\omega,\tau}^{(i)} & \text{(otherwise)} \end{cases} \tag{4.46}$$

そして出力音 $\hat{\mathbf{S}}^{(i,k)}$ を式4.39) に基づき計算し，聴感評点 $\mathcal{Z}(\hat{\mathbf{S}}^{(i,k)}, \mathbf{X}^{(i)})$ と評価値 $\mathcal{B}(\hat{\mathbf{S}}^{(i,k)}, \mathbf{X}^{(i)})$ を式 (4.41) で求める．これらの処理を $\mathcal{I}$ 発話分繰り返したのち，$\Theta_{\mathcal{M}}$ を式 (4.36) から求まる勾配に基づき更新する．

4.3.6 評価実験

強化学習に基づく時間周波数マスク生成の有効性を確かめるために，客観評価試験を行った．まず，提案法により聴感評点が向上するかを調査するために，学習回数と聴感評点の変化の関係性を調査した．次いで，提案法による聴感評点の値を，二乗誤差最小化に基づく目的関数での DNN 学習 [91]，および最尤推定法による DNN 学習と比較した．

(a) 実験条件

データセット

所望の源信号は日本語の発話音声とし，学習データには ATR 音声データベース [141] から男性 11 名，女性 11 名による全 6640 発話を利用した．雑音は CHiME-3 の雑音データ

表 4.3: 実験条件.

Parameters for signal processing	
Sampling rate	16.0 kHz
FFT length	512 pts
FFT shift length	256 pts
# of mel-filterbanks	64
Smoothing parameter β	0.3
Lower threshold $G^{\min}$	0.158 (= −16 dB)
Training SNR (dB)	-6, 0, 6, 12
DNN architecture	
# of hidden layers for DNNs	3
# of hidden units for DNNs	1024
Activation function (T-F mask, ϕ_g)	sigmoid
Activation function (variance, ϕ_σ)	exponential
Activation function (hidden, ϕ_h)	ReLU
Context window size P	5
Variance regularization parameter C_σ	10^{-4}
Parameters for ML-based DNN training	
Initial step-size	10^{-4}
Step-size threshold for early-stopping	10^{-7}
Dropout probability (input layer)	0.2
Dropout probability (hidden layer)	0.5
L_2 normalization parameter	10^{-4}
Parameters for proposed DNN training	
Step-size	10^{-6}
# of utterance $\mathcal{I}$	10
# of T-F mask sampling K	20
Clipping parameter λ	0.05
ϵ-greedy parameter ϵ	0.05

セット [142] より，“cafes”，“street junctions”，“public transport (buses)”，“pedestrian areas” の 4 種類を用いた．これらのデータは 16kHz でサンプリングした．テストデータには，所望の源信号は学習データとは異なる男性 3 名，女性 3 名による全 300 発話，雑音には空港，アミューズメントパーク，オフィス，パーティ会場で収録した環境雑音を利用した．観測信号は，源信号と雑音を SNR が -6, 0，6，12 dB のどれかとなるようにランダムに重畳することで生成した．

DNN 構造

従来法と提案法のニューラルネットワークには全結合多層ニューラルネットワーク（DNN）を用いた．DNN の構造は隠れ層 3 層，隠れユニット数 1024 とした．$\boldsymbol{x}_\tau$ は，各次元が平均 0，分散 1 となるよう正規化した．活性化関数 ϕ_g，ϕ_σ，ϕ_h にはそれぞれ，シグモイド関数，指数関数，ランプ関数（ReLU: rectified linear unit) を用いた．コンテキスト窓は $P = 5$ とし，また $C_\sigma = 10^{-4}$ とした．勾配法には Adam 法 [84] を用いた．また過適合を避けるために，DNN の入出力変数は $\mathsf{B} = 64$ のメルフィルタバンク行列で圧縮をした．また DNN で推定した時間周波数マスクと分散パラメータはメルフィルタバンク行列の疑似逆行列を用いて線形周波数領域へ復号した [145]．

二乗誤差最小化に基づく目的関数での DNN 学習には，PSA を推定するための目的関数を利用した [91]．PSA の推定には分散パラメータが不要なため，$\boldsymbol{\sigma}(\boldsymbol{x}_\tau)$ は推定せず，DNN は $\boldsymbol{G}(\boldsymbol{x}_\tau)$ のみを推定した．最尤推定法に基づく DNN 音源強調関数の学習では，4.3.1 節で説明した方法を利用した．これらの手法の学習には，過適合を防ぐために，ドロップアウトと早期終了（early-stopping）アルゴリズムを用いた．ドロップアウト確率は，入力層が 0.2，隠れ層が 0.5 とした．また，早期終了は 3 章で利用したものと同様のアルゴリズムを用いた．

提案法の学習は，最尤推定法で学習した DNN パラメータを初期値として行った．ステップサイズは 10^{-6} とし，$\nabla_\Theta \mathcal{J}(\Theta)$ を求めるためのいてレーションパラメータは $\mathcal{I} = 10$，$K = 20$ とした．また ϵ-greedy アルゴリズムのパラメータは $\epsilon = 0.05$ とし，時間周波数マスクのクリッピングパラメータは $\lambda = 0.05$ とした．聴感評点には音声品質の指標である PSEQ と，音声明瞭度の指標である short-time intelligibility measure (STOI) [146] を用いた．これらの聴感評点は，最小値が 0，最大値が 100 となるように正規化して利用した．

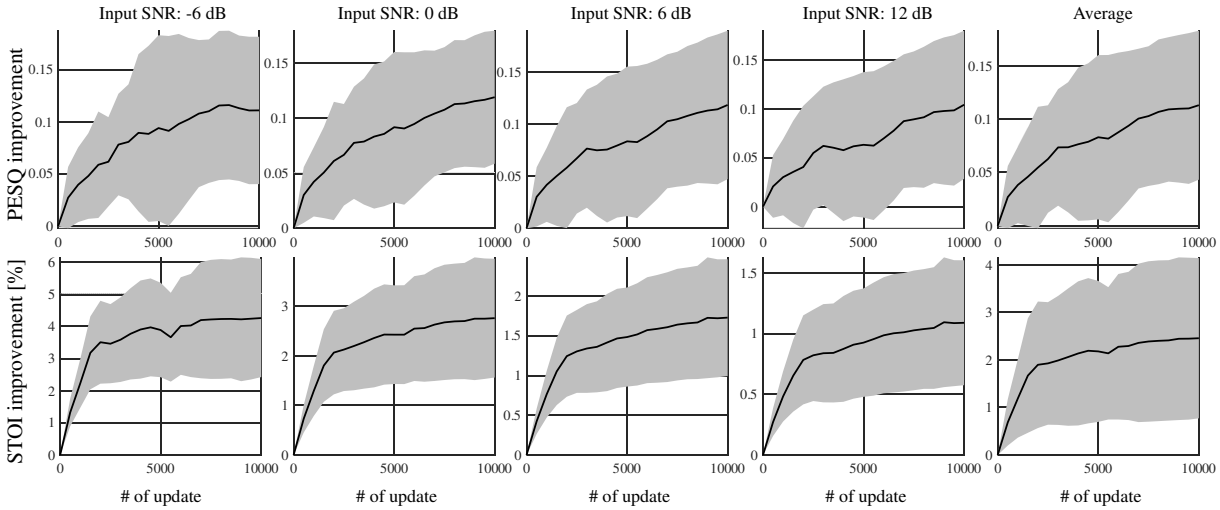


図 4.9: 学習回数と聴感評点の関係性．x 軸が学習回数，y 軸が最尤推定法で学習した DNN から求めた出力音の聴感評点との差分 (PSI: perceptual score improvement) を表す．実線が PSI の平均値，グレーの領域が PSI の標準偏差を表す．

その他の実験条件

時間周波数マスク処理による非線形歪みを低減させるために，時間周波数マスクのスムージングとフロアリングを以下で行った．

$$\hat{G}_{\omega,\tau} \leftarrow \max\left(G^{\min}, \hat{G}_{\omega,\tau}\right) \tag{4.47}$$

$$\hat{G}_{\omega,\tau} \leftarrow \beta\hat{G}_{\omega,\tau} + (1-\beta)\hat{G}_{\omega,\tau-1} \tag{4.48}$$

ただし，時間周波数マスクのフロア値は $G^{\min} = 0.158$ とし，平滑化パラメータは $\beta = 0.3$ とした．フーリエ変換長は 512 点，シフト長は 256 点とした．その他の実験条件は表 4.3 に示した．

(b) 動作実験

提案法によって聴感評点が向上するかを確認するために，学習回数と聴感評点の関係性を評価した．もし提案法により方策関数が正しく学習できているならば，学習回数が増加するに従い聴感評点も向上するはずである．学習の指標として，最尤推定法で学習した DNN から求めた出力音の聴感評点との差分（PSI: perceptual score improvement）を利用した．この理由は，提案法の初期化は最尤推定法で行うため，学習回数 0 の PSI は 0 となり，学習が成功しているならば PSI が正の値をとるためである．

図 4.9 に PSEQ と STOI の PSI を示す．学習回数に応じて PSI が向上していることが確認できる．このことから，提案法は聴感評点などのブラックボックスな評価指標を向上させるようにニューラルネットワークのパラメータを学習できることがわかる．

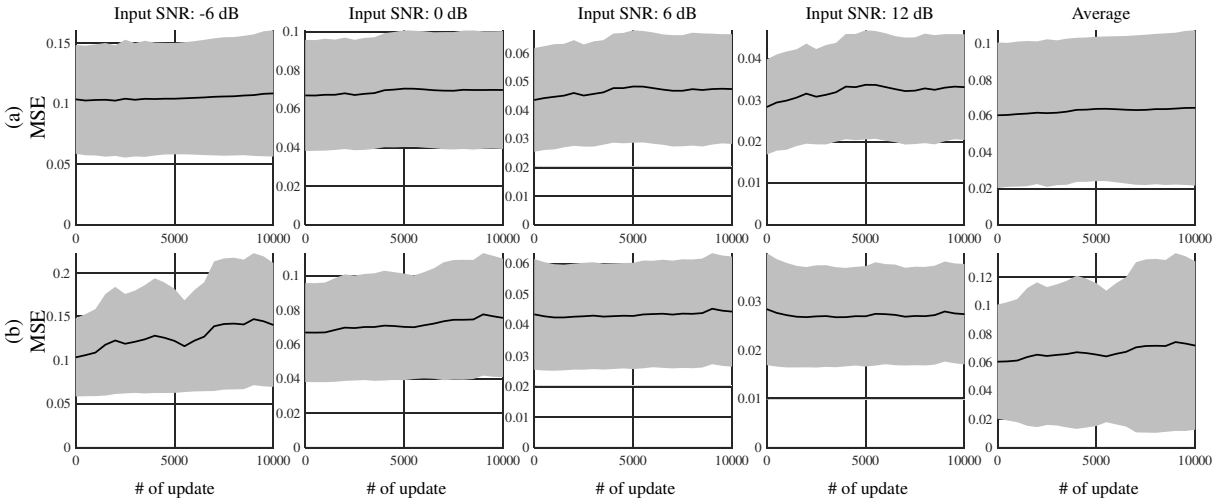


図 4.10: 学習回数と MSE の関係性．聴感評点はそれぞれ，(a) が PESQ，(b) が STOI である．x 軸が学習回数，y 軸が MSE を表す．実線が MSE の平均値，グレーの領域が MSE の標準偏差を表す．

また学習回数と以下の式で表される MSE の関係性も評価した．

$$\mathrm{MSE} = \frac{1}{\Omega T}\sum_{\tau=1}^{T}\sum_{\omega=1}^{\Omega}\left(|S_{\omega,\tau}| - |G_{\omega,\tau}X_{\omega,\tau}|\right)^2 \tag{4.49}$$

図 4.10 にその結果を示す．いくつかの SNR 条件では，聴感評点が向上しているにも関わらず，MSE が減少しなかった．これらの結果から，提案法によりニューラルネットワークが，従来用いられてきた二乗誤差の最小化とは異なる基準で学習されていることが確認できる．

(c) 客観評価実験

提案法の音源強調性能を，従来法と比較した．性能評価指標として，信号対歪比（SDR: signal-to-distortion ratio），PESQ，および STOI を利用した．SDR は以下の式で定義し，“BSS-Eval toolbox [134]” を利用して求めた．

$$\mathrm{SDR}\ [\mathrm{dB}] := 10\log_{10}\frac{\sum_{\tau=1}^{T}\sum_{\omega=1}^{\Omega}|S_{\omega,\tau}|^2}{\sum_{\tau=1}^{T}\sum_{\omega=1}^{\Omega}|S_{\omega,\tau}-\hat{S}_{\omega,\tau}|^2}, \tag{4.50}$$

表 4.4: 評価結果（平均 ± 標準偏差）．各アスタリスクは，そのスコアが他の手法のスコアと比べて有意に高いことを表す．

Input SNR: -6 dB

	Performance measurement		
Objective function	SDR [dB]	PESQ	STOI [%]
MMSE [91]	$\mathbf{5.77 \pm 1.97}$	1.89 ± 0.25	73.9 ± 5.50
ML (4.20)	5.76 ± 2.13	1.92 ± 0.24	75.2 ± 5.77
Prop. (PESQ)	5.61 ± 2.32	$\mathbf{*2.03 \pm 0.25}$	76.8 ± 5.55
Prop. (STOI)	4.45 ± 2.42	1.78 ± 0.25	$\mathbf{*79.5 \pm 5.15}$

Input SNR: 0 dB

	Performance measurement		
Objective function	SDR [dB]	PESQ	STOI [%]
MMSE [91]	10.4 ± 1.81	2.33 ± 0.20	85.1 ± 4.00
ML (4.20)	$\mathbf{10.7 \pm 1.86}$	2.36 ± 0.19	86.3 ± 3.81
Prop. (PESQ)	$\mathbf{10.7 \pm 1.84}$	$\mathbf{*2.48 \pm 0.19}$	86.7 ± 3.92
Prop. (STOI)	10.42 ± 2.01	2.27 ± 0.19	$\mathbf{*89.1 \pm 3.52}$

Input SNR: 6 dB

	Performance measurement		
Objective function	SDR [dB]	PESQ	STOI [%]
MMSE [91]	14.4 ± 1.80	2.67 ± 0.18	91.7 ± 2.94
ML (4.20)	15.1 ± 1.67	2.72 ± 0.17	92.8 ± 2.67
Prop. (PESQ)	14.8 ± 1.57	$\mathbf{*2.84 \pm 0.17}$	92.5 ± 2.91
Prop. (STOI)	$\mathbf{*15.5 \pm 1.83}$	2.65 ± 0.16	$\mathbf{*94.6 \pm 2.51}$

Input SNR: 12 dB

	Performance measurement		
Objective function	SDR [dB]	PESQ	STOI [%]
MMSE [91]	17.9 ± 1.88	2.95 ± 0.18	95.1 ± 2.23
ML (4.20)	19.2 ± 1.59	3.05 ± 0.15	96.2 ± 1.94
Prop. (PESQ)	18.4 ± 1.57	$\mathbf{*3.15 \pm 0.16}$	95.5 ± 2.24
Prop. (STOI)	$\mathbf{*20.3 \pm 1.65}$	3.00 ± 0.14	$\mathbf{*97.3 \pm 1.85}$

表 4.4 に，評価結果を示す．表中のアスタリスクは，そのスコアが他の手法のスコアと比べて，対応のある片側 t-検定において有意差が認められたことを表す ($\alpha = 0.05$)．

SDR は，従来の MMSE や最尤法に基づく目的関数で DNN 音源強調関数を学習したときに高くなる傾向が見られた．一方，聴感評点である PESQ と STOI は，聴感評点に基づく目的関数を用いて提案法により学習することで，全ての SNR 条件で，従来法と比べて有意に向上していることがわかる．

図 4.11 に，観測信号，各手法で推定された時間周波数マスク，出力音の例を示す．また，表 4.5 に各スコアを示す．最尤推定法に基づく時間周波数マスクは，その目的関数の特性から，全周波数帯域にわたって目的音の歪を平均的に減少させている．一方で，PESQ を聴感評点として提案法により学習すると，広域の残留雑音を強く抑圧するように学習が働いていることがわかる（図 4.11 (b)）．一方，STOI を ESQ を聴感評点として提案法により学習すると，目的音の歪を低減させるために，雑音の抑圧量を低下させていることがわかる（Fig. 4.11 (c)）．この結果は，広域の残留雑音が音声品質を低下させる一方で，目的音の歪みは音声明瞭度を低下させることに起因しており，学習に用いた各聴感評点の特性を反映した時間周波数マスクが推定で来ていることを示唆している．以上より，聴感評点に基づき目的関数を設計し，強化学習の一種である方策勾配法を応用して DNN 音源強調関数を学習することにより，明示的にラベルデータを設計できない主観品質最大化の問題においても，ニューラルネットワークを学習することができることがわかった．

表 4.5: 図 4.11 の出力音の定量評価値．

	Performance measurement		
Objective function	SDR [dB]	PESQ	STOI [%]
ML (4.20)	**11.9**	2.44	88.4
Prop. (PESQ)	11.3	**2.62**	87.4
Prop. (STOI)	10.2	2.25	**89.4**

4.3.7 時間周波数マスクの生成に基づく音源強調のまとめ

本節では，強化学習に基づき時間周波数マスクの生成するためのニューラルネットワークを学習するために，方策関数を勾配法で学習する “REINFORCE algorithm”[139, 102] を利用した目的関数を提案した．時間周波数マスクの選択と同様に，DNN-RL の出力音の音質評価と DNN-MMSE の出力音の音質評価の比較報酬を利用して報酬関数と目的関数を設計した．客観評価実験および主観評価実験の結果から，聴感評点を報酬関数として音源強調のためのニューラルネットワークを強化学習の枠組みで目的関数を設計/学習することにより，明示的にラベルデータを設計できない主観品質最大化の問題においても，ニューラルネットワークを学習することができることがわかった．

4.4 本章のまとめ

本章では高品質な音声通信や聴覚補助の実現に向け，DNN を利用した音源強調の出力音の主観品質を向上させるために，ラベルデータが定義できない源信号を強調するための手法について研究した．従来の DNN 音源強調では，源信号の振幅スペクトルなどをラベルデータとし，DNN の出力とラベルデータの二乗誤差を最小化するように DNN を学習させるため，出力音に歪が生じて主観品質が低下するという問題があった．一方，主観品質と相関の高い評価値（聴感評点）を計算機を用いて評価する手法は存在したものの，その計算方法は複雑かつ微分可能な関数の合成関数で設計されていないため，DNN を学習するための目的関数としてそのまま利用することはできなかった．そこで本章では強化学習のフレームワークを応用し，ラベルデータを用意する代わりに主観評価値と相関の高い音質評価値 [50, 51, 52] を報酬関数に利用し，それを最大化するようための目的関数を設計した．その実装法として 4.2 節では時間周波数マスクの選択に基づく音源強調，4.3 節では時間周波数マスクの生成に基づく音源強調を提案した．定量評価試験では，提案する目的関数を利用することで，どちらの手法でも聴感評点を最大化するようにニューラルネットワークを学習できることを確認した．また主観評価試験では，提案法は従来の MMSE に基づく目的関数を利用した音源強調よりも高い主観品質で音源強調できることを示した．

4.2 節で提案した時間周波数マスクの選択に基づく音源強調は，時間周波数マスクテンプレートの数を少なくすることで解空間を小さくすることができるため，比較的小規模なニューラルネットワークを用いても音源強調ができる．しかしテンプレートの数が少ない場合，表現できる時間周波数マスクの自由度も小さくなるため，DNN-MMSE と比べた音質の改善量は小さかった．一方，4.3 節で提案した時間周波数マスクの生成に基づく音源強調は，従来の DNN を用いた回帰に基づく音源強調と同等の柔軟性で時間周波数マスクが生成可能であり，DNN-MMSE と比べて音質が大きく改善した．しかした時間周波数マスクの選択に基づく音源強調と比べ解空間が広いため，方策関数には大規模なニューラルネットワークを用いる必要がある．提案法を実環境で動作させる際には，学習データ量や計算時間，また音源強調を行う計算機パワーを考慮して手法を選択する必要がある．

また本手法は，従来の誤差逆伝搬法では利用が困難であった，微分不可能な評価関数を用いてニューラルネットワークを学習できる枠組みである．本章では聴感評点を例に挙げ本手法の有効性を示したが，枠組み自体は聴感評点だけでなく，人間の評価や別のセンサー値なども用いることができる．今後は，聴感評点だけでなく様々な評価指標を用いて提案法の有効性を示していきたい．

4.4.1 本章の貢献と関連研究

本章の内容は，研究業績リスト [C-2] の内容をまとめたものである．この研究の貢献は，従来の DNN 音源強調の代表的な学習法である誤差逆伝搬法では利用が困難であった，微分不可能な評価関数を用いてニューラルネットワークを学習できる枠組みを提供した点にある．

微分不可能な評価関数を用いてニューラルネットワークを学習する枠組みとして，方策関数や方策関数をニューラルネットワークで表現し，それを強化学習で学習する手法は広く知られているが，多くの手法では用いられる “行動” は離散的に表現できるものであった [100, 101]．また，連続的な行動方策を強化学習で学習する手法として，policy gradient 法が知られているが，これをディープニューラルネットワークの学習に利用した研究は少ない．さらに，音源強調における時間周波数マスク処理は観測信号に影響を及ぼすことはなく，強化学習における “環境へのフィードバック” が存在しないため，提案法は厳密には強化学習ではない．

ゆえに本研究は，微分不可能な評価関数を用いてニューラルネットワークを学習する枠組みである強化学習に着想を得た，音源強調のためのニューラルネットワークを学習する新たな枠組みを提案した研究といえる．この貢献により，これまで音源強調の学習に利用できなかった聴感評点や人間の評価などの，より “高次” な評価尺度を目的関数として利用できるようになり，ニューラルネットワークを用いた音源強調の応用範囲を広げることができる．

4.5 本節の付録

4.4.1 式 (4.36) の導出

本節では，式 (4.36) の導出を説明する．まず，式 (4.30)(4.31) より，目的関数を $\mathcal{B}(\hat{\mathbf{S}}, \mathbf{X})$ の期待値として以下のように定義する．

$$\mathcal{J}(\Theta_{\mathcal{M}}) = \mathbb{E}_{\hat{\mathbf{S}},\mathbf{X}}\left[\mathcal{B}(\hat{\mathbf{S}}, \mathbf{X})\right] \tag{4.51}$$

$$= \int p(\mathbf{X}) \int \mathcal{B}(\hat{\mathbf{S}}, \mathbf{X}) q(\hat{\mathbf{S}}|\mathbf{X}, \Theta_{\mathcal{M}}) d\hat{\mathbf{S}} d\mathbf{X} \tag{4.52}$$

すると式 (4.52) の勾配は，対数微分則を用いて以下のように変形できる．

$$\nabla_{\Theta_{\mathcal{M}}} \mathcal{J}(\Theta_{\mathcal{M}}) = \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{\hat{\mathbf{S}}|\mathbf{X}}\left[\mathcal{B}(\hat{\mathbf{S}}, \mathbf{X}) \nabla_{\Theta_{\mathcal{M}}} \ln q(\hat{\mathbf{S}}|\mathbf{X}, \Theta_{\mathcal{M}})\right]\right] \tag{4.53}$$

ここで $\mathbf{X}$ に関する期待値を $\mathcal{I}$ 個の発話データに関する音源強調の結果の平均値に，$\hat{\mathbf{S}}$ に関する期待値を K 回のサンプリングに関する平均値に置き換えることで，式 (4.53) を以

下のように近似計算する．

$$\nabla_{\Theta_{\mathcal{M}}}\mathcal{J}(\Theta_{\mathcal{M}}) \approx \frac{1}{\mathcal{I}}\sum_{\tau=1}^{\mathcal{I}}\frac{1}{K}\sum_{k=1}^{K}\mathcal{B}(\hat{\mathbf{S}}^{(i,k)},\mathbf{X}^{(i)})\nabla_{\Theta_{\mathcal{M}}}\ln q(\hat{\mathbf{S}}^{(i,k)}|\mathbf{X}^{(i)},\Theta_{\mathcal{M}}) \tag{4.54}$$

ここで，出力音は各時間フレームごとに独立に求められることを仮定すると $\ln q(\hat{\mathbf{S}}|\mathbf{X},\Theta_{\mathcal{M}})$ は以下のように書き換えられる．

$$\ln q(\hat{\mathbf{S}}|\mathbf{X},\Theta_{\mathcal{M}}) = \sum_{\tau=1}^{T}\ln q(\hat{\boldsymbol{S}}_{\tau}|\boldsymbol{X}_{\tau},\Theta_{\mathcal{M}}) \tag{4.55}$$

またその勾配は以下のように求められる．

$$\nabla_{\Theta_{\mathcal{M}}}\ln q\left(\hat{\mathbf{S}}^{(i,k)}|\mathbf{X}^{(i)},\Theta_{\mathcal{M}}\right) = \sum_{\tau=1}^{T^{(i)}}\nabla_{\Theta_{\mathcal{M}}}\ln q(\hat{\boldsymbol{S}}_{\tau}^{(i,k)}|\boldsymbol{X}_{\tau}^{(i)},\Theta_{\mathcal{M}}) \tag{4.56}$$

$$\approx \frac{1}{T^{(i)}}\sum_{\tau=1}^{T^{(i)}}\nabla_{\Theta_{\mathcal{M}}}\ln q(\hat{\boldsymbol{S}}_{\tau}^{(i,k)}|\boldsymbol{X}_{\tau}^{(i)},\Theta_{\mathcal{M}}) \tag{4.57}$$

ただし，各発話ごとの時間フレーム数 $T^{(i)}$ を正規化するために，元の勾配に $1/T^{(i)}$ を乗じた．すると，式 (4.36) は式 (4.54)(4.57) を用いて求められる．

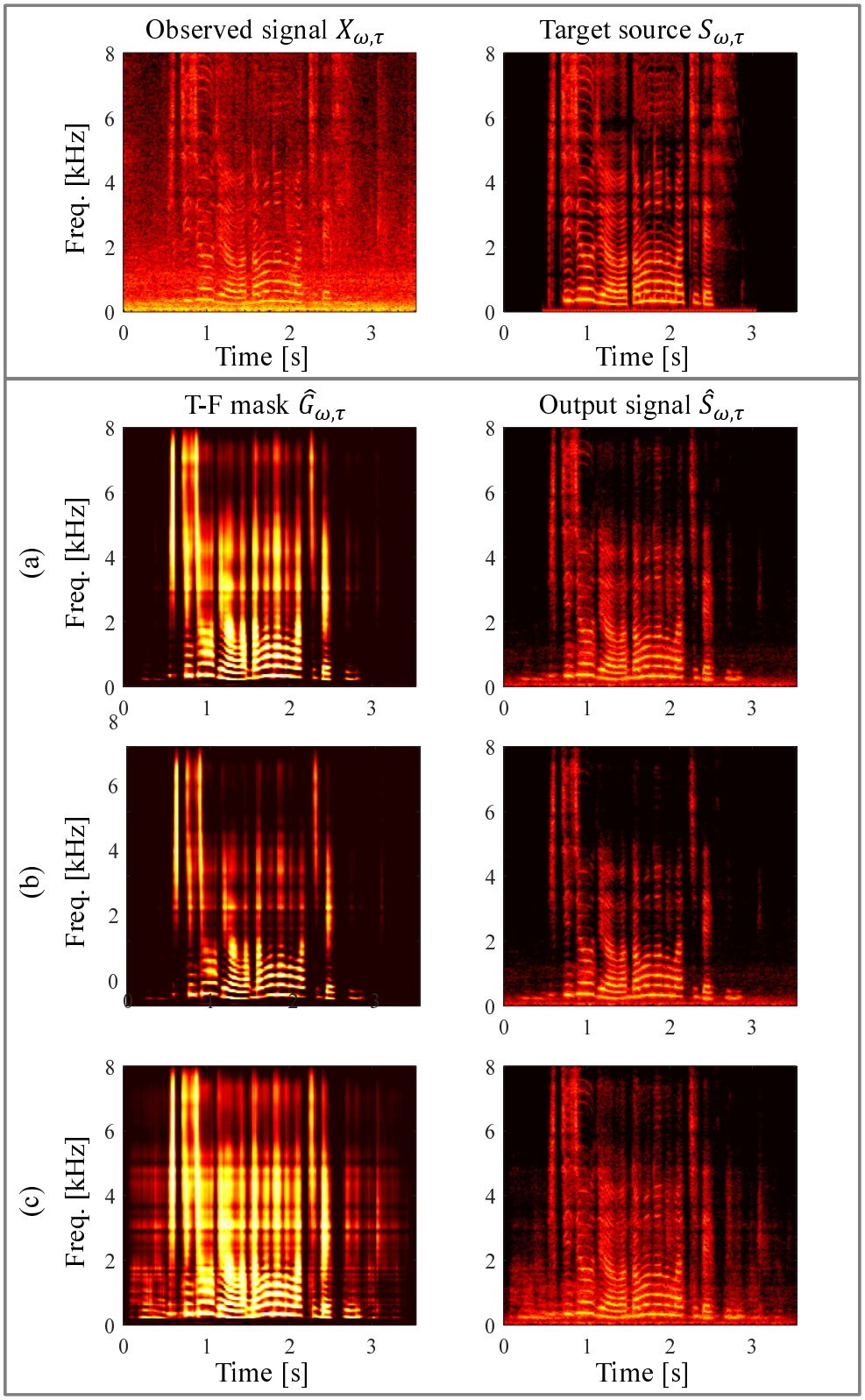


図 4.11: 推定された時間周波数マスクの例．上左図は観測信号 $X_{\omega,\tau}$，上右図は目的音 $S_{\omega,\tau}$ のスペクトログラムを表す．下左図は推定された時間周波数マスク $\hat{G}_{\omega,\tau}$，下右図は出力音 $\hat{S}_{\omega,\tau}$ のスペクトログラムを表し，(a) は最尤推定法，(b) は聴感評点に PESQ を利用した提案法，(c) は聴感評点に STOI を利用した提案法の結果を表す．

第 5 章

異常音検知の音響特徴量抽出のための目的関数

モーターの異常回転音やベアリングのぶつかり音などの普段発生しない音（異常音）を検知し，機器動作の状態が正常か異常かを判定することで機器の故障を検知する「異常音検知」を目指す．多くの異常音検知で採用されている外れ値検知 [53] に基づく異常音検知では，正常音が従う確率分布と統計的に差異がある音を異常音と仮定する．本章では，図 2.3 に示したように，ニューラルネットワークを，フレーム結合した観測信号を入力とし，音響特徴量を出力とするような音響特徴量抽出関数として用いる（図 5.1 上）．以降では，外れ値検知に基づく異常音検知の異常検知精度を最大化するために，音響特徴量を抽出するニューラルネットワークの学習のための目的関数を提案する．外れ値検出に基づく異常音検知を仮説検定とみなし，仮説検定の最適化基準であるネイマン・ピアソンの補題 [147] から，ニューラルネットワークを学習するための目的関数である “ネイマン・ピアソン指標” を導出する．ネイマン・ピアソン指標による学習の実装例として，変分オートエンコーダを応用して異常音データを疑似生成する手法を提案する（図 5.1 下）．

5.1　ネイマン・ピアソン指標

本節では，異常音検知のための音響特徴量を抽出するニューラルネットワークを学習するための目的関数である “ネイマン・ピアソン指標” を導出する．5.1.1 節では，外れ値検出に基づく異常音検知を仮説検定とみなし，異常音検知の音響特徴量が満たすべき性質である，ネイマン・ピアソン指標を導出する．5.1.2 節では，ネイマン・ピアソン指標を学習データから最適化可能な形へ具体化する．5.1.3 節では，具体化したネイマン・ピアソン指標を用いてニューラルネットワークを安定して学習するための，異常音データの疑似生成アルゴリズムを提案する．そして 5.1.4 節では，変分オートエンコーダを用いた実装方法を説明し，5.1.5 節で具体的な学習アルゴリズムを説明する．

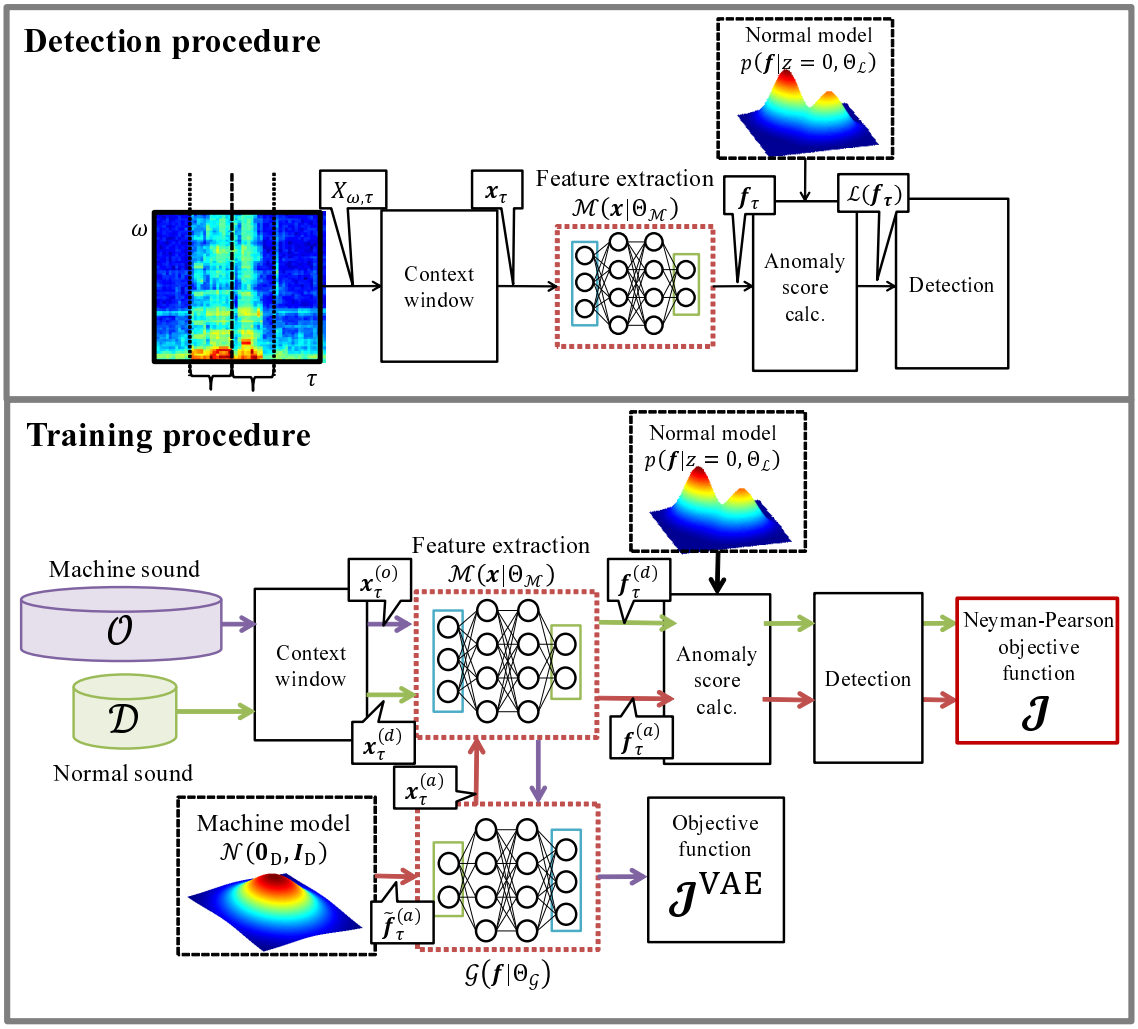


図 5.1: 提案法の概要．異常音検知のフロー（上）と変分オートエンコーダを応用したネイマン・ピアソン指標による学習の実装（下）．下図において，紫矢印が機械音の処理フロー，緑矢印が正常音の処理フロー，赤線が異常音の処理フローを表す．

5.1.1 異常音検知の音響特徴量が満たすべき性質

2.2 節で説明した外れ値検出に基づく異常音検知の式 (2.26) と式 (2.28) を変形すると，$\boldsymbol{x}_\tau$ が以下の不等式を満たす時，観測信号は異常音と判定される．

$$p(\mathcal{M}(\boldsymbol{x}_\tau|\Theta_{\mathcal{M}})|z=0,\Theta_{\mathcal{L}}) < \exp(-\phi) \tag{5.1}$$

式中の各記号を再掲すると，$\mathcal{M}$ は音響特徴量抽出関数，$\Theta_{\mathcal{M}}$ はそのパラメータ，$\Theta_{\mathcal{L}}$ は正常モデルのパラメータ，ϕ は異常判定閾値である．この式の意味を定性的にとらえる．閾値 ϕ が十分に大きな値をとると仮定すると，式 (5.1) の右辺は非常に小さな値となる．これはつまり，異常音の定義を，“抽出された音響特徴量 $\mathcal{M}(\boldsymbol{x}_\tau|\Theta_{\mathcal{M}})$ が，正常モデル $p(\mathcal{M}(\boldsymbol{x}_\tau|\Theta_{\mathcal{M}})|z=0,\Theta_{\mathcal{L}})$ から生成されたサンプルである確率が非常に小さい音” としているといえる．つまり，外れ値検出に基づく異常音検知は，帰無仮説と対立仮説を以下

とした，仮説検定の一種であるとみなせる．

帰無仮説: 抽出された音響特徴量 $\mathcal{M}(\boldsymbol{x}_\tau|\Theta_\mathcal{M})$ は，正常モデル $p(\mathcal{M}(\boldsymbol{x}_\tau|\Theta_\mathcal{M})|z=0,\Theta_\mathcal{L})$ から生成されたサンプルである．

対立仮説: 抽出された音響特徴量 $\mathcal{M}(\boldsymbol{x}_\tau|\Theta_\mathcal{M})$ は，正常モデル $p(\mathcal{M}(\boldsymbol{x}_\tau|\Theta_\mathcal{M})|z=0,\Theta_\mathcal{L})$ から生成されたサンプルでない．

ゆえに，異常音検知の音響特徴量抽出関数の目的関数は，仮説検定関数の最適化指標から導出できると考えた．

ネイマン・ピアソンの補題 [147] は，異常音検知のような単純仮説検定において，有意水準 ρ となる検定の中で検出力を最大化する検定が満たす性質を示す定理である．この性質とは，偽陽性率 (FPR: false positive rate) を ρ と固定したときに，真陽性率 (TPR: true positive rate) を最大にすることある．異常音検知において，TPR と FPR は以下の式で計算できる．

$$\mathrm{TPR}(\Theta_\mathcal{M},\phi)=\mathbb{E}\left[\mathcal{H}(\mathcal{L}(\mathcal{M}(\boldsymbol{x}_\tau|\Theta_\mathcal{M})),\phi)\right]_{\boldsymbol{x}|z\neq 0} \tag{5.2}$$

$$\mathrm{FPR}(\Theta_\mathcal{M},\phi)=\mathbb{E}\left[\mathcal{H}(\mathcal{L}(\mathcal{M}(\boldsymbol{x}_\tau|\Theta_\mathcal{M})),\phi)\right]_{\boldsymbol{x}|z=0} \tag{5.3}$$

ただし $\mathbb{E}[\cdot]_{\boldsymbol{x}}$ は $\boldsymbol{x}$ に関する期待値演算を表す．いま，ϕ_ρ が $\mathrm{FPR}(\Theta_\mathcal{M},\phi_\rho)=\rho$ を満たす閾値とする．すると検出力を最大化する検知器は，以下の値を最大化する検知器となる．

$$\mathrm{TPR}(\Theta_\mathcal{M},\phi_\rho)+\left\{\rho-\mathrm{FPR}(\Theta_\mathcal{M},\phi_\rho)\right\}. \tag{5.4}$$

ここで問題の簡単のために，ϕ_ρ を $\Theta_\mathcal{M}$ に関係のない定数とすると，$\Theta_\mathcal{M}$ を最適化する目的関数は以下のように書ける．

$$\mathcal{J}=\mathrm{TPR}(\Theta_\mathcal{M},\phi_\rho)-\mathrm{FPR}(\Theta_\mathcal{M},\phi_\rho) \tag{5.5}$$

上記の目的関数は，外れ値検出に基づく異常音検知を仮説検定とみなしてネイマン・ピアソンの補題から導出したものであるため，“ネイマン・ピアソン指標” と名付けることにする．ネイマン・ピアソン指標は，“異常音検知の精度を最大化する音響特徴量の満たすべき性質” は “FPR を ρ とする制約のもとで，TPR を最大化する” ことを意味する目的関数である．

5.1.2 ネイマン・ピアソン指標の具現化

ネイマン・ピアソン指標に基づき，音響特徴量を抽出するニューラルネットワークを学習するために，ネイマン・ピアソン指標を微分可能かつ，学習データから最大化可能な目的関数へ変形する．このような変形を行うことで，$\Theta_\mathcal{M}$ は勾配法などの非線形最適

化により学習可能となる．まず，$p(\mathcal{M}(\boldsymbol{x}_\tau|\Theta_\mathcal{M})|z=0)$ は，$\Theta_\mathcal{M}$ について微分可能な形で実装されていることを仮定する．この仮定を満たす実装には，たとえば式 (2.27) の混合ガウス分布（GMM: Gaussian mixture model）などがある．次に，$\Theta_\mathcal{M}$ について微分不可能なステップ関数である $\mathcal{H}(\mathcal{L}(\boldsymbol{x}_\tau),\phi)$ を，シグモイド関数を用いて微分可能な形に近似する．

$$\tilde{\mathcal{H}}(\mathcal{L}(\mathcal{M}(\boldsymbol{x}|\Theta_\mathcal{M}))),\phi) = \frac{1}{1+\exp\{\mathcal{L}(\mathcal{M}(\boldsymbol{x}|\Theta_\mathcal{M}))-\phi\}} \tag{5.6}$$

また，$\Theta_\mathcal{M}$ を学習データから最適化するために，TPR と FPR の算出に用いる期待値演算を，正常音の学習データ $\boldsymbol{x}_k^{(d)}$ と異常音の学習データ $\boldsymbol{x}_k^{(a)}$

$$\mathcal{D} = \{\boldsymbol{x}_k^{(d)} \in \mathbb{R}^\mathsf{Q} | k=1,...,K_d\} \tag{5.7}$$

$$\mathcal{A} = \{\boldsymbol{x}_k^{(a)} \in \mathbb{R}^\mathsf{Q} | k=1,...,K_a\} \tag{5.8}$$

の算術平均で近似する．ただし，$\mathsf{Q}=\Omega\times(P_b+P_f+1)$ であり K_d, K_a は，正常音の学習データと異常音の学習データの総サンプル数を表す．また，$\mathrm{FPR}(\Theta_\mathcal{M},\phi_\rho)=\rho$ を満たすために閾値 ϕ_ρ は，正常音の学習データ $\mathcal{D}$ から求めた異常度を降順ソートしたものの $\lfloor\rho K_d\rfloor$ 番目の値とする．ただし $\lfloor\cdot\rfloor$ は床関数とする．すると，式 (5.5) $\mathcal{M}$ のパラメータについて微分可能な形として以下のように記述できる．

$$\mathcal{J} = \frac{1}{K_a}\sum_{k=1}^{K_a}\tilde{\mathcal{H}}(\mathcal{L}(\mathcal{M}(\boldsymbol{x}_k^{(a)}|\Theta_\mathcal{M})),\phi_\rho) - \frac{1}{K_d}\sum_{k=1}^{K_d}\tilde{\mathcal{H}}(\mathcal{L}(\mathcal{M}(\boldsymbol{x}_k^{(d)}|\Theta_\mathcal{M})),\phi_\rho) \tag{5.9}$$

つまり式 (5.9) が異常音検知の音響特徴量抽出関数を最適化するための目的関数である．

5.1.3 異常音データの疑似生成

異常音のデータを収集することは困難なため，式 (5.9) のように単に期待値演算を算術平均で近似しては TPR を精度よく近似計算することはできない．ゆえに，$\Theta_\mathcal{M}$ の最適化に影響を及ぼすことが想定される．期待値演算の近似精度を向上させるために，異常音データを疑似生成する．

外れ値検出に基づく異常音検知では，異常音は “抽出された音響特徴量 $\mathcal{M}(\boldsymbol{x}_\tau|\Theta_\mathcal{M})$ が，正常モデル $p(\mathcal{M}(\boldsymbol{x}|\Theta_\mathcal{M})|z=0,\Theta_\mathcal{L})$ から生成されたサンプルである確率が非常に小さい音” と定義している．つまり異常音とは，“機器動作音ではあるが監視対象機器の正常音とは異なる音” と考えることもできる．本節では，この定義を用いて異常音の音響特徴量を疑似生成する．まず監視対象機器の正常音か否かを示す確率変数である確率変数 z を周辺化し，あらゆる機器の正常な機器動作音や異常な機器動作音

$$\mathcal{O} = \{\boldsymbol{x}_k^{(o)} \in \mathbb{R}^\mathsf{Q} | k=1,...,K_o\}. \tag{5.10}$$

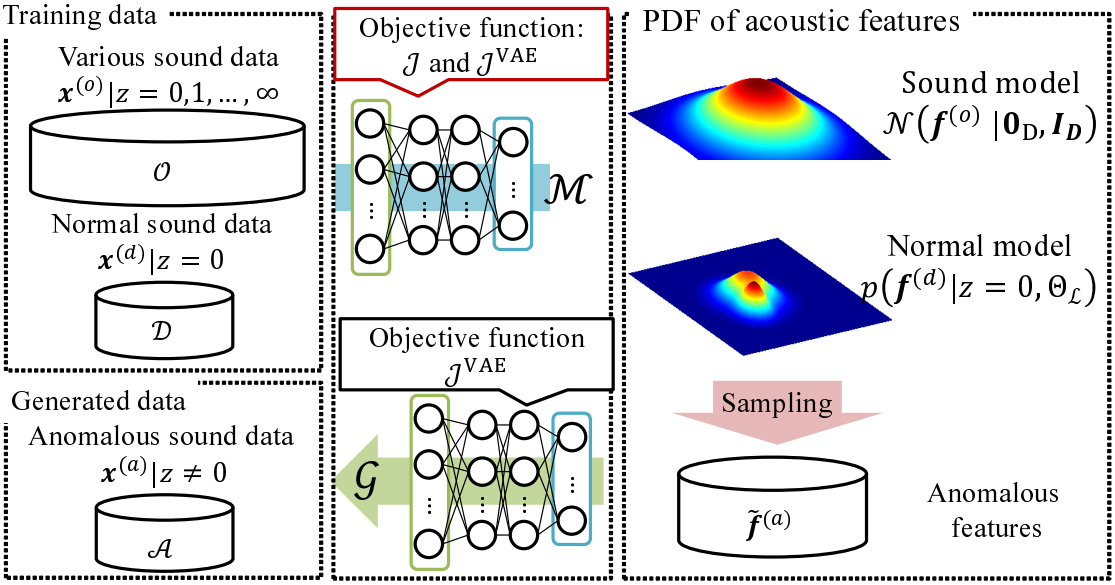


図 5.2: 変分オートエンコーダを用いた提案法の実装.

Algorithm 3 異常音データの音響特徴量の疑似生成アルゴリズム

Input: $p(\mathcal{M}(\boldsymbol{x}|\Theta_{\mathcal{M}})|z = 0, \Theta_{\mathcal{L}})$, and ϕ_ρ

Output: $\boldsymbol{f}_k^{(a)}$

while $\mathcal{L}(\tilde{\boldsymbol{f}}_k^{(a)}) \leq \phi_\rho$ **do**

$\tilde{\boldsymbol{f}}_k^{(a)} \sim \mathcal{N}(\mathbf{0}_{\mathrm{D}}, \boldsymbol{I}_{\mathrm{D}})$

end while

から抽出した音響特徴量の従う確率密度関数 $p(\mathcal{M}(\boldsymbol{x}|\Theta_{\mathcal{M}}))$ を導入する．ここで，あらゆる機器の正常な機器動作音や異常な機器動作音とは，監視対象機器が設置されている場所以外で収録した大量かつ様々な機器動作音であり，$K_o \gg K_d$ である．そして $p(\mathcal{M}(\boldsymbol{x}|\Theta_{\mathcal{M}}))$ から乱数生成し，$p(\mathcal{M}(\boldsymbol{x}|\Theta_{\mathcal{M}})|z = 0, \Theta_{\mathcal{L}})$ を用いてそのデータが正常音の音響特徴量である尤度を計算する．そして，その尤度が一定値以下であったなら，そのデータを異常音の音響特徴量 $\tilde{\boldsymbol{f}}_k^{(a)}$ として採用する．最後に，音響特徴量抽出関数の逆関数 $\mathcal{G}$ を用いて異常音データを疑似生成する．

$$\boldsymbol{x}_k^{(a)} \leftarrow \mathcal{G}(\tilde{\boldsymbol{f}}_k^{(a)}|\Theta_{\mathcal{G}}) \tag{5.11}$$

次節では，以上のアルゴリズムの実現方法として，変分オートエンコーダを利用した実装法を説明する．

5.1.4 変分オートエンコーダを用いた実装

異常音の音響特徴量 $\tilde{\boldsymbol{f}}_k^{(a)}$ の生成手順を簡素化し，さらに音響特徴量抽出関数の逆関数 $\mathcal{G}$ を精度よく求めるために，$\mathcal{M}$ と $\mathcal{G}$ の実装に，変分オートエンコーダを応用する [94].

図 5.2 に提案法の概要を示す．本実装では $p(\mathcal{M}(\boldsymbol{x}|\Theta_{\mathcal{M}}))$ からの乱数生成を容易にするために，$p(\mathcal{M}(\boldsymbol{x}|\Theta_{\mathcal{M}}))$ が $\mathcal{N}(\boldsymbol{f}^{(o)}|\mathbf{0}_{\mathsf{D}}, \boldsymbol{I}_{\mathsf{D}})$ となるように変分オートエンコーダを学習する．ここで $\mathbf{0}_{\mathsf{D}}$ と $\boldsymbol{I}_{\mathsf{D}}$ はそれぞれ，D のゼロベクトルと単位行列である．$\mathcal{N}(\boldsymbol{f}^{(o)}|\mathbf{0}_{\mathsf{D}}, \boldsymbol{I}_{\mathsf{D}})$ からの乱数生成には一般的な正規乱数発生器を利用できるため，異常音の音響特徴量 $\tilde{\boldsymbol{f}}_k^{(a)}$ の疑似生成は **Algorithm 3** のように容易に実行できるようになる．ただし $\sim$ は右辺の確率密度関数からの乱数生成である．そして $\Theta_{\mathcal{M}}, \Theta_{\mathcal{G}}$ は，式 (5.9) のネイマン・ピアソン指標 $\mathcal{J}$ と，以下で説明する変分オートエンコーダの目的関数 $\mathcal{J}^{\mathrm{VAE}}$ を交互に最大化するように学習する．

まず，$\mathcal{M}(\boldsymbol{x}|\Theta_{\mathcal{M}})$ の出力を多次元ガウス分布パラメータである平均ベクトルと分散ベクトルとする．つまり $\mathcal{M}(\boldsymbol{x}|\Theta_{\mathcal{M}})$ は観測信号を得た下での音響特徴量の事後分布のパラメータを出力しているとみなし，

$$\begin{aligned}
\boldsymbol{\mu}(\boldsymbol{x}_k) &= (\mu_{1,k}, ..., \mu_{\mathsf{D},k})^\top \\
&= \mathbf{W}^{(\mu)}\boldsymbol{u}_\tau^{(L-1)} + \mathbf{b}^{(\mu)} && (5.12)\\
\boldsymbol{\sigma}(\boldsymbol{x}_k) &= \left(\ln \sigma_{1,k}^2, ..., \ln \sigma_{\mathsf{D},k}^2\right) \\
&= \mathbf{W}^{(\sigma)}\boldsymbol{u}_\tau^{(L-1)} + \mathbf{b}^{(\sigma)} && (5.13)\\
\boldsymbol{z}_k^{(l)} &= \sigma_\theta\left\{\boldsymbol{u}_k^{(l)}\right\} && (5.14)\\
\boldsymbol{u}_k^{(l)} &= \mathbf{W}^{(l)}\boldsymbol{z}_k^{(l-1)} + \mathbf{b}^{(l)} && (5.15)\\
\Theta_{\mathcal{M}} &= \left\{\mathbf{W}^{(l)}, \mathbf{b}^{(l)}, \mathbf{W}^{(\mu)}, \mathbf{b}^{(\mu)}, \mathbf{W}^{(\sigma)}, \mathbf{b}^{(\sigma)} | l = 2, ..., L-1\right\} && (5.16)
\end{aligned}$$

のように実装する．ただし $\boldsymbol{z}_k^{(1)} = \boldsymbol{x}_k$ であり，また $\mathcal{M}$ を音響特徴量抽出関数として用いる場合は，$\boldsymbol{f}_k = \boldsymbol{\mu}(\boldsymbol{x}_k)$ とする．そして音響特徴量抽出関数の逆関数 $\mathcal{G}$ は，$\mathcal{M}$ の逆向きの構造を持ったニューラルネットワークとして，以下のように実装する．

$$\begin{aligned}
\hat{\boldsymbol{x}}_k &= \mathbf{W}^{(L,\mathcal{G})}\boldsymbol{u}_\tau^{(L-1,\mathcal{G})} + \mathbf{b}^{(L,\mathcal{G})} && (5.17)\\
\boldsymbol{z}_k^{(l,\mathcal{G})} &= \sigma_\theta\left\{\boldsymbol{u}_k^{(l,\mathcal{G})}\right\} && (5.18)\\
\boldsymbol{u}_k^{(l,\mathcal{G})} &= \mathbf{W}^{(l,\mathcal{G})}\boldsymbol{z}_k^{(l-1,\mathcal{G})} + \mathbf{b}^{(l,\mathcal{G})} && (5.19)\\
\Theta_{\mathcal{G}} &= \left\{\mathbf{W}^{(l,\mathcal{G})}, \mathbf{b}^{(l,\mathcal{G})} | l = 2, ..., L\right\} && (5.20)
\end{aligned}$$

ただし $\boldsymbol{z}_k^{(1,\mathcal{G})} = \boldsymbol{f}_k$ である．

$\mathcal{M}(\boldsymbol{x}|\Theta_{\mathcal{M}})$ の出力を多次元ガウス分布パラメータとした場合，変分オートエンコーダの目的関数は再構成誤差 $\boldsymbol{E}$ と $\mathcal{M}(\boldsymbol{x}|\Theta_{\mathcal{M}})$ の出力と $\mathcal{N}(\boldsymbol{f}|\mathbf{0}_{\mathsf{D}}, \boldsymbol{I}_{\mathsf{D}})$ の KL ダイバージェンスを同時に最小化することになる．本研究では，$p(\mathcal{M}(\boldsymbol{x}|\Theta_{\mathcal{M}}))$ が $\mathcal{N}(\boldsymbol{f}^{(o)}|\mathbf{0}_{\mathsf{D}}, \boldsymbol{I}_{\mathsf{D}})$ となるように変分オートエンコーダを学習するために，あらゆる機器の正常な機器動作音や異常な機器動作音 $\mathcal{O}$ を用いて $\Theta_{\mathcal{M}}, \Theta_{\mathcal{G}}$ を学習する．すなわち式 (2.56) における目的関数以下

Algorithm 4 提案法の学習アルゴリズム

Input: $\mathcal{D}$ and $\mathcal{O}$

Output: $\Theta_{\mathcal{M}}$ and $\Theta_{\mathcal{L}}$

Initialize $\Theta_{\mathcal{M}}$, $\Theta_{\mathcal{G}}$ and $\Theta_{\mathcal{L}}$

while *until algorithm convergence* **do**

$\boldsymbol{x}^{(d)}_{1,\dots,K}, \boldsymbol{x}^{(o)}_{1,\dots,K} \leftarrow$ Random draw K samples from $\mathcal{D}$ and $\mathcal{O}$

$\Theta_{\mathcal{M}}, \Theta_{\mathcal{G}} \leftarrow$ Maximize $\mathcal{J}^{\mathrm{VAE}}$ using $\boldsymbol{x}^{(o)}_{1,\dots,K}$

$\boldsymbol{f}^{(d)}_{1,\dots,K_d} \leftarrow \mathcal{M}(\boldsymbol{x}^{(d)}_{1,\dots,K_d}|\Theta_{\mathcal{M}})$

$\Theta_{\mathcal{L}} \leftarrow$ EM-algorithm using $\boldsymbol{f}^{(d)}_{1,\dots,K_d}$

$\phi_\rho \leftarrow \lfloor \rho K_d \rfloor^{\mathrm{th}}$ value of descend sorted $\mathcal{L}(\boldsymbol{f}^{(d)}_{1,\dots,K_d})$

$\boldsymbol{x}^{(a)}_{1,\dots,K} \leftarrow$ Generate K samples using **Algorithm 3** and $\mathcal{G}$

$\Theta_{\mathcal{M}} \leftarrow$ Maximize $\mathcal{J}$ using $\boldsymbol{x}^{(d)}_{1,\dots,K}$ and $\boldsymbol{x}^{(a)}_{1,\dots,K}$

end while

のように具現化する．

$$\mathcal{J}^{\mathrm{VAE}} = -\boldsymbol{E}(\mathcal{O}, \Theta_{\mathcal{M}}, \Theta_{\mathcal{G}}) - \mathrm{KLD}(\mathcal{O}, \Theta_{\mathcal{M}}, \Theta_{\mathcal{G}}) \tag{5.21}$$

ここで再構成誤差は

$$\boldsymbol{E}(\mathcal{O}, \Theta_{\mathcal{M}}, \Theta_{\mathcal{G}}) = \sum_{k_o=1}^{K_o} ||\mathcal{G}(\boldsymbol{\zeta}^{(o)}_k) - \boldsymbol{x}^{(o)}_k||^2 \tag{5.22}$$

$$\boldsymbol{\zeta}^{(o)}_k = \boldsymbol{\mu}(\boldsymbol{x}^{(o)}_k) + \boldsymbol{\sigma}(\boldsymbol{x}^{(o)}_k) \odot \boldsymbol{\epsilon}_k \tag{5.23}$$

$$\boldsymbol{\epsilon}_k \sim \mathcal{N}(\boldsymbol{\epsilon}_k|\mathbf{0}_{\mathsf{D}}, \boldsymbol{I}_{\mathsf{D}}) \tag{5.24}$$

として求め，KL ダイバージェンスを

$$\mathrm{KLD}(\mathcal{O}, \Theta_{\mathcal{M}}, \Theta_{\mathcal{G}}) = \frac{1}{2}\sum_{k=1}^{K_o}\sum_{d=1}^{D}\left(1 + \ln((\sigma^{(o)}_{k,d})^2) - (f^{(o)}_{k,d})^2 - (\sigma^{(o)}_{k,d})^2\right) \tag{5.25}$$

のように求める．ただし $\odot$ は要素積を表す．

5.1.5 学習アルゴリズム

本節では，提案法の詳細な実行手順を **Algorithm 4** に沿って説明する．アルゴリズムへの入力は正常音の学習データ $\mathcal{D}$ と様々な機器動作音データ $\mathcal{O}$ であり，出力は音響特徴量を抽出するニューラルネットワークのパラメータ $\Theta_{\mathcal{M}}$ と正常モデルのパラメータ $\Theta_{\mathcal{L}}$ である．なお以下では学習アルゴリズムを簡潔に記述するために，音響特徴量を抽出するニューラルネットワークは DNN で実装し，正常モデルは GMM で実装することを前提に説明をする．

まず学習データから，正常音のミニバッチ $\boldsymbol{x}^{(d)}_{1,\dots,K}$ および様々な機器動作音のミニバッチ $\boldsymbol{x}^{(o)}_{1,\dots,K}$ を K サンプル，学習データセット $\mathcal{D}$，$\mathcal{O}$ からランダムに取り出す．次に $\boldsymbol{x}^{(o)}_{1,\dots,K}$ を用いて，$\mathcal{J}^{\mathrm{VAE}}$ を増加させるように，$\Theta_{\mathcal{M}}, \Theta_{\mathcal{G}}$ を勾配法で 1 ステップ更新する[1]．そして正常モデルのパラメータ $\Theta_{\mathcal{L}}$ を，正常音の学習データ $\mathcal{D}$ の全てのデータから抽出した音響特徴量 $\boldsymbol{f}^{(d)}_{1,\dots,K_d}$ を用いて，期待値最大化（EM: expectation-maximization）アルゴリズムで更新する．次いで FPR が ρ となるように異常判定閾値 ϕ_ρ を更新するために，正常音の学習データ $\mathcal{D}$ の全てのデータから計算した異常度 $\mathcal{L}(\boldsymbol{f}^{(d)}_{1,\dots,K_d})$ を降順ソートし，$\lfloor \rho K_d \rfloor$ 番目の異常度を ϕ_ρ に設定する．最後に異常音データ $\boldsymbol{x}^{(d)}_{1,\dots,K}$ を **Algorithm 3** と式 (5.11) で K サンプル疑似生成し，$\mathcal{J}$ を増加させるように $\boldsymbol{x}^{(d)}_{1,\dots,K}$ と $\boldsymbol{x}^{(a)}_{1,\dots,K}$ を用いて $\Theta_{\mathcal{M}}$ を勾配法で 1 ステップ更新する．

5.2 評価実験

提案法 (PROP) の有効性を示すために，定量評価実験および実環境動作実験を行った．定量評価試験において，提案法は以下の 3 つの従来法と比較した．

AE： $\Theta_{\mathcal{M}}, \Theta_{\mathcal{G}}$ を二乗誤差最小化基準で学習．

VAE： $\Theta_{\mathcal{M}}, \Theta_{\mathcal{G}}$ を変分下界 $\mathcal{J}^{\mathrm{VAE}}$ 最大化基準で学習．

CE − VAE： $\Theta_{\mathcal{M}}$ の学習に，ネイマン・ピアソン指標の代わりに交差エントロピーを用いる．与えられた音響特徴量が正常音から抽出されたものか，$\mathcal{N}(\boldsymbol{f}|\boldsymbol{0}_{\mathrm{D}}, \boldsymbol{I}_{\mathrm{D}})$ からのサンプル（異常音）かの識別（2 クラス分類）を行うニューラルネットワークを用意し，このニューラルネットワークと $\Theta_{\mathcal{M}}$ を交差エントロピー最小化基準で学習する．出力層の活性化関数を softmax とし，正常音の事後確率に負の対数尤度をとったものを異常度として異常検知を行う．

AE および VAE は，異常検知結果からのフィードバックのない従来の音響特徴量抽出法である．これらと比べ PROP の性能が向上するならば，提案法のような，異常検知結果からのフィードバックのある学習方式が有効といえる．CE − VAE は提案法と同じ，異常検知結果からのフィードバックのある学習方式であるが，目的関数がネイマン・ピアソン指標ではなく交差エントロピーとなっている．これと比べ PROP の性能が向上するならば，異常音検知の音響特徴量抽出の学習にはネイマン・ピアソン指標が有効であるといえる．

[1] 正常音データのデータ数が多い場合，$\Theta_{\mathcal{L}}$ の更新は $\mathcal{D}$ の一部のデータのみから行ってもよい．また，勾配法のステップサイズを小さく設定する場合は，勾配法の毎ステップごとに $\Theta_{\mathcal{L}}$ を更新しなくてもよい．

5.2.1 実験条件

音響特徴量の次元数は $\mathsf{D} = 32$ とし，フレーム結合サイズは $P_b = P_f = 10$ とした．全ての手法において $\mathcal{M}$ は隠れ層数 3，隠れユニット数 512，活性化関数がランプ関数（ReLU: rectified linear unit）の DNN で実装した．また $\mathcal{G}$ は $\mathcal{M}$ の逆向きの DNN 構造とした．$\mathtt{CE-VAE}$ の識別のネットワーク構造は正常の確率と異常の確率を出力するために，隠れ層数 3，隠れユニット数 512，活性化関数が ReLU，出力層がユニット数 2 で活性化関数が softmax の DNN で実装した．DNN の入出力の次元数を抑えるために，$\boldsymbol{X}_\tau$ と $\boldsymbol{G}_\tau$ は $\mathsf{B} = 64$ のメルフィルタバンクで圧縮し，時間周波数マスク設計の際にスプライン補間で線形周波数に補間した．すなわち，$\boldsymbol{x}$ の次元数は $\mathsf{Q} = 64 \times (P_b + P_f + 1) = 1344$ となる．勾配法のアルゴリズムには Adam [84] を用いた．また過適合を防ぐために正則化パラメータが $\lambda = 10^{-5}$ の L_2 正則化と，入力層のドロップ確率が 0.2，隠れ層のドロップ確率が 0.5 のドロップアウトを利用した．ミニバッチサイズは $K = 100$ とし，500 が終了した段階で学習を終了した．FPR の設定値は実験的に $\rho = 0.05$ とした．正常モデルに用いた GMM の混合数は $\mathsf{C} = 16$ とし，学習の安定のために共分散行列は対角行列に制限した．全てのデータは 16 kHz でサンプリングし，短時間フーリエ変換のフレームサイズは 512 サンプルとし，シフト幅は 256 サンプルとした．また，$\mathtt{AE}$，$\mathtt{VAE}$ の学習は $\mathcal{D}$ と $\mathcal{O}$ の両方のデータを用いて行った．

5.2.2 定量評価実験

実環境で異常音が含まれたデータを用意することが困難なため，人工的に生成した異常音データを用いて評価した．正常音の学習データ $\mathcal{D}$ には実環境で収集したエンジン音を利用し，また様々な機器動作音データ $\mathcal{O}$ には，正常音の学習データを収録した工場とは異なる工場で収集したエンジン音やベアリング音，ポンプ音などを利用した．$\mathcal{D}$ と $\mathcal{O}$ のデータ量はそれぞれ 1 時間と 20 時間とした．異常音データは 45 種類の音を用い，15 種類のエンジンの回転音，15 種類のエンジンの加速音， また 15 種類の金属やコンクリートのぶつかり音を用いた．これらの異常音は正常音との SNR が -10，-5，0，5 dB となるようにテストデータと重畳した．

評価尺度には，適合率 (Prec.)，再現率 (Rec.)，調和平均 (F_1) を用いた．これらの値を算出するのに用いた異常判定閾値は各値の和が最大になるものを用いた．評価結果を表 5.1 に示す．全ての SNR で，提案法の調和平均が最も高いことが見て取れる．調和平均は適合率と再現率のトレードオフを示す値であり，異常検知器の安定性を示す指標である．また，様々な閾値で異常判定制度を評価し，受信者動作特性（ROC: receiver operating characteristic）曲線として描画した結果を図 5.3 に示す．この図から，提案法の TPR はすべての FPR の条件で従来法をよりも高いことがわかり，この理由は，ネイマン・ピア

表 5.1: 実験結果.

SNR (dB)	-10 dB			-5 dB			0 dB			5 dB		
score	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1
AE	0.62	0.98	0.76	**0.94**	0.76	0.84	0.95	0.87	0.91	**0.98**	0.91	0.94
VAE	0.51	**0.94**	0.67	0.52	**1.0**	0.68	0.61	**0.98**	0.75	0.86	0.84	0.85
CE − VAE	0.51	**0.94**	0.67	0.78	0.80	0.79	0.94	0.73	0.83	0.95	0.78	0.85
PROP	**0.76**	0.91	**0.82**	0.84	0.96	**0.90**	**0.96**	0.89	**0.93**	0.92	**1.0**	**0.96**

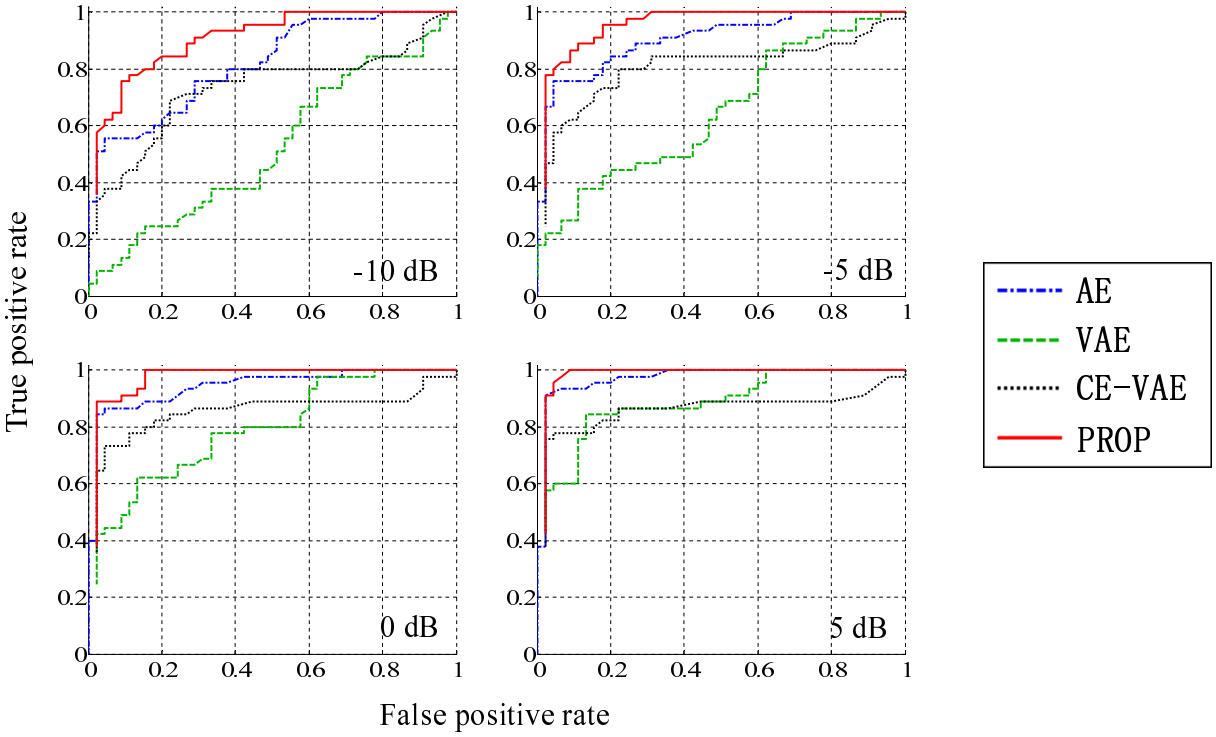


図 5.3: 各 SNR での ROC 曲線.

ソン指標が任意の FPR ρ で TPR を最大化することを要請する目的関数であるためと考えられる．これらの結果から，ネイマン・ピアソン指標により異常音検知の性能が向上することがわかり，外れ値検知に基づく異常音検知では，ネイマン・ピアソン指標が他の目的関数と比べて検知性能を向上させることを実験的に示した．

5.2.3 実環境動作実験

提案法が実環境において異常音を検知できるかを，(a) 3D プリンタ，(b) 送風ポンプ，(c) 給水ポンプの 3 つの機器動作音で評価した．

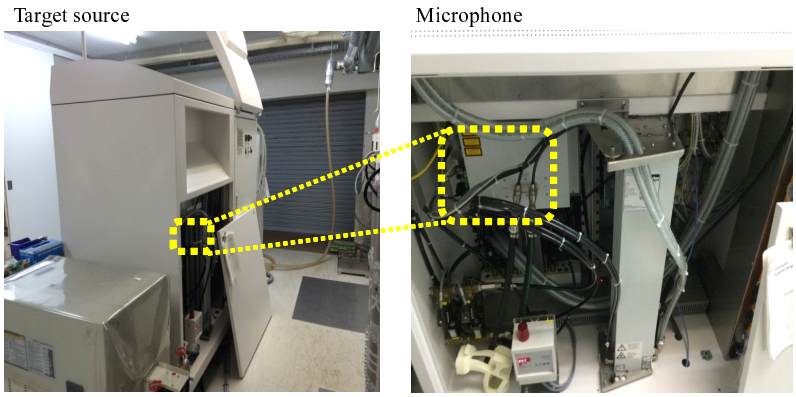


図 5.4: 3D プリンタの全体像（左）とマイクロホンの配置位置（右）.

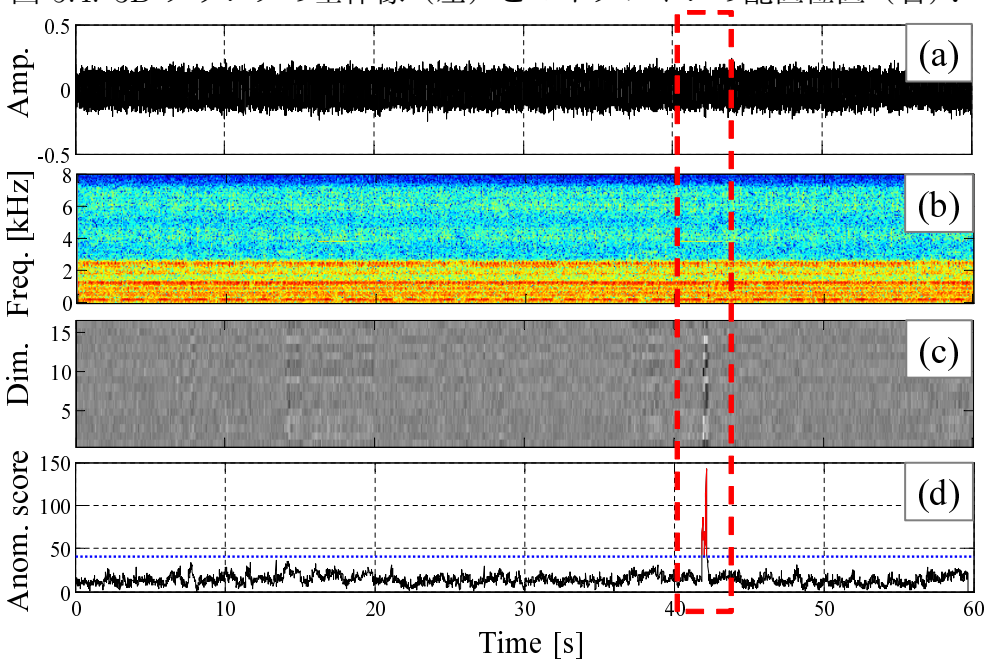


図 5.5: 3D プリンタの異常音検知結果．各図はそれぞれ，(a) 観測波形，(b) スペクトログラム，(c) 提案法により抽出された音響特徴量，(d) 異常度を表す.

(a) 3D プリンタ

図 5.4 の，光造形方式の 3D プリンタで異常音を検知できるかを評価した．マイクロホンは，3D プリンタの内部に配置した．異常音は，スイーパと造形物が衝突した音を利用した[2]．正常音データには，30 分間の 3D プリンタの正常動作音を用いた．様々な機器動作音データには，定量評価実験と同じものを用いた．正常音の学習データ量が少ないため，DNN の構造は隠れ層が 2 層，隠れユニット数が 256 の小さな構造のものを用いた．その他の実験条件は定量評価試験と同じものを用いた.

図 5.5 に検知結果を示す．43 秒付近に異常衝突音が発生しているものの，非常に小さな

[2]この 3D プリンタは，この衝突が原因で約 5 分後に異常停止した.

衝突音であるため波形やスペクトログラムからは異常音を見つけることができない（図 5.5 (a)(b)）．一方，提案法で抽出した音響特徴量には明確な変化が現れており（図 5.5 (c)），異常度も上昇していることがわかる（図 5.5 (d)）．

(b) 送風ポンプ

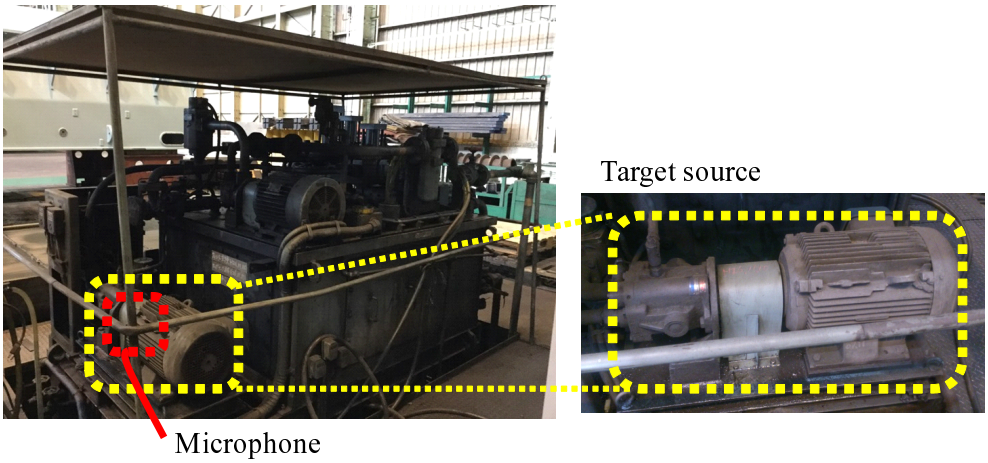


図 5.6: 送風ポンプの全体像とマイクロホンの配置位置（左），およびポンプの拡大図（右）．

図 5.6 に示す，送風ポンプで異常音を検知できるかを評価した．マイクロホンは，ポンプに隣接するポールに貼り付けて配置した．異常音は，送風ダクトに異物が混入しダクトが詰まった際の音を利用した[3]．正常音データには，20 分間の送風ポンプの正常動作音を用いた．様々な機器動作音データには，定量評価実験と同じものを用いた．正常音の学習データ量が少ないため，DNN の構造は隠れ層が 2 層，隠れユニット数が 256 の小さな構造のものを用いた．その他の実験条件は定量評価試験と同じものを用いた．

図 5.7 に検知結果を示す．5 秒付近に異常衝突音が発生している．3D プリンタの例と異なり聴感的にもはっきりと聞き取れる異常音であるためスペクトログラムからは異常音を見つけることができる（図 5.7 (b)）．提案法で抽出した音響特徴量にも明確な変化が現れており（図 5.7 (c)），異常度も上昇していることがわかる（図 5.7 (d)）．

(c) 給水ポンプ

ビルの設備である給水ポンプで異常音を検知できるかを評価した．異常音は，ベアリングの傷に起因する異常音を利用した[4]．ビルの機械室では周囲に様々な機械が雑音を発しており，対象の機器動作音のみの収音が難しい．そこで本実験では，マイクロホンを

[3]すぐに故障につながる異常ではないが，頻出する場合にはダクト内の検査が必要となる．
[4]定期検査でこういった異常音が検知されると，修理の必要があると判断される．

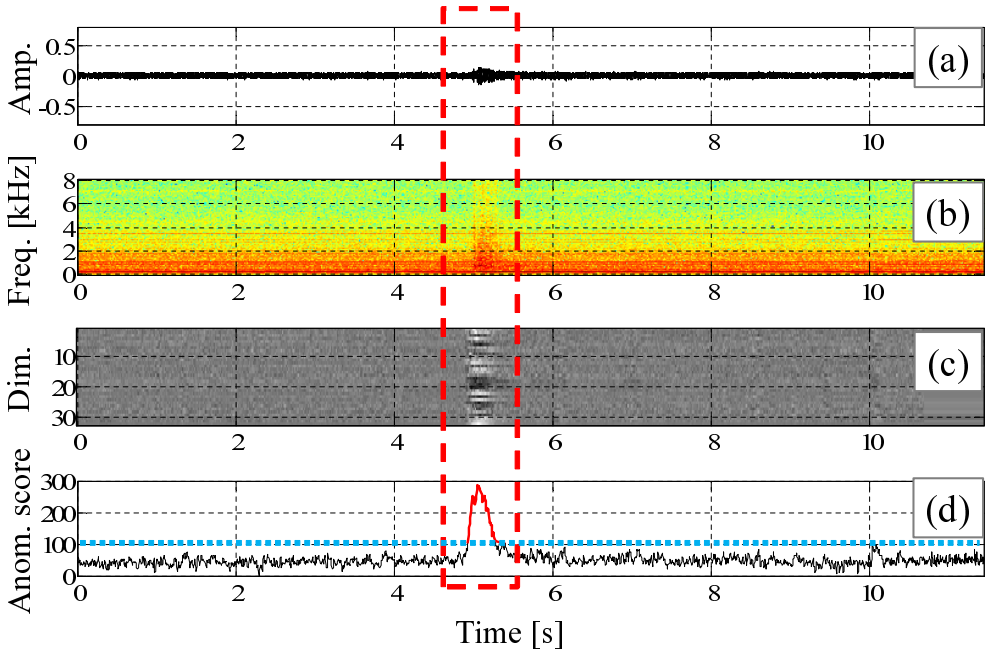


図 5.7: 送風ポンプの異常音検知結果．各図はそれぞれ，(a) 観測波形，(b) スペクトログラム，(c) 提案法により抽出された音響特徴量，(d) 異常度を表す．

図 5.8 のように配置し，付録 A に記載した音源強調法を用いて給水ポンプ音を音源強調した．また，正常動作時の音データが存在しなかったため，隣に配置してある，正常動作している同型同種の給水ポンプの稼働音を正常音データとして用いた．正常音データの長さは 2 時間とし，様々な機器動作音データには定量評価実験と同じものを用いた．音響特徴量の次元は $D = 16$ とし，その他の実験条件は定量評価試験と同じものを用いた．

図 5.9 に検知結果を示す．比較のために，前半 60 秒に正常動作中の給水ポンプの動作音，後半 60 秒に異常動作中の給水ポンプの動作音をつなげて表示している．波形やスペクトログラムからには大きな違いは見当たらないものの（図 5.9 (a)(b)），音源強調後のスペクトログラムには高域のパワーに違いが表れている（図 5.9 (b')）．機器の個体差による違いの可能性もあるが，検査員の指摘した “高域に存在する異音” の特徴とも一致する．提案法で音響特徴量を抽出したところ明確な変化が現れており（図 5.9 (c)），異常度も大きく上昇した（図 5.9 (d)）．

以上の結果から，提案法は実環境において，3D プリンタや送風ポンプの突発的な異常音や，ベアリングの傷などに起因する持続的な異常音を検知できることがわかった．これらの結果から，ネイマン・ピアソン指標によりニューラルネットワークを学習することで，実環境においても機器動作の異常音を検知できることを実験的に示した．

5.3 本章のまとめ

本章では，モーターの異常回転音やベアリングのぶつかり音などの普段発生しない音（異常音）を検知し，機器動作の状態が正常か異常かを判定することで機器の故障を検知する「異常音検知」の実現を目指した．機器の動作音からその機器が正常動作しているか，異常動作しているかを判定する異常音検知技術は，機器監視/保守業務の自動化を実現する技術として，産業界から大きな期待が寄せられている．この問題の難しさとして，機器の故障頻度がきわめて低いため，機器の異常動作音（ラベルデータ）が収集できず，一般的な識別のためのニューラルネットワークの目的関数である交差エントロピーが利用できない点にある．ゆえに従来では，オートエンコーダなどの DNN の中間層を音響特徴量とし，外れ値検出に基づくアルゴリズムで異常音を検知していた．しかし，DNN の学習は MMSE などの異常音検知とは無関係の目的関数で行われており，ニューラルネットワークが適切に学習されている保証はなかった．そこで本研究では，正常音が従う確率分布と統計的に差異がある音を異常音と定義することで異常音検知を仮説検定とみなし，仮説検定の最適化基準であるネイマン・ピアソンの補題 [147] から新たな目的関数である “ネイマン・ピアソン指標” を導出した．ニューラルネットワークをネイマン・ピアソン指標で学習するために，変分オートエンコーダに基づく実装法を提案した．定量評価試験では，従来法と比べ調和平均が 0.02 から 0.06 ポイント向上することを示し，実環境実験では 3D プリンタや送風ポンプの突発的な異常音や，ベアリングの傷などに起因する持続的な異常音を検知できることを示した．

5.3.1 本章の貢献と関連研究

本章の内容は，研究業績リスト [C-1] の内容をまとめたものである．この研究の貢献は，負例データの収集が困難な識別問題における音源情報の推定精度について定義し，外れ値検知に基づく異常音検知を最適化するための目的関数を設計したことにある．本研究では，これまで 正常音と異常音の識別問題として扱われてきた異常音検知を，正常音と正常音でない音の識別問題と捉えなおすことで，異常音検知を仮説検定問題として再定義した．この貢献は，銃声検知や未知話者検出などのセキュリティのための音源情報推定技術など，負例データの収集が困難な様々な音源情報推定へと応用ができる．

なお，識別問題において負例をニューラルネットワークで生成する枠組みという観点で，提案法は敵対的生成ネットワーク（GAN）と関連が深い．一般的な GAN では，識別の考え方に基づき負例を一様分布などの任意の分布から生成し，JS ダイバージェンスの最小化を目的関数として識別器を学習している一方で，提案法では仮説検定の考え方に基づき負例に関する確率分布や集合を定義して生成し，ネイマン・ピアソン基準で学習を行っている．

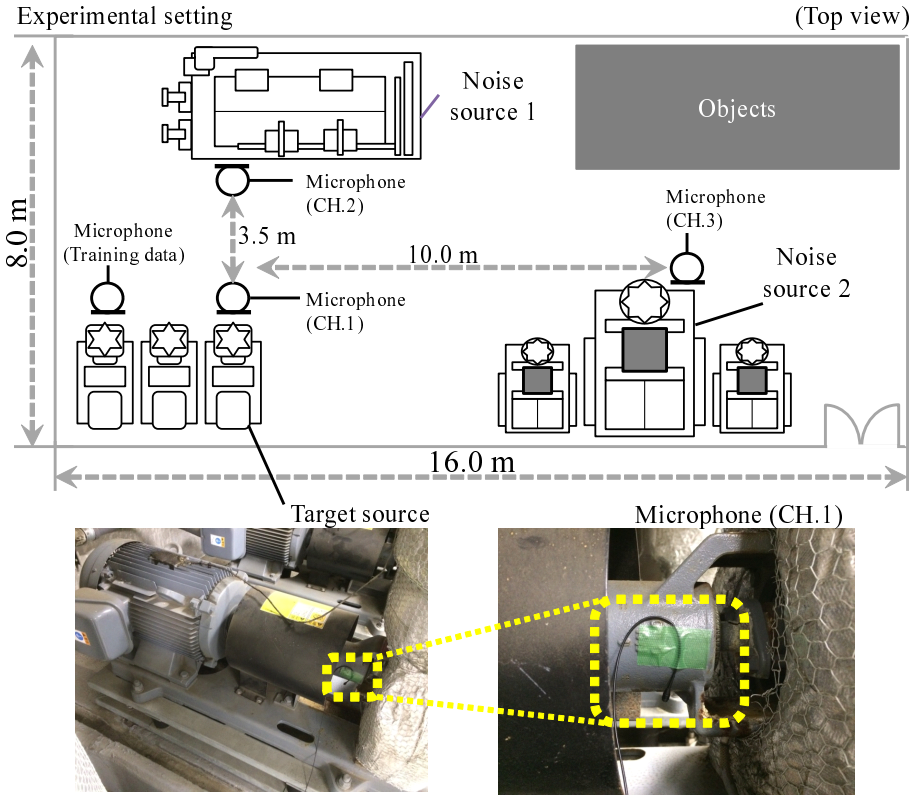


図 5.8: 実験条件およびマイクロホンの配置図.

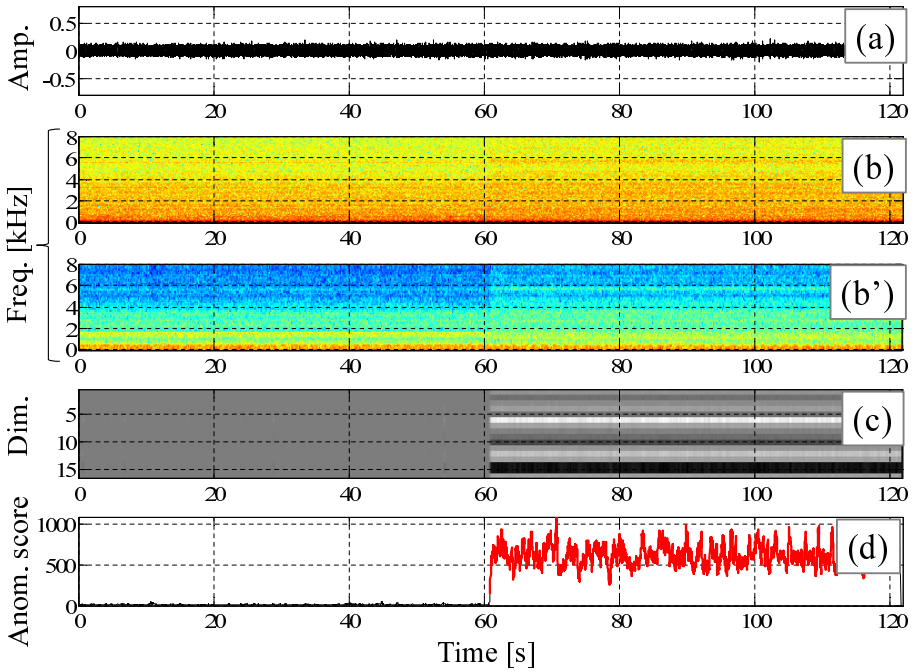


図 5.9: 給水ポンプの異常音検知結果．前半 60 秒が正常動作中の給水ポンプの動作音，後半 60 秒が異常動作中の給水ポンプの動作音である．各図はそれぞれ，(a) 観測波形，(b) 観測スペクトログラム，(b') 雑音抑圧後のスペクトログラム，(c) 提案法により抽出された音響特徴量，(d) 異常度を表す．

第 6 章

結論

本論文は，音源の位置，方向などの物理的かつ顕在的な音源情報に加え，音源の種類や状態などの情報的かつ潜在的な音源情報の推定を目指し，ニューラルネットワークの学習のための目的関数をどのように設計すべきかを研究した成果をまとめたものである．これまで，深層学習を音源情報推定に適用し，ニューラルネットワークのネットワーク構造を工夫することで，音声認識をはじめとする様々な音源情報の推定精度が向上してきた．しかし従来の枠組みでは，決定論的な目的関数では音源情報の性質や推定精度を定義できない，もしくは定義することが妥当ではない音源情報は推定が困難であった．そこで本研究では，音源情報推定において目的関数は，所望の音源情報の性質や推定精度を定義するものであることから，決定論的な目的関数でその性質や推定精度を決定論的に記述できなくとも，推定したい音源情報の特性や解きたい問題に応じて入出力値がとるべき値の確率分布や集合は記述できるはずであると考えた．そしてこの着眼点のもと，ニューラルネットワークの入出力が満たすべき統計的な性質を目的関数とする，「確率論的目的関数」を設計するという着想からこの問題に取り組んだ．

1 章では，研究背景，従来の音源情報推定の研究，および本研究の方針について述べた．

2 章では，本研究で題材とした音源強調と異常音検知の従来研究，深層学習の利用法，およびその問題点と，本研究で取り組む課題について述べた．

3 章では，スポーツの競技音などラベルデータが十分に存在しない源信号を強調するための手法について研究した．DNN を用いてラベルデータが十分に存在しない源信号を強調するためには，MFCC や LPC などの音響特徴量の候補から人手で事前に設計/選択した音響特徴量を観測信号から抽出することでネットワークのサイズを小さくして DNN の自由度を下げる必要がある．しかし，音響特徴量の候補の次元は大きく最適な組み合わせを探索的に決定することは困難であり，また音源の種類によって適切な音響特徴量が異なるため，音響特徴量を人手で選択することは困難だった．そこで 3 章では DNN の

推定誤差について確率分布を明示的に仮定し，その分布の性質から確率論的に目的関数を設計することで，適切な音響特徴量を自動選択する方法を目指した．源信号の推定誤差がガウス分布に従うと仮定し，源信号の推定誤差を最小化するための音響特徴量選択の目的関数として相互情報量を利用した．相互情報量を正確に計算するための手段として，福水らの提案した相互情報量を再生核ヒルベルト空間で計算する “カーネル次元圧縮 [128, 129]” を採用した．特徴量候補の次元数が大きい音響特徴量選択にカーネル次元圧縮法を適用するために，スパース正則化法に基づく微分可能な目的関数を導出し，大量な音響特徴量候補から適切な音響特徴量を勾配法により選択できる音響特徴量選択法を提案した．定量評価試験では，提案法を用いて音響特徴量を選択することで従来の音響特徴量選択法と比べ SDR が向上することを示した．また主観評価試験では，提案法を用いて音響特徴量を選択することで従来の音響特徴量選択法と比べ源信号の明瞭度が向上することを示した．

3 章の研究の貢献は，福水らの提案したカーネル次元圧縮法を特徴量候補の次元数が大きい音響特徴量選択に応用するために，スパース正則化法を利用した微分可能な目的関数を導出した点にある．この目的関数を用いることで，音響特徴量選択を勾配法で最適化できるようになるため，ニューラルネットワークへ入力する音響特徴量を現実的な計算時間で選択できるようになった．この貢献により，これまで推定が困難とされてた，学習データが十分に得られないような源信号や，適切な音響特徴量が未知な源信号も推定できるようになった．

4 章では，ラベルデータを一意に定めることができず，二乗誤差などの目的関数で推定精度を定義することが妥当でない源信号を強調するための手法について研究した．高品質な音声通信や聴覚補助の実現に向け，DNN を利用した音源強調の出力音の主観品質を向上させることを目指した．従来の DNN 音源強調では，源信号の振幅スペクトルなどをラベルデータとし，DNN の出力とラベルデータの二乗誤差を最小化するように DNN を学習させるため，出力音に歪が生じて主観品質が低下するという問題があった．一方，主観品質と相関の高い評価値（聴感評点）を計算機を用いて評価する手法は存在したものの，その計算方法は複雑かつ微分可能な関数の合成関数で設計されていないため，DNN を学習するための目的関数としてそのまま利用することはできなかった．そこで 4 章では，強化学習のフレームワークを応用し，音源強調のための DNN を，主観評価値と相関の高い音質評価値 [50, 51, 52] を利用して学習するための目的関数を設計した．その実装法として 4.2 節では時間周波数マスクの選択に基づく音源強調，4.3 節では時間周波数マスクの生成に基づく音源強調を提案した．定量評価試験では，提案する目的関数を利用することで，これまで目的関数として利用できなかった聴感評点を最大化するようにニューラルネットワークを学習できることを確認した．また主観評価試験では，

提案法は従来の MMSE に基づく目的関数を利用した音源強調よりも高い主観品質で音源強調できることを示した．

4 章の研究の貢献は，従来の DNN 音源強調の代表的な学習法である誤差逆伝搬法では利用が困難であった，微分不可能な評価関数を用いてニューラルネットワークを学習できる枠組みを提供した点にある．この貢献により，これまで音源強調の学習に利用できなかった聴感評点や人間の評価などの，より “高次” な評価尺度を目的関数として利用できるようになり，ニューラルネットワークを用いた音源強調の応用範囲を広げることができるようになった．

5 章では，機器の異常動作音などラベルデータが収集できない音源情報を推定するための手法について研究した．モーターの異常回転音やベアリングのぶつかり音などの普段発生しない音（異常音）を検知し，機器の動作音からその機器が正常動作しているか，異常動作しているかを判定する異常音検知技術は，機器監視/保守業務の自動化を実現する技術として，産業界から大きな期待が寄せられている．この問題の難しさとして，機器の故障頻度がきわめて低いため，機器の異常動作音（ラベルデータ）が収集できず，一般的な識別のためのニューラルネットワークの目的関数である交差エントロピーが利用できない点にある．ゆえに従来では，オートエンコーダなどの DNN の中間層を音響特徴量とし，外れ値検出に基づくアルゴリズムで異常音を検知していた．しかし，DNN の学習は MMSE などの異常音検知とは無関係の目的関数で行われており，ニューラルネットワークが適切に学習されている保証はなかった．そこで本研究では，正常音が従う確率分布と統計的に差異がある音を異常音と定義することで異常音検知を仮説検定とみなし，仮説検定の最適化基準であるネイマン・ピアソンの補題 [147] から新たな目的関数である “ネイマン・ピアソン指標” を導出した．ニューラルネットワークをネイマン・ピアソン指標で学習するために，変分オートエンコーダに基づく実装法を提案した．定量評価試験では，従来法と比べ調和平均が 0.02 から 0.06 ポイント向上することを示し，実環境実験では 3D プリンタや送風ポンプの突発的な異常音や，ベアリングの傷などに起因する持続的な異常音を検知できることを示した．

5 章の研究の貢献は，負例データの収集が困難な識別問題における音源情報の推定精度について定義し，外れ値検知に基づく異常音検知を最適化するための目的関数を設計したことにある．本研究では，これまで 正常音と異常音の識別問題として扱われてきた異常音検知を，正常音と正常音でない音の識別問題と捉えなおすことで，異常音検知を仮説検定問題として再定義した．この貢献は，銃声検知や未知話者検出などのセキュリティのための音源情報推定技術など，負例データの収集が困難な様々な音源情報推定へと応用ができる．

本論文では，確率論的に導かれる目的関数を利用することで音源情報推定精度が向上することを示すために，単純な構造のニューラルネットワークを利用してきた．目的関数の研究は，昨今著しい発展を遂げているネットワーク構造の研究と独立性が高い．双方の技術の進展を組み合わせることで，より高度な音源情報推定が可能になると考えている．また本研究では，これまで困難とされてきた音源情報推定の題材として音源強調と異常音検知を採り上げ，その推定するための新たな切り口として，確率論的な目的関数の高度化という一つの方針を示した．今後は本研究を発展させ，環境音や音声の個人性などより多様かつ複雑な音源情報を推定するための目的関数を研究していく．

音響信号処理をはじめとするメディア処理において目的関数の設計は，所望の情報の性質や推定精度を定義することと等価であり，その研究は非常に重要なものである．通信技術や計算機能力の発展に伴い，知的な音情報処理は，産業界や家庭内において今後ますます必要とされていくと考えられる．この発展のためには，潜在的な音源情報の推定は必須であり，また人間の感性や個人的な嗜好のような，より高次かつあいまいな音源情報も推定してく必要がある．こういった音源情報の多くはラベルデータの収集が困難であるし，決定論的な目的関数でその性質や推定精度を定義することは妥当ではないため，その推定のためには目的関数の発展が不可欠である．本研究は音源情報推定のための目的関数の高度化に先駆的に取り組んだものであり，今後の音情報処理の基礎/応用研究，アプリケーションの開発，および実用化の一助になると考えている．

付録 A

遠方配置したマイクロホンを連携させる音源強調法

本章では，5 章の実環境実験で用いた音源強調法を説明する．製造工場など大規模な空間である監視対象機器の稼働音（源信号）のみを収音したい状況を考える．しかし製造工場では，他の製造機の動作音が複数の方向から到来することが多く，観測信号には源信号と雑音の両方が収含まれてしまうため，時間周波数マスキングを用いて源信号を強調したい．製造工場では，製造機などの雑音源の位置が固定されているため，雑音源の近くにマイクロホンを配置し，雑音をマイクロホンで直接観測できるアドバンテージがある．すると雑音の振幅スペクトルが既知となるため，容易に時間周波数マスクを設計できそうである．しかし実際には，製造工場では監視対象機器と他の製造機が離れて配置されていることが多いため，以下の理由で瞬時混合が仮定できず，雑音を観測した信号をそのまま雑音の振幅スペクトルとしては時間周波数マスクは設計できない．

残響時間の問題 製造工場などは残響時間が長く，残響（インパルス応答）の時間長が短時間フーリエ変換（STFT: short time Fourier transform）の分析幅より長くなってしまう．

到達時間差の問題 マイクロホン間隔に応じた到達時間差が STFT の分析シフト幅より大きくなると，源信号を観測するマイクロホンと雑音を観測するマイクロホンの STFT 分析結果の間で時間フレーム差が生じる．

本付録では，上記の問題を解決するために，製造工場など大規模な空間での観測信号を，振幅スペクトル領域での瞬時混合でモデル化することで，所望の源信号を推定する手法を説明する．

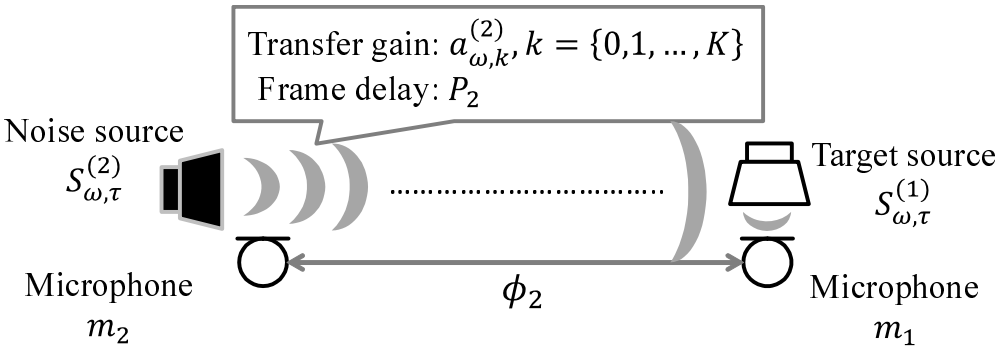


図 A.1: 観測信号のモデル化（$M = 2$）．マイクロホン m_2 で観測した雑音 $S^{(2)}_{\omega,\tau}$ が時間フレーム差 P_2 と伝達特性ゲイン $a^{(2)}_{\omega,k}$ をもって到来する．

A.1 観測信号のモデル化

今，源信号 $S^{(1)}_{\omega,\tau} \in \mathbb{C}^{\Omega\times T}$ を，M 本のマイクロホンの観測信号から推定する問題を考える．$S^{(1)}_{\omega,\tau}$ に最も近い位置に配置したマイクロホンの番号を 1 とし，その観測音 $X^{(1)}_{\omega,\tau} \in \mathbb{C}^{\Omega\times T}$ を以下で記述する．

$$X^{(1)}_{\omega,\tau} = S^{(1)}_{\omega,\tau} + N_{\omega,\tau} \tag{A.1}$$

ここで $N_{\omega,\tau} \in \mathbb{C}^{\Omega\times T}$ は，複数の他の製造機から到来する雑音であり，空間内の $M-1$ 個の雑音源 $S^{(2,\ldots,M)}_{\omega,\tau}$ が混ぜ合わさったものと考える．1 番目のマイクロホンと m 番目雑音源が空間的に離れて配置されている場合，長い残響や到達時間差の影響で，$N_{\omega,\tau}$ を瞬時混合の形式で記述できない．この問題を解決するために，残響と到達時間差を時間周波数領域でモデル化したい．瞬時混合形式で記述できないほど長い残響を時間周波数表現するために，エコーキャンセラ [148] や音響イベント検出 [149] の分野では，残響を近似的に，時間周波数領域の畳み込みで記述する方法がある．また到達時間差は，時間フレームの遅延（シフト）として記述する．すると雑音 $N_{\omega,\tau}$ を以下のように記述できる．

$$N_{\omega,\tau} \approx \sum_{m=2}^{M}\sum_{k=0}^{K} A^{(m)}_{\omega,k} S^{(m)}_{\omega,\tau-P_m-k} \tag{A.2}$$

ここで，$P_m \in \mathbb{N}_+$ は，1 番目のマイクロホンと m 番目雑音源 $S^{(m)}_{\omega,\tau}$ の位置差に応じて生じる時間フレーム差，$A^{(m)}_{\omega,k}$ は，1 番目と m 番目のマイクロホンと，m 番目雑音源 $S^{(m)}_{\omega,\tau}$ から 1 番目のマイクロホンまでの伝達特性である．式 (A.1)(A.2) より，1 番目のマイクロホンの観測を瞬時混合の形式で記述できるようになるため，時間周波数マスクによる源信号推定を実行できる．

時間周波数マスクを設計するために，$|N_{\omega,\tau}|$ を $M-1$ 個のマイクロホンの観測を用いて記述する．m 番目の雑音源の近傍には m 番目のマイクロホンが配置されているとす

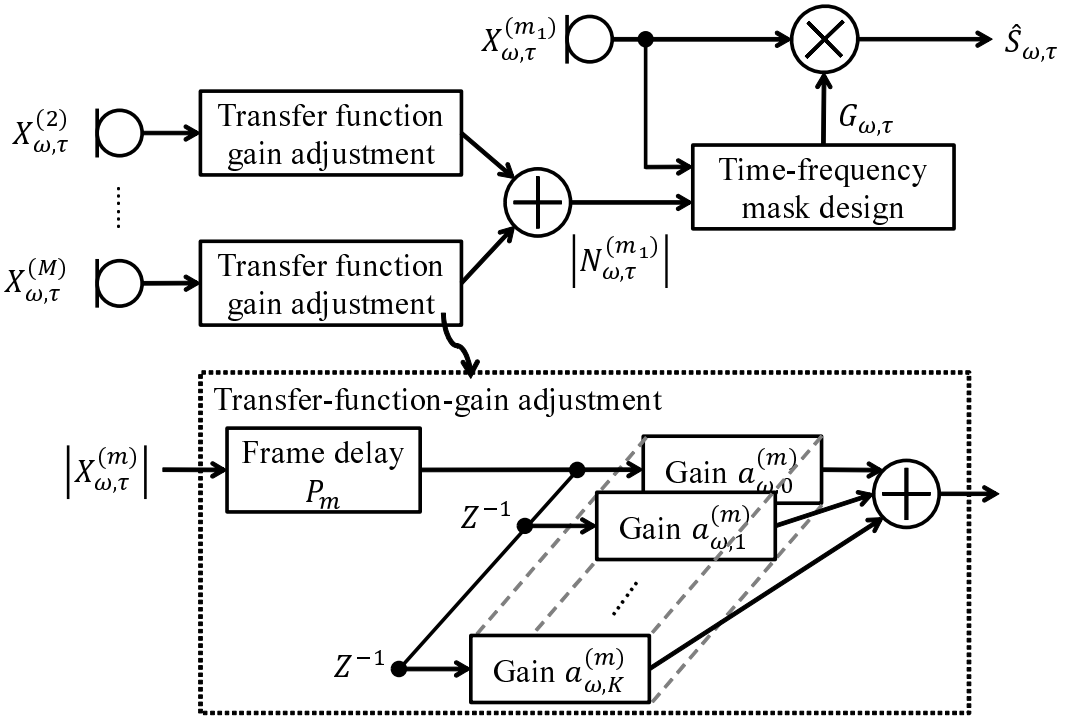


図 A.2: 提案法の音源強調処理手順.

る．製造工場では，大音量を発生させる大型機器同士は離して配置することが多いため，m 番目のマイクロホンとその他のマイクロホンは十分に離して配置されていると仮定する．すると m 番目のマイクロホンの近傍では，$|S^{(m)}_{\omega,\tau}| \gg |S^{(1,\ldots,M,\neq m)}_{\omega,\tau}|$ が成り立ち，その観測信号 $X^{(m)}_{\omega,\tau}$ は近似的に，

$$X^{(m)}_{\omega,\tau} \approx S^{(m)}_{\omega,\tau} \tag{A.3}$$

と記述できる．すると，各雑音源が互いに無相関と仮定すると，$N_{\omega,\tau}$ の振幅スペクトルも近似的に以下のように記述できる（図 A.1）．

$$|N_{\omega,\tau}| \approx \sum_{m=2}^{M}\sum_{k=0}^{K} a^{(m)}_{\omega,k}|X^{(m)}_{\omega,\tau-P_m-k}| \tag{A.4}$$

ここで $a^{(m)}_{\omega,k} = |A^{(m)}_{\omega,k}| \in \mathbb{R}_+$ である（以降，伝達ゲインと呼ぶ）．式 (A.4) より，各雑音源の時間フレーム差 $P_{2,\ldots,M}$ と伝達特性ゲイン $\boldsymbol{a}^{(2,\ldots,M)}_{1,\ldots,K}$ が推定できれば，雑音の振幅スペクトルが推定できるため時間周波数マスクでき，音源強調が可能になる（図 A.2）．次節では，$\Theta = \{\boldsymbol{a}^{(2,\ldots,M)}_{1,\ldots,K}, P_{2,\ldots,M}\}$ を推定するため枠組みを説明する．

A.2 雑音推定のための目的関数の設計

本節では，$\Theta = \{\boldsymbol{a}^{(2,\ldots,M)}_{1,\ldots,K}, P_{2,\ldots,M}\}$ を推定する．Θ の事前分布と Θ を得た下での観測信号の確率分布を設定し，事後確率最大化推定の枠組みで $P_{2,\ldots,M}$ と $\boldsymbol{a}^{(2,\ldots,M)}_{1,\ldots,K}$ を推定する．

まず，各確率分布の記述を簡潔にするために，観測信号のモデルである式 (A.1) を行列形式で記述する．式 (A.1) が振幅スペクトル領域でも成り立つと仮定すると，$|X^{(1)}_{\omega,\tau}|$ を近似的に以下のように記述できる．

$$|X^{(1)}_{\omega,\tau}| = |S^{(1)}_{\omega,\tau}| + |N_{\omega,\tau}| \tag{A.5}$$

すると式 (A.5) は以下の行列演算で表現できる．

$$\boldsymbol{X}^{(1)}_{\tau} \approx \boldsymbol{S}^{(1)}_{\tau} + \boldsymbol{N}_{\tau} \tag{A.6}$$

$$\boldsymbol{X}^{(m)}_{\tau} \approx \boldsymbol{S}^{(m)}_{\tau} \tag{A.7}$$

$$\begin{aligned}\boldsymbol{N}_{\tau} &\approx \sum_{m=2}^{M}\sum_{k=0}^{K} \boldsymbol{a}^{(m)}_{k} \odot \boldsymbol{X}^{(m)}_{\tau-P_m-k} \\ &\approx \mathbf{X}_{\tau}\mathbf{a}\end{aligned} \tag{A.8}$$

ここで

$$\boldsymbol{X}^{(i)}_{\tau} = \left(|X^{(i)}_{1,\tau}|, |X^{(i)}_{2,\tau}|, ..., |X^{(i)}_{\Omega,\tau}|\right)^{\top} \tag{A.9}$$

$$\boldsymbol{S}^{(i)}_{\tau} = \left(|S^{(i)}_{1,\tau}|, |S^{(i)}_{2,\tau}|, ..., |S^{(i)}_{\Omega,\tau}|\right)^{\top} \tag{A.10}$$

$$\boldsymbol{N}_{\tau} = \left(|N_{1,\tau}|, |N_{2,\tau}|, ..., |N_{\Omega,\tau}|\right)^{\top} \tag{A.11}$$

$$\boldsymbol{a}^{(i)}_{k} = \left(a^{(i)}_{1,k}, a^{(i)}_{2,k}, ..., a^{(i)}_{\Omega,k}\right)^{\top} \tag{A.12}$$

$$\mathbf{X}_{\tau} = \left(\mathbf{X}^{(2)}_{\tau}, ..., \mathbf{X}^{(M)}_{\tau}\right) \tag{A.13}$$

$$\mathbf{X}^{(m)}_{\tau} = \left(\mathrm{diag}(\boldsymbol{X}^{(m)}_{\tau-P_m}), ..., \mathrm{diag}(\boldsymbol{X}^{(m)}_{\tau-P_m-K})\right) \tag{A.14}$$

$$\mathbf{a} = \left(\mathbf{a}^{(2)}, ..., \mathbf{a}^{(M)}\right) \tag{A.15}$$

$$\mathbf{a}^{(m)} = \left(\boldsymbol{a}^{(m)}_{0}, ..., \boldsymbol{a}^{(m)}_{K}\right) \tag{A.16}$$

である．ただし $\odot$ はアダマール積，$\mathrm{diag}(\boldsymbol{x})$ はベクトル $\boldsymbol{x}$ を対角要素に持つ対角行列を表す．

次に観測信号 $\boldsymbol{X}^{(1)}_{\tau}$ に対する確率分布を設計する．ここで $X^{(1)}_{\omega,\tau}$ に源信号 $S^{(1)}_{\omega,\tau}$ が含まれる時間が短い（すなわち，監視対象の機器がほとんど動作していない）データセットが手に入るとする．式 (A.6) より，そのデータセットでは，ほとんどの時間インデックス τ で

$$\boldsymbol{X}^{(1)}_{\tau} = \boldsymbol{N}_{\tau} \tag{A.17}$$

が成り立つ．この仮定から $\boldsymbol{X}^{(1)}_{\tau}$ の確率分布を，$\boldsymbol{N}_{\tau}$ を平均，共分散行列 $\mathrm{diag}(\boldsymbol{\sigma})$ を持つ

ガウス分布 $\mathcal{N}\left(\boldsymbol{N}_\tau, \mathrm{diag}(\boldsymbol{\sigma}^2)\right)$ でモデル化する．

$$\boldsymbol{X}_\tau^{(1)} \sim \mathcal{N}\left(\boldsymbol{X}_\tau^{(1)} | \boldsymbol{N}_\tau, \mathrm{diag}(\boldsymbol{\sigma})\right) \tag{A.18}$$

$$\sim \frac{|\boldsymbol{\Lambda}|^{1/2}}{(2\pi)^{\Omega/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{X}_\tau^{(1)} - \boldsymbol{N}_\tau)^\top \boldsymbol{\Lambda} (\boldsymbol{X}_\tau^{(1)} - \boldsymbol{N}_\tau)\right\} \tag{A.19}$$

ここで $\boldsymbol{\Lambda} = (\mathrm{diag}(\boldsymbol{\sigma}))^{-1}$ であり，$\boldsymbol{\sigma} = (\sigma_1, ..., \sigma_\Omega)^\top$ は $\boldsymbol{X}_\tau^{(1)}$ の各周波数ごとのパワーとして

$$\sigma_\omega = \frac{1}{T} \sum_{\tau=1}^{T} |X_{\omega,\tau}^{(1)}| \tag{A.20}$$

で求める．これは，各周波数インデックスごとの振幅の差を補正することを目的としている．

最後に Θ の事前分布を，各パラメータの物理的な特性を元に設計する．まず時間フレーム差 $P_{2,...,M}$ の事前分布を設計する．m 番目のマイクロホンは m 番目の雑音源の近傍に配置されているならば，P_m はおおよそ，1 番目のマイクロホンと m 番目のマイクロホンの距離で推測できると考えられる．つまり，1 番目のマイクロホンと m 番目のマイクロホンの距離を ϕ_m，音速を C m/s，標本化周波数を f_s，STFT のシフト幅を f_{shift} としたとき，おおよその時間フレーム差 D_m は

$$D_m = \mathrm{round}\left\{\frac{\phi_m}{C} \cdot \frac{f_s}{f_{\mathrm{shift}}}\right\} \tag{A.21}$$

で求まる．ここで round$\{\cdot\}$ は整数への四捨五入を表す．ただし実際には m 番目のマイクロホンは m 番目の雑音源の距離がゼロではないため，P_m は D_m の近傍で確率的に揺らぐと考えられる．このことをモデル化するために，時間フレーム差の事前分布を，平均値 D_m を持つポアソン分布で設計する．

$$P_m \sim \mathrm{Poisson}(P_m | D_m)$$

$$\sim \frac{D_m^{P_m}}{P_m!} \exp\{-D_m\} \tag{A.22}$$

次いで伝達ゲイン $\boldsymbol{a}_{1,...,K}^{(2,...,M)}$ の事前分布を設計する．$a_{\omega,k}^{(m)}$ は伝達ゲインを表すため正の実数であるまた一般に伝達特性は時間減衰するため，そのゲインである伝達ゲイン $a_{\omega,k}^{(m)}$ も時間 k が進むほど小さくなる．このことをモデル化するために，伝達ゲインの確率分布は平均値 α_k を持つ指数分布でモデル化する．

$$a_{\omega,k}^{(m)} \sim \mathrm{Exponential}\left(a_{\omega,k}^{(m)} | \alpha_k\right)$$

$$\sim \frac{1}{\alpha_k} \exp\left\{-\frac{a_{\omega,k}^{(m)}}{\alpha_k}\right\} \tag{A.23}$$

ここで α_k はフレームの経過に従って減少させるために，以下のように計算する．

$$\alpha_k = \max(\alpha - \beta k, \epsilon) \tag{A.24}$$

ここで α は 0 フレーム目の α_k の値，β は減衰重み，ϵ はゼロ除算を避けるための小さな係数である．

以上の議論より，観測信号と各パラメータについて確率分布が定義できたため，最大化すべき目的関数は以下のように記述できる．

$$\begin{aligned}\mathcal{J} &= p(\Theta|\boldsymbol{X}_{1,\ldots,T}) \propto p(\boldsymbol{X}_{1,\ldots,T}, \Theta) \\ &= p(\boldsymbol{X}_{1,\ldots,T}|\Theta) p(\boldsymbol{a}_{1,\ldots,K}^{(2,\ldots,M)}) p(P_{1,\ldots,M})\end{aligned} \tag{A.25}$$

$$p(\boldsymbol{X}_{1,\ldots,T}|\Theta) = \prod_{\tau=1}^{T} \mathcal{N}\left(\boldsymbol{X}_{\tau}^{(1)}|\boldsymbol{N}_{\tau}, \mathrm{diag}(\boldsymbol{\sigma})\right) \tag{A.26}$$

$$p(\boldsymbol{a}_{1,\ldots,K}^{(2,\ldots,M)}) = \prod_{\omega=1}^{\Omega} \prod_{m=2}^{M} \prod_{k=1}^{K} \mathrm{Exponential}(a_{\omega,k}^{(m)}|\alpha_k) \tag{A.27}$$

$$p(P_{2,\ldots,M}) = \prod_{m=2}^{M} \mathrm{Poisson}(P_m|D_m) \tag{A.28}$$

ここで，$\boldsymbol{a}_{1,\ldots,K}^{(2,\ldots,M)}$ は非負の値である必要があるため，この最適化は，以下のような L の制約付き最大化問題となる．

$$\Theta \leftarrow \arg\max_{\Theta} \mathcal{J} \quad \text{subject to} \quad 0 \leq a_{1,\ldots,\Omega,1,\ldots,K}^{(2,\ldots,M)} \tag{A.29}$$

ここで $\mathcal{J}$ は確率値の積の形になっているため，計算機の途中でアンダーフローを起こす可能性がある．ここで，対数関数が単調増加関数であることを利用し，尤度関数の代わりに対数尤度関数を最大化する．

$$\Theta \leftarrow \arg\max_{\Theta} \mathcal{J} \quad \text{subject to} \quad 0 \leq a_{1,\ldots,\Omega,1,\ldots,K}^{(2,\ldots,M)} \tag{A.30}$$

$$\mathcal{J} = \ln p(\boldsymbol{X}_{1,\ldots,T}|\Theta) + \ln p(\boldsymbol{a}_{1,\ldots,K}^{(2,\ldots,M)}) + \ln p(P_{2,\ldots,M}) \tag{A.31}$$

ここで各要素は以下のように記述できる．

$$\ln p(\boldsymbol{X}_{1,\ldots,T}|\Theta) \propto -\frac{1}{2} \sum_{\tau=1}^{T} \left(\boldsymbol{X}_{\tau}^{(1)} - \mathbf{X}_{\tau}\mathbf{a}\right)^{\top} \boldsymbol{\Lambda} \left(\boldsymbol{X}_{\tau}^{(1)} - \mathbf{X}_{\tau}\mathbf{a}\right) \tag{A.32}$$

$$\ln p(\boldsymbol{a}_{1,\ldots,K}^{(2,\ldots,M)}) \propto \sum_{\omega=1}^{\Omega} \sum_{m=2}^{M} \sum_{k=1}^{K} -\ln \alpha_k - \frac{a_k^{(m)}}{\alpha_k} \tag{A.33}$$

$$\ln p(P_{2,\ldots,M}) \propto \sum_{m=2}^{M} -\ln(P_m!) + P_m \ln(D_m) - D_m \tag{A.34}$$

以上の変形により，$\mathcal{J}$ を構成する各尤度関数の最大化は容易になった．しかし式 (A.30) は $P_{2,...,M}$ と $\boldsymbol{a}_{1,...,K}^{(2,...,M)}$ の 2 変数の関数であり，2 変数に関して同時に最適化を行うことは困難である．そこで本手法では，式 (A.30) を cordinate descent (CD) 法を用いて最大化する．具体的には尤度関数を，$\boldsymbol{a}_{1,...,K}^{(2,...,M)}$ に関する項と $P_{2,...,M}$ に関する項に分解し，

$$\mathcal{J}_a = \ln p(\boldsymbol{X}_{1,...,T}|\Theta) + \ln p(\boldsymbol{a}_{1,...,K}^{(2,...,M)}) \tag{A.35}$$

$$\mathcal{J}_P = \ln p(\boldsymbol{X}_{1,...,T}|\Theta) + \ln p(P_{2,...,M}) \tag{A.36}$$

各変数を交互に最適化することで，$\mathcal{J}$ を最大化する．

$$\boldsymbol{a}_{1,...,K}^{(2,...,M)} \leftarrow \underset{\boldsymbol{a}_{1,...,K}^{(2,...,M)}}{\arg\max}\, \mathcal{J}_a \quad \text{subject to} \quad 0 \leq a_{1,...,\Omega,1,...,K}^{(2,...,M)} \tag{A.37}$$

$$P_{2,...,M} \leftarrow \underset{P_{2,...,M}}{\arg\max}\, \mathcal{J}_P \tag{A.38}$$

式 (A.37) は制約付き最大化のため，近接勾配法を用いて最適化する．具体的には $\mathcal{J}_a$ の $\mathbf{a}$ に関する勾配ベクトルを以下の式で求め，

$$\frac{\partial \mathcal{J}_a}{\partial \mathbf{a}} = \frac{1}{T}\sum_{\tau=1}^{T} \mathbf{X}_\tau^\top \boldsymbol{\Lambda}\left(-\boldsymbol{X}_\tau^{(1)} + \mathbf{X}_\tau \mathbf{a}\right) - \boldsymbol{\alpha} \tag{A.39}$$

$$\boldsymbol{\alpha} = \left(\underbrace{\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\alpha}}, ..., \tilde{\boldsymbol{\alpha}}}_{M-1}\right) \tag{A.40}$$

$$\tilde{\boldsymbol{\alpha}} = \left(\underbrace{\frac{1}{\alpha_0}, ..., \frac{1}{\alpha_0}}_{\Omega}, \underbrace{\frac{1}{\alpha_1}, ..., \frac{1}{\alpha_1}}_{\Omega}, ..., \underbrace{\frac{1}{\alpha_K}, ..., \frac{1}{\alpha_K}}_{\Omega}\right) \tag{A.41}$$

式 (A.42) の勾配法と，式 (A.43) のフロアリングを交互に行う繰り返し最適化で実行する．

$$\mathbf{a} \leftarrow \mathbf{a} + \lambda \frac{\partial \mathcal{L}_a}{\partial \mathbf{a}} \tag{A.42}$$

$$a_{1,...,\Omega,1,...,K}^{(2,...,M)} \leftarrow \max(0, a_{1,...,\Omega,1,...,K}^{(2,...,M)}) \tag{A.43}$$

ここで λ は更新のステップサイズである．式 (A.38) は離散変数の組み合わせ最適化であるため，グリッドーサーチを行う．具体的には，すべての m について P_m のとりえる最大値と最小値を定義し，すべての P_m の最小から最大の組み合わせについて $\mathcal{J}_P$ を評価し，これが最大となる組み合わせで P_m を更新する．実用的には，各マイク距離 $\phi_{2,...,M}$ から推測される音源距離の最小値 $\phi_{2,...,M}^{\min}$ と最大値 $\phi_{2,...,M}^{\max}$ を入力し，そこから P_m のとりえる最大値と最小値を計算する．音源距離の最大値と最小値は，$\phi_m^{\min} = \phi_m - 20$，$\phi_m^{\max} = \phi_m + 20$ 程度に設定する．

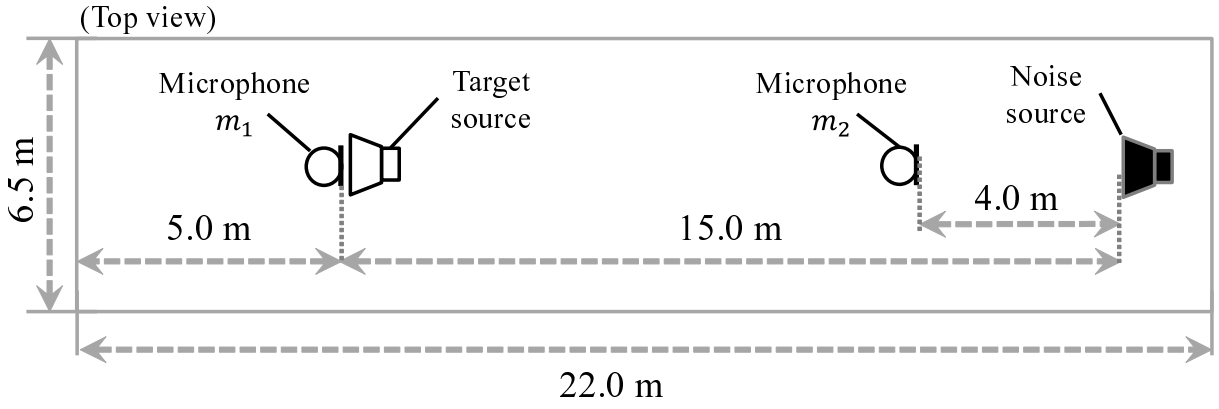


図 A.3: マイクロホンとスピーカの配置図.

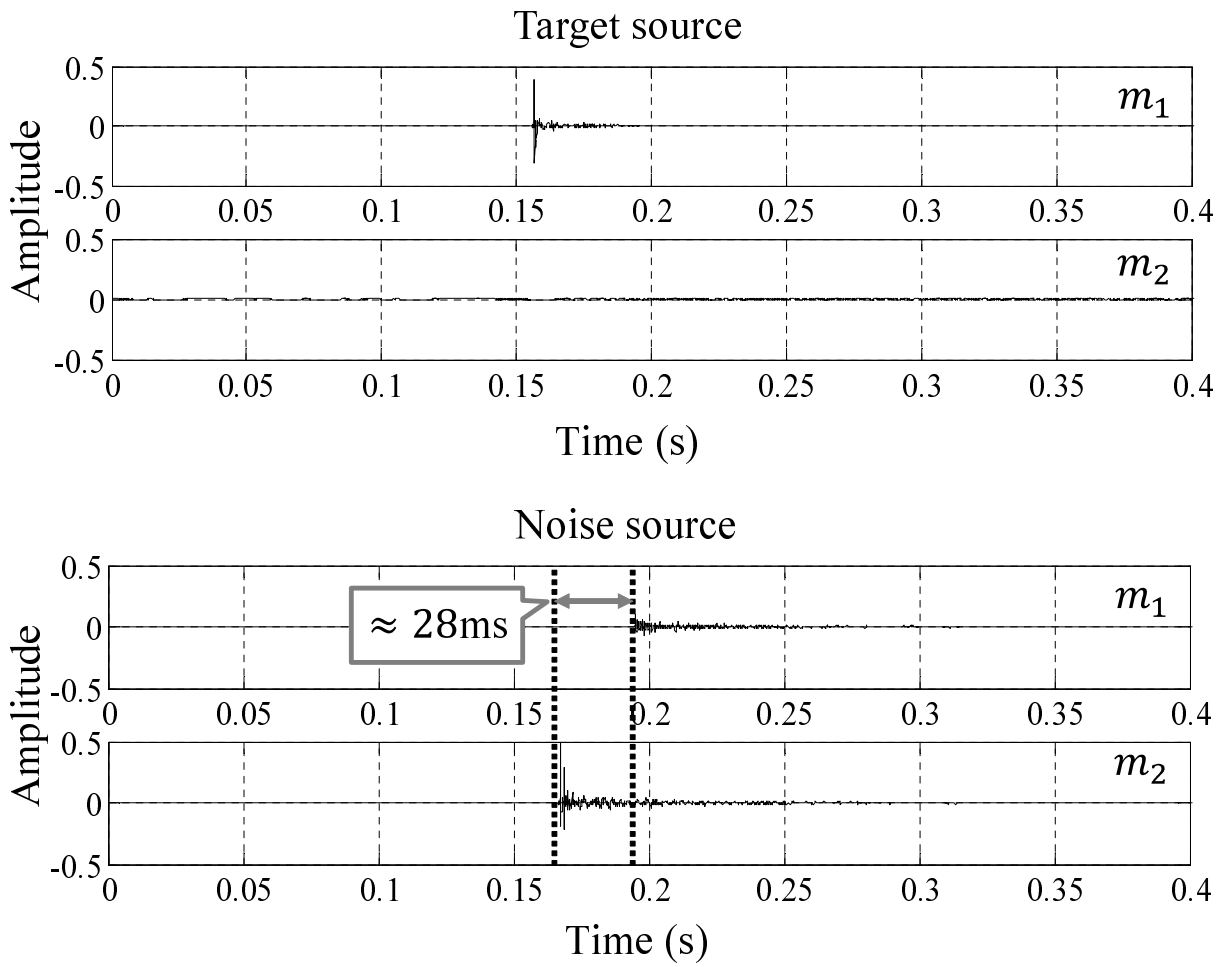


図 A.4: 各スピーカから各マイクロホンまでのインパルス応答.

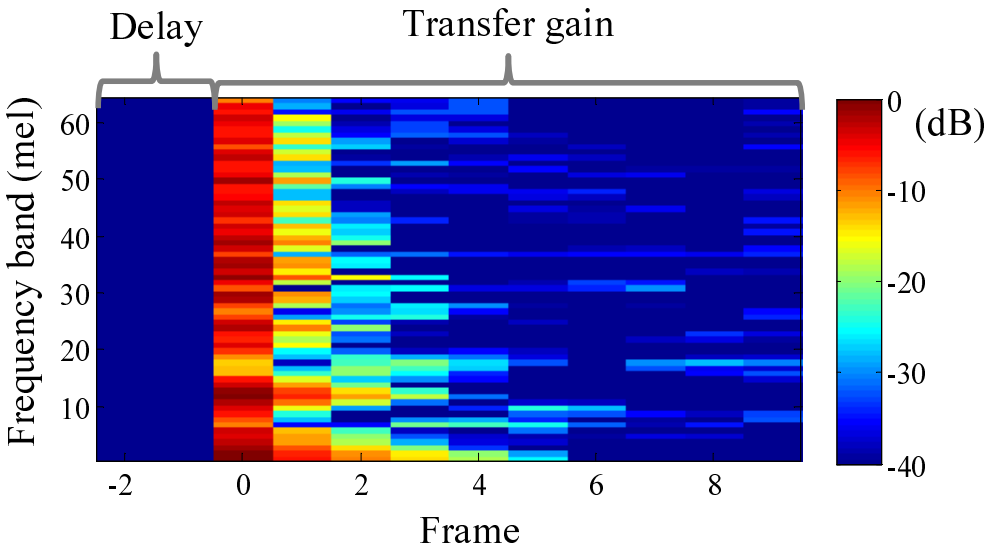


図 A.5: 時間フレーム差（$P_2 = 2$）と伝達ゲインの推定結果.

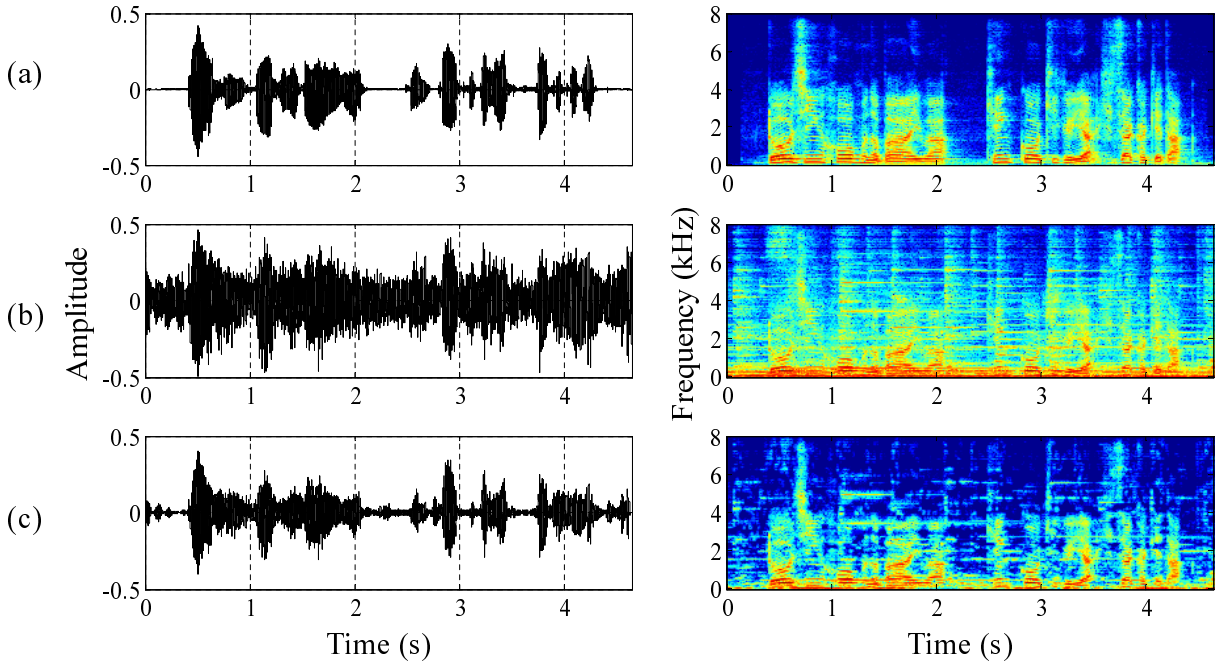


図 A.6: 処理結果例．左図が波形，右図がスペクトログラムを表す．各図はそれぞれ (a) 源信号，(b) 観測信号，(c) 出力信号である．

Algorithm 5 提案法の学習アルゴリズム

Input: $\boldsymbol{X}_{1,\ldots,T}^{(1,\ldots,M)}$

Output: $\boldsymbol{a}_{1,\ldots,K}^{(2,\ldots,M)}, P_{2,\ldots,M}$

Initialize $\boldsymbol{a}_{1,\ldots,K}^{(2,\ldots,M)}, P_{2,\ldots,M}$

while *until algorithm convergence* **do**

$\mathsf{a} \leftarrow \mathsf{a} + \lambda \frac{\partial \mathcal{L}_a}{\partial \mathsf{a}}$

$a_{1,\ldots,\Omega,1,\ldots,K}^{(2,\ldots,M)} \leftarrow \max(0, a_{1,\ldots,\Omega,1,\ldots,K}^{(2,\ldots,M)})$

$P_{2,\ldots,M} \leftarrow \arg\max_{P_{2,\ldots,M}} \mathcal{L}_P$ by grid-search algorithm.

end while

A.3 実験

本手法による音源強調の挙動を，実験室での動作実験，及び実環境での動作結果を用いて示す．

A.3.1 動作実験

まず本手法の音源強調性能を定量評価実験により評価した．実験は図 A.3 に示すようにマイクロホンとスピーカの配置した室内で行った．目的音（target source）から各マイク位置までと，雑音（noise source）から各マイク位置までのインパルス応答を図 A.4 に示す．雑音のインパルス応答からわかるように，雑音は両方のマイクロホン m_1, m_2 に含まれており，また m_1 と m_2 の間にはマイク間距離（約 12.0 m）に応じた 28 ms 程度の時間差が生じている．

目的音の学習データには ATR 音声データベース [141] から男性 4 名，女性 4 名による全 400 発話を利用した．また，テストデータには New Japan 音声データベースより男性 2 名，女性 2 名による全 200 発話を利用した．雑音は音楽とし，学習データは SiSEC データセット（MSD100）[150] より男性ボーカル 3 曲，女性ボーカル 3 曲とした．またテストデータは，学習データとは異なるアーティストの男性ボーカル 1 曲，女性ボーカル 1 曲とした．すべて目的音および雑音のデータは 16kHz にダウンサンプリングした．学習の SNR は -12 dB，-6 dB，0 dB，6 dB，12 dB とし，評価は-6 dB，0 dB，6 dB，12 dB で行った．パラメータの次元数を抑えるために観測信号は $\mathsf{B} = 64$ のメルフィルタバンクで圧縮し，時間周波数マスク設計の際にスプライン補間で線形周波数に補間した．短時間フーリエ変換のフレームサイズは 512 サンプルとし，シフト幅は 256 サンプルとした．また学習に用いた各種パラメータは実験的に $\alpha = 1.0$，$\beta = 5.0 \times 10^{-2}$，$K = 10$，$\lambda = 10^{-5}$ とした．

まず，時間フレーム差と伝達ゲインの推定結果を図 A.5 に示す．時間フレーム差は

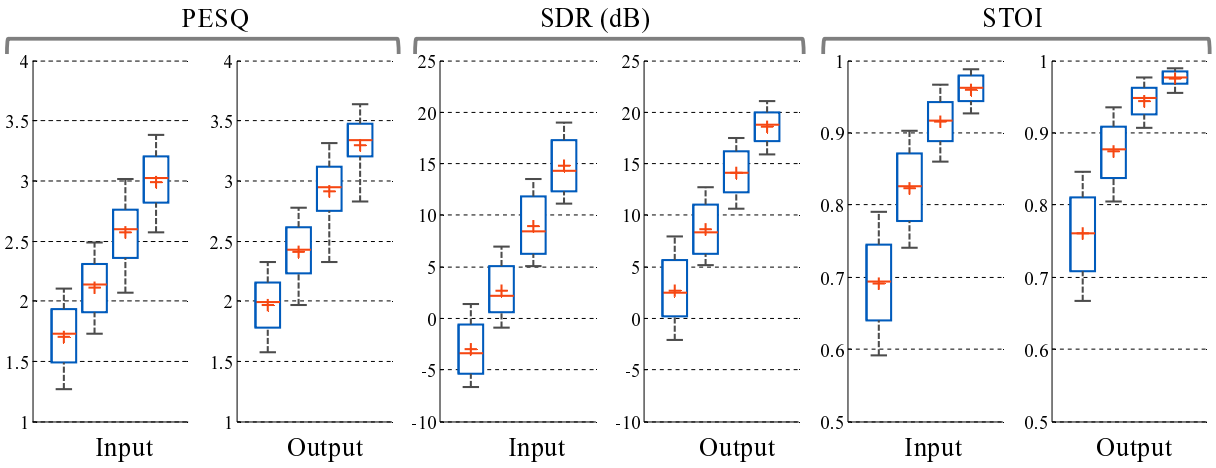


図 A.7: PESQ，SDR，STOI の定量評価結果．縦軸はそれぞれの評価値の大きさを表す．各 4 つの箱ひげ図は左から，観測信号の SNR が-6dB, 0dB, 6dB, 12dB の結果を表す．

$P_2 = 2$ と推定された．これは本実験条件においては 1 フレームが 16 ms (256 pts /16000 Hz) であり，インパルス応答におけるマイク間時間差が約 28 ms であったことから，妥当な値が推定されていることがわかる．また伝達ゲインも時間的に減衰していることがわかる．推定されたパラメータを用いて音源強調を行った例を図 A.6 に示す．時間波形およびスペクトログラムより源信号が強調されていることがわかる．

次に，音源強調性能を PESQ（音質），SDR（スペクトル歪），STOI [146]（音声明瞭度）を用いて定量評価した結果を図 A.7 に示す．全ての評価尺度において，提案法で処理を行うことで評価値が改善しており，PESQ で約 0.4 point，SDR で約 5〜10 dB，STOI で 0.08〜0.02 point の改善が見られた．これらのことから，本手法により遠方配置したマイクロホンを連携させて音源強調を行えることがわかる．

参考文献

[1] K. Kobayashi, Y. Haneda, K. Furuya, and A. Kataoka, "A hands-free unit with noise reduction by using adaptive beamformer," *IEEE Transactions on Consumer Electronics*, Vol.54-1, 2008.

[2] Y. Hioka, K. Furuya, K. Kobayashi, S. Sakauchi, and Y. Haneda, "Angular region-wise speech enhancement for hands-free speakerphone," *IEEE Transactions on Consumer Electronics*, Vol.58-4, 2012.

[3] M. Fukui, K. Kobayashi, Y. Haneda, and H. Ohmuro, "Low-complexity dereverberation for hands-free audio conferencing unit," *IEEE Transactions on Consumer Electronics*, Vol.61-4, 2015.

[4] A. Farina, A. Capra, L. Chiesi, and L. Scopece, "A spherical microphone array for synthesizing virtual directive microphones in live broadcasting and in post production," *40-th AES International Conference*, 2010.

[5] H. Wittek, C. Faller, C. Langen, A. Favrot, and C. Tournery, "Digitally enhanced shotgun microphone with increased directivity," *129-th AES Convention,* 2010.

[6] R. Oldfield, B. Shirley, and J. Spille, "Object-based audio for interactive football broadcast," *Multimedia Tools and Applications*, 2015.

[7] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and Gunshot Detection and Localization for Audio-Surveillance Systems," *In Proc. of IEEE Conference on Advanced Video and Signal Based Surveillance* (AVSS), pp.21–26, 2007.

[8] D. Conte, P. Foggia, G. Percannella, A. Saggese, and M. Vento, "An Ensemble of Rejecting Classifiers for Anomaly Detection of Audio Events," *In Proc.* of AVSS, 2012.

[9] M. K. Nandwana, A. Ziaei, and J. H. L. Hansen, “Robust Unsupervised Detection of Human Screams in Noisy Acoustic Environments,” *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), pp.161–165, 2015.

[10] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, “Audio Surveillance of Roads: A System for Detecting Anomalous Sounds,” *IEEE Transactions on Intelligent Transportation Systems*, Vol.17-1, pp.279–288, 2016.

[11] 丹羽 健太, “音源情報を推定するための拡散受音に関する研究,” 名古屋大学 博士論文, 2014.

[12] T. Oba, K. Kobayashi, H. Uematsu, T. Asami, K. Niwa, N. Kamado, T. Kawase, and T. Hori, “Media Processing Technology for Business Task Support,” *NTT Technical Review,* Vol.13-4, 2015.

[13] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. F. C. Yu, W. J. Fabian, M. Espi, T. Higuchi, S. Araki, and T. Nakatani, “The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices,” *in Proc. IEEE Automatic Speech Recognition and Understanding Workshop* (ASRU), 2015.

[14] J. Du, Y. H. Tu, L. Sun, F. Ma, H. K. Wang, J. Pan, C. Liu, and C. H. Lee, “The USTC-iFlytek System for CHiME-4 Challenge,” *Technical report of CHiME-4,* 2016.

[15] K. Niwa, T. Nishino, and K. Takeda, “Encording large array signals into a 3D sound field representation for selective listening point audio based on blind source separation,” *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), pp.181–184, 2008.

[16] S. Koyama, K. Furuya, Y. Hiwasaki, and Y. Haneda, “Analytical approach to wave field reconstruction filtering in spatio-temporal frequency domain,” *IEEE Transactions on Audio, Speech, and Language Processing*, Vol.21-4, 2013.

[17] Y. Jiaxing, M. Iwata, T. Kobayashi, M. Murakawa, T. Higuchi, Y. Kubota, Y. Toshiya, and K. Mori, “Statistical Impact-Echo Analysis based on Grassmann Manifold Learning: Its Preliminary Results for Concrete Condition Assessment,” *in Proc. of European Workshop on Structural Health Monitoring*, 2014.

[18] Y. Kubota, Y. E. Jiaxing, M. Iwata, M. Murakawa, and T. Higuchi, “Defect Detection for RC Slab based on Hammering Echo Acoustic Analysis,” *In Proc. of the 30th US-Japan Bridge Engineering Workshop.*

[19] Y. Haneda, S. Makino and Y. Kaneda, “Common acoustical pole and zero modeling of room transfer functions,” *IEEE Transactions on Speech and Audio Processing*, pp.320–328, 1994.

[20] Y. Haneda, S. Makino, Y. Kaneda and N. Kitawaki, “Common-acoustical-pole and zero modeling of head-related transfer functions,” *IEEE Transactions on Speech and Audio Processing*, pp.188–196, 1999.

[21] S. Koyama, K. Furuya, Y. Hiwasaki and Y. Haneda, “Reproducing virtual sound sources in front of a loudspeaker array using inverse wave propagator,” *IEEE Transactions on Speech and Audio Processing*, pp.1746–1758, 2012.

[22] S. Koyama, K. Furuya, Y. Hiwasaki, Y. Haneda and Y. Suzuki, “Wave field reconstruction filtering in cylindrical harmonic domain for with-height recording and reproduction,” *IEEE Transactions on Speech and Audio Processing*, pp.1546–1557, 2014.

[23] 浅野 太, “音のアレイ信号処理-音源の定位・追跡と分離-(第 1 版),” コロナ社, 2012.

[24] Y. Hioka, K. Furuya, K. Kobayashi, K. Niwa, and Y. Haneda, “Underdetermined sound source separation using power spectrum density estimated by combination of directivity gain,” *IEEE Transactions on Audio, Speech, and Language Processing*, Vol.21-6, pp.1240–1250, 2013.

[25] P. Coucke, B. De. Ketelaere, and J. De. Baerdemaeker, “Experimental analysis of the dynamic, mechanical behavior of a chicken egg,” *Journal of Sound and Vibration*, Vol. 266, pp.711–721, 2003.

[26] 陳 鵬, 馮 芳, 豊田 利夫, 劉 信芳, 嶋津 弘志, 平野 竜也, “歯車装置異常時の動特性及び異常診断法に関する研究 (第 1 報, 平歯車異常時の振動方程式および偏心状態の動特性),” 日本機械学会論文集（C 編）, pp.3258–3263, 2000.

[27] L. R. Bahl, F. Jelinek and R. L. Mercer, “A Maximum Likelihood Approach to Continuous Speech Recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol.5-2, pp.179–190, 1983.

[28] M. H. Johnson, " Pattern recognition in acoustic signal processing," *The Journal of the Acoustical Society of America*, 125(4), 2009.

[29] S. Koyama, K. Furuya, Y. Hiwasaki and Y. Haneda, "MAP estimation of driving signals of loudspeakers for sound field reproduction from pressure measurements," *in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (WASPAA), 2013.

[30] L. Deng, J. Droppo, and A. Acero, "Dynamic Compensation of HMM Variances Using the Feature Enhancement Uncertainty Computed From a Parametric Model of Speech Distortion," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol.13-3, pp.412–421, 2005.

[31] T. Nakatani, S. Araki, T. Yoshioka, M. Delcroix, and M. Fujimoto, "Dominance based integration of spatial and spectral features for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol.21-12, pp.2516–2531, 2013.

[32] H. Erdogan, J. R. Hershey, S. Watanabe, and J. L. Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 2015.

[33] K. Niwa, Y. Koizumi, T. Kawase, K. Kobayashi, and Y. Hioka, "Pinpoint extraction of distant sound source based on DNN mapping from multiple beamforming outputs to prior SNR," *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), pp.435–439, 2016.

[34] T. Kawase, K. Niwa, M. Fujimoto, N. Kamado, K. Kobayashi, S. Araki, and T. Nakatani, "Real-time integration of statistical model-based speech enhancement with unsupervised noise PSD estimation using microphone array," *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), pp.604–608, 2016.

[35] Y. Kubo, S. Wiesler, R. Schlueter, H. Ney, S. Watanabe, A. Nakamura and T. Kobayashi "Subspace pursuit method for kernel-log-linear models," *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 2011.

[36] Y. LeCun, Y. Bengio and G. Hinton “Deep learning,” *Nature*, pp.436–444, 2015.

[37] 岡谷 貴之, “深層学習,” 講談社, 2015.

[38] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, “Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups,” *IEEE Signal Processing Magazine*, Vol.29-6, pp.82–97, 2012.

[39] Y. LuCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard and L. D. Jackel, “Back-propagation applied to handwritten zip code recognition,” *Neural Computation,* 1(4), pp.541–551, 1989.

[40] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE,* Vol.86-11, 1998.

[41] J. J. Hopfield, “Neural network and physical systems with emergent collective computational abilities,” *in Proc. of the National Academy of Sciences of the United States of America*, 79 (8): 1982.

[42] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, Vol.9, Issue 8, pp. 1725–1780, 1997.

[43] O. Abdel-Hamid, A. Mohamed, H. Jiang, and G. Penn, “Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition,” *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), pp.4277–4280, 2012.

[44] H. Sak, A. W. Senior and F. Beaufays, “Long short-term memory recurrent neural network architectures for large scale acoustic modeling,” *in Proc. Interspeech,* pp.338–342, 2014.

[45] K. Niwa, Y. Koizumi, T. Kawase, K. Kobayashi and Y. Hioka, “Supervised Source Enhancement Composed of Non-negative Auto-Encoders and Complementarity Subtraction” *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 2017.

[46] P. Smaragdis and S. Venkataramani, “A Neural Network Alternative to Non-Negative Audio Models,” *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 2017.

[47] Y. Xu, J. Du, L. R. Dai and C. H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, pp.7–19, 2015.

[48] D. Bagchi, M. Mandel, Z. Wang, Y. He, A. Plummer and E. F. Lussier, "Combining spectral feature mapping and multi-channel model-based source separation for noise-robust automatic speech recognition," *in Proc. IEEE Automatic Speech Recognition and Understanding Workshop* (ASRU), 2015.

[49] V. Pulkki and M. Karjalainen, "Communication Acoustics: An Introduction to Speech, Audio and Psycoacoustics," *John Wiley and Sons, Ltd*, 2015.

[50] ITU-T Recommendation P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," 2001.

[51] ITU-T Recommendation P.863, "Perceptual objective listening quality assessment," 2011.

[52] V. Emiya, E. Vincent, N. Harlander and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, Vol.19-7, pp. 2046–2057, 2011.

[53] V. Chandola, A. Banerjee, and V. Kumar "Anomaly detection: A survey," *ACM Computing Surveys*, 2009.

[54] D. H. Jonson and D. E. Dudgeon, "*Array signal processing: concepts and techniques,*" *Prentice-Hall Series in Signal Processing,* 1993.

[55] M. Brandstein and D. Ward (Eds.), "*Microphone Arrays,*" *Digital Signal Processing,* Springer, 2001.

[56] J. L. Flanagan, D. A. Berkley, G. W. Elko, J. E. West, and M. M. Sondhi, "Autodirective microphone systems," *Acoustica,* Vol.73-2, pp.58–71, 1991.

[57] J. Benesty, S. Makino and J. Chen (Eds.), "*Speech Enhancement*" *Springer*, 2005.

[58] K. Kobayashi, K. Furuya, and A. Kataoka, "A Talker-Tracking Microphone Array for Teleconferencing," *113-th AES convention,* 2002.

[59] K. Niwa, Y. Hioka, K. Furuya, and Y. Haneda, “Diffused sensing for sharp directive beamforming,” *IEEE Transactions on Audio, Speech, and Language Processing*, Vol.21, pp.2346–2355, 2013.

[60] K. Niwa, Y. Hioka, and K. Kobayashi, “Optimal Microphone Array Observation for Clear Recording of Distant Sound Sources,” *IEEE Transactions on Audio, Speech, and Language Processing*, Vol.24-10, pp.1785–1795, 2016.

[61] Y. Koyano, K. Yatabe and Y. Oikawa, “Infinite-Dimensional SVD for Analyzing Microphone Array,” *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 2017.

[62] H. Sawada, S. Araki, R. Mukai and S. Makino, “Grouping separated frequency components with estimating propagation model parameters in frequency-domain blind source separation,” *IEEE Transactions on Audio, Speech, and Language Processing*, Vol.15-5, pp.1592–1604, 2007.

[63] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, “Multichannel extensions of non-negative matrix factorization with complex-valued data,” *IEEE Transactions on Audio, Speech, and Language Processing*, Vol.21-5, pp.971–982, 2013.

[64] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, “Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization,” *IEEE Transactions on Audio, Speech, and Language Processing*, Vol.24-9, pp.1626–1641, 2016.

[65] S. Makino, T. W. Lee and H. Sawada (Eds.), “*Blind Speech Separation,*” *Springer,* 2007.

[66] R. Zelinski, “A microphone array with adaptive post-filtering for noise reduction in reverberant rooms,” *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), pp.2578–2581, 1988.

[67] M. Aoki, M. Okamoto, S. Aoki, H. Matsui, T. Sakurai, and Y. Kaneda, “Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones,” *Acoustical science and technology,* Vol.22-2, pp.149–157, 2001.

[68] I. A. McCowan abd H. Bourlard, “Microphone array post-filter based on noise field coherence,” *IEEE Transactions on Audio, Speech, and Language Processing*, Vol.11, pp.709–716, 2003.

[69] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator,” *IEEE Transactions on Audio, Speech and Language Processing*, pp.1109–1121, 1984.

[70] A. Narayanan and D. Wang, “Ideal ratio mask estimation using deep neural networks for robust speech recognition,” *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 2013.

[71] O. Yilmaz and S. Rickard, “Blind Separation of Speech Mixtures via Time-Frequency Masking,” *IEEE Transactions on Audio, Speech and Language Processing*, pp.1830–1847, 2004.

[72] P. Smaragdis, and J. C. Brown, “Non-negative Matrix Factorization for Polyphonic Music Transcription,” *in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (WASPAA), 2003.

[73] S. Voran, “Exploring of Additivity Approximation for Spectral Magnitude,” *in Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (WASPAA), 2015.

[74] S. F. Boll, “Suppression of Acoustic Noise in Speech using Spectral Subtraction,” *IEEE Transactions on Audio, Speech and Signal Processing*, pp.113–120, 1979.

[75] K. Yamanishi, J. Takeuchi, G. J. Williams, and P. Milne, “On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms,” *In Proc. of ACM Conference on Knowledge Discovery and Data Mining* (SIGKDD), pp.320–324, 2000.

[76] T. Ide and H. Kashima, “Eigenspace-based anomaly detection in computer systems,” *In Proc. of ACM Conference on Knowledge Discovery and Data Mining* (SIGKDD), pp.440–229, 2004.

[77] M. Takimoto, M. Matsugu, and M. Sugiyama, “Visual inspection of precision instruments by least-squares outlier detection,” *In Proc. of IEEE Workshop on Distributed Mobile Systems & IoT Services* (DMSS), pp.22–26, 2009.

[78] S. Liu, T. Suzuki and M. Sugiyama, “Support consistency of direct sparse-change learning in Markov networks,” *In Proc. of Conference on Artificial Intelligence* (AAAI), pp.2701–2725, 2015.

[79] T. Ide, A. Khandelwal, and J. Kalagnanam, “Sparse Gaussian Markov Random Field Mixtures for Anomaly Detection,” *In Proc. of the IEEE International Conference on Data Mining series* (ICDM), pp.955–960, 2016.

[80] D. Chakrabarty and M. Elhilali, “Abnormal Sound Event Detection using Temporal Trajectories,” *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 2016.

[81] A. Mohamed, G. E. Dahl, G. Hinton, “Acoustic Modeling using Deep Belief Networks,” *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 20, Issue 1, pp.14–22, 2012.

[82] D. E. Rumelhart, G. Hinton, R. J. Williams, “Learning representations by back-propagating errors,” *Nature* Vol. 323(6088): pp.533–536, 1986.

[83] J. Duchi, E. Hazan and Y. Singer, “Adaptive Subgradient Methods for Online Learning and Stochastic Optimization,” *The Journal of Machine Learning Research*, pp.2121-2159, 2011.

[84] D. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization” *In Proc. of the 3rd International Conference for Learning Representations* (ICLR), 2015.

[85] G. E. Hinton, A. Osindero, Y. W. Teh, “A fast learning algorithm for deep belief nets,” *Neural Computation,* 18(7), pp.1527–1554, 2006.

[86] F. Seide, G. Li, X. Chen and D. Yu “Feature engineering in context-dependent deep neural networks for conversational speech transcription,” *in Proc. IEEE Automatic Speech Recognition and Understanding Workshop* (ASRU), pp. 24–29, 2011.

[87] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” *Journal of Machine Learning Research*, 2015.

[88] J. Schluter and S. Bock, “Improved Musical Onset Detection with Convolutional Neural Networks,” *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 2014.

[89] P. Werbos, “Backpropagation through time: What it does and how to do it,” *in Proceedings of the IEEE,* 78(10), pp.1550–1560, 1990.

[90] A. Graves, G. Wayne, M. Reynolds, T. Harley, I. Danihelka, A. G. Barwinska, S. G. Colmenarejo, E. Grefenstette, T. Ramalho, J. Agapiou, A. P. Badia, K. M. Hermann, Y. Zwols, G. Ostrovski, A. Cain, H. King, C. Summerfield, P. Blunsom, K. Kavukcuoglu and D. Hassabis, “Hybrid computing using a neural network with dynamic external memory,” *Nature* Vol. 538, pp.471–476, 2016.

[91] F. Weninger , H. Erdogan, S. Watanabe, E. Vincent, J. L. Roux, J. R. Hershey, and B. Schuller, “ Speech Enhancement with LSTM Recurrent Neural Networks and its Application to Noise-Robust ASR,” *in Proc. of International Conference on Latent Variable Analysis and Signal Separation* (LVA/ICA), 2015.

[92] D. D. Lee and S. Seung, “Algorithms for Non-negative Matrix Factorization,’ *in Proc. of Proceedings of Neural Information Processing Systems* (NIPS), pp.556–562, 2000.

[93] C. Fevotte, N. Bertin and J. L. Durrieu, “Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis,” *Neural Computation*, Vol.21(3), pp.793–830, 2009.

[94] D. P. Kingma, and M. Welling, “Auto-encoding variational Bayes,” *in Proc. of the International Conference on Learning Representations* (ICLR), 2014.

[95] I. Goodfellow, J. P. Abadie, M. Mirza, B. Xu, D. W. Farley, S. Ozair, A. Courville and Y. Bengio, “Generative Adversarial Nets,” *in Proc. of Advances in Neural Information Processing Systems* (NIPS), pp.2672–2680,

[96] P. Dayan, G. Hinton, R. Neal and R. Zemel “The Helmholtz Machine,” *Neural Computation*, 7(5), pp.889–904, 1995.

[97] G. Hinton, P. Dayan, B. J. Frey and R. Neal, “The wake-sleep algorithm for unsupervised neural networks,” *Science,* 268(5214), pp.1158–1161, 1995.

[98] M. Sugiyama, “*Statistical reinforcement learning: modern machine learning approaches,*” *Chapman and Hall/CRC*, 2015.

[99] G. J. Tesauro, “Temporal difference learning and TDGammon,” *Communications of the ACM*, pp.58–68, 1995.

[100] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, 518, pp. 529–533, 2015.

[101] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel and D. Hassabis, "Mastering the game of Go with deep neural networks and tree search," *Nature*, pp.484–489, 2016.

[102] T. Zhao, H. Hachiya, G. Niu and M. Sugiyama, "Analysis and Improvement of Policy Gradient Estimation," *in Proc. of Advances in Neural Information Processing Systems* (NIPS), 2011.

[103] Z. Q. Wang and D. Wang, "Recurrent Deep Stacking Networks for Supervised Speech Separation," *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 2017.

[104] K. Kinoshita, M. Delcroix, A. Ogawa, T. Higuchi and T. Nakatani, " Deep mixture density network for statistical model-based feature enhancement," *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 2017.

[105] M. Kim, Collaborative Deep Learning for Speech Enhancement: A Run-time Model Selection Method using Auto-encoders," *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 2017.

[106] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 2016.

[107] Y. Luo, Z. Chen, J. R. Hershey, J. L. Roux, and N. Mesgarani, "Deep Clustering and Conventional Networks for Music Separation: Stronger Together," *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 2017.

[108] D. S. Williamson, Y. Wang and D. L. Wang, "Complex Ratio Masking for Monaural Speech Separation," *IEEE Transactions on Audio, Speech and Language Processing*, pp.483–492, 2016.

[109] C. V. Cotton and D. P. W. Ellis, "Spectral vs. Spectrotemporal Features for Acoustic Event Detection," *in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (WASPAA), 2011.

[110] X. Lu, Y. Tsao, S. Matsuda and C. Hori, "Sparse Representation based on a Bag of Spectral Exemplars for Acoustic Event Detection," *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), pp.6255–6259, 2014.

[111] J. Schroder, S. Goetze and J. Anemuller, "Spectro-Temporal Gabor Filterbank Features for Acoustic Event Detection" *IEEE Transactions on Audio, Speech, and Language Processing*, 2015.

[112] M. Espi, M. Fujimoto, K. Kinoshita and T. Nakatani, "Exploiting Spectro-Temporal Locality in Deep Learning based Acoustic Event Detection," *Journal on Audio, Speech, and Music Processing* (EURASIP), 2015.

[113] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," *in Proc. of Proceedings of Neural Information Processing Systems* (NIPS) pp.1096–1104, 2009.

[114] K. Hamasaki, K. Hiyama and R. Okumura, "The 22.2 Multichannel Sound System and Its Application," *AES 118th Convention*, 2005.

[115] K. Matsui and A. Ando, "Binaural Reproduction of 22.2 Multichannel Sound with Loudspeaker Array Frame," *AES 135th Convention*, 2013.

[116] C. Q. Robinson, S. Mehta and N. Tsingos, "Scalable Format and Tools to Extend the Possibilities of Cinema Audio," *in Proc. of Society of Motion Picture and Television Engineers* (SMPTE), pp. 1–12, 2012.

[117] J. Engdegard, B. Resch, C. Falch, O. Hellmuth, J. Hilpert, A. Holzer, L. Terentiev, J. Breebaart, J. Koppens, E. Schuijers, and W. Oomen, "Spatial audio object coding (SAOC) - The upcoming MPEG standard on parametric object based audio coding," *AES 124th Convention*, 2008.

[118] J. Herre, J. Hilpert, A. Kuntz and J. Plogsties, “MPEG-H 3D Audio - The New Standard for Coding of Immersive Spatial Audio,” *IEEE Journal of Selected Topics in Signal Processing*, Vol.9, Issue 5, pp.770–779, 2015.

[119] ISO/IEC 14496-3:2009, Information technology - Coding of audio-visual objects - Part 3: Audio (4th Edition), Subpart 11.

[120] A. Hilton, J. Y. Guillemaut, J. Kilner, O. Grau and G. Thomas, “3D-TV Production From Conventional Cameras for Sports Broadcast,” *IEEE Transactions on Broadcasting*, Vol. 57, pp.462–476, 2011.

[121] M. Tanimoto, “FTV: Free-viewpoint Television,” *Image Communication*, Vol. 27, No. 6, pp. 555-570, 2012.

[122] A. Farina, A. Capra, L. Chiesi and L. Scopece, “A Spherical Microphone Array for Synthesizing Virtual Directive Microphones in Live Broadcasting and in Post Production,” *AES 40th International Conference*, 2010.

[123] H. Wittek , C. Faller, A. Favrot, C. Tournery, C. Langen, “Digitally Enhanced Shotgun Microphone with Increased Directivity,” *AES 129th Convention*, 2010.

[124] T. Hastie, R. Tibshirani and J. Friedman, “*The Elements of Statistical Learning: Data Mining, Inference, and Prediction*,” *Springer*, 2009.

[125] D. Guo, S. Shamai (Shitz) and S. Verdu, “Mutual information and minimum mean-square error in Gaussian channels,” *IEEE Transactions on Information Theory*, Vol.51, No.4, pp.1261–1282, 2005.

[126] H. Hino and N. Murata, “A conditional entropy minimization criterion for dimensionality reduction and multiple kernel learning,” *Neural Computation*, vol. 22, pp.2887–2923, 2010.

[127] T. Suzuki and M. Sugiyama, “Sufficient dimension reduction via squared-loss mutual information estimation,” *in Proc. of International Conference on Artificial Intelligence and Statistics* (AISTATS), pp.804–811, 2010.

[128] K. Fukumizu, F.R. Bach and M.I. Jordan, “Dimension Reduction for Supervised Learning with Reproducing Kernel Hilbert Space,” *Journal of Machine Learning Research*, vol.5, pp.73–99, 2004.

[129] K. Fukumizu, F.R. Bach and M.I. Jordan, "Kernel Dimension Reduction in Regression," *Annals of Statistics*, Vol.37-4, pp.1871–1905, 2009.

[130] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society*, Series B, pp.49–67, 2007.

[131] A. Beck and M. Teboulle, "Fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J, Imaging Sciences,*, pp.183–202, 2008.

[132] D.A. Harville, "*Matrix Algebra From a Statistician's Perspective*," *Springer-Verlag New York,* 1997.

[133] M.D. Zeiler, "ADADELTA: An adaptive learning rate method ," arXiv:1212.5701, 2012. `http://arxiv.org/abs/1212.5701`

[134] E. Vincent, R. Gribonval and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, 14(4), pp.1462–1469, 2006.

[135] J.P. Cunningham and Z. Ghahramani, "Linear Dimensionality Reduction: Survey, Insights, and Generalizations," *Journal of Machine Learning Research*, pp. 2859–2900, 2015.

[136] S. Srinivasan, N. Roman, and D.L. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication*, vol.48, pp.1486–1501, 2006.

[137] Y. Stylianou, O. Cappe and E. Moulines, "Continuous Probabilistic Transform for Voice Conversion," *IEEE Transaction on Speech, Audio Processing*, Vol.6, pp.131–142, 1998.

[138] A. Hyvarinen, et al., "*Independent Component Analysis*, " *J. Wiley, New York*, 2001.

[139] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine Learning*, 1992.

[140] T. Tieleman and G. Hinton, "Lecture 6.5 - RMSprop, *COURSERA: Neural Networks for Machine Learning*, 2012.

[141] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech communication*, pp.357–363, 1990.

[142] J. Barker, R. Marxer, E. Vincent and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: dataset, task and baseline," *in Proc. IEEE Automatic Speech Recognition and Understanding Workshop* (ASRU), 2015

[143] M. Matsumoto and T. Nishimura, "Mersenne Twister: A 623-dimensionally Equidistributed Uniform Pseudorandom Number Generator," ACM Trans. on Modeling and Computer Simulations, 1998.

[144] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy Gradient Methods for Reinforcement Learning with Function Approximation," In *Proc. NIPS*, 1999.

[145] F. Weninger, J. R. Hershey, J. L. Roux and B. Schuller, "Discriminatively Trained Recurrent Neural Networks for Single-Channel Speech Separation," in *Proc. GlobalSIP*, 2014.

[146] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech," *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 19, pp.2125–2136, 2011.

[147] J. Neyman and E. S. Pearson, "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical Transactions of the Royal Society*, 1933.

[148] J. S. Soo, and K. K. Pang, "Multidelay Block Frequency Domain Adaptive Filter," *IEEE Transactions on Acoustic, Speech and Signal Processing*, Vol.38-2, pp.373–376, 1990.

[149] T. Higuchi and H. Kameoka, "Joint audio source separation and dereverberation based on multichannel factorial hidden Markov model," *in Proc IEEE International Workshop on Machine Learning for Signal Processing* (MLSP), 2014.

[150] N. Ono, Z. Rafii, D. Kitamura, N. Ito and A. Liutkus. "The 2015 Signal Separation Evaluation Campaign," *in Proc. of International Conference on Latent Variable Analysis and Signal Separation* (LVA/ICA) pp. 387–395, 2015.

謝辞

本論文は，筆者が日本電信電話株式会社（NTT）メディアインテリジェンス研究所および電気通信大学で行った 3 年間の研究成果をまとめたものです．本研究を遂行するにあたり，数多くの方々に御指導と御援助を賜りました．特に御世話になった方々をここに記し，深い感謝の意を表します．

指導教官である電気通信大学 羽田陽一教授には，本論文の構成や内容について丁寧な御教示や御指導を賜りました．本論文全般において数々の議論をさせていただきましたが，特に 4 章にあたる内容は，音の物理的性質や主観評価などの統計的信号処理とは異なる観点から何度も議論させていただいて生まれた研究成果です．お忙しい中，審査委員を務めていただいた電気通信大学の南泰浩教授，柳井啓司教授，庄野逸教授，橋本直己准教授からは，本論文の内容や構成に関して有益な意見を頂きました．

法政大学 伊藤克亘教授には，筆者が高校 3 年生から修士課程を卒業するまで，音響信号処理の基礎から応用までを幅広くご教授頂きました．また伊藤教授には研究だけでなく，海外の大学訪問や就職活動など，公私ともに様々なご支援をいただきました．伊藤教授に教えていただいた，研究や議論の楽しさがあったからこそ研究者として生きていく決意ができました．

NTT において研究内容について特に多くのご指導をいただいたのが丹羽健太博士です．本論文の全ての内容は丹羽博士にいただいた熱心なご指導と議論の末に完成した内容であり，丹羽博士なくして本研究は完遂しえなかったと確信しております．また研究だけでなく，研究者としての姿勢や物事の捉え方など，丹羽博士から受けた影響は多大なものでした．またオークランド大学の日岡祐輔博士にも多くのご指導をいただきました．日岡博士には音源強調に関して熱心に議論して頂くと共に，論文の書き方について懇切丁寧にご指導頂きました．

NTT メディアインテリジェンス研究所 小澤英昭所長には，本研究ご理解頂くと共に，温かい激励を頂きました．音声言語メディアプロジェクトの 高橋敏プロジェクトマネージャ、および音環境情報処理グループの元グループリーダである 大室仲氏と現グループリーダである 原田登博士には，上長として本研究にご理解頂き，本研究遂行の機会を与えて頂きました．音環境情報処理グループの小林和則博士には，音源強調に関して様々

な議論をしていただきました．3 章にあたる研究の源流は筆者が入社時に小林博士と丹羽博士が進めていた研究であり，実際のスポーツフィールドでの実験やシステム構築など，実環境でシステムを動かすための技術を広く教えていただきました．音環境情報処理グループの植松尚博士，中川朗氏，齊藤翔一郎氏，河内祐太氏には，5 章と付録 A の内容について多くの議論をしていただきました．異常音検知技術を実環境で運用している中で，技術的にもビジネス的にも様々な課題に直面しました．同氏らの協力なくしては本研究の遂行は不可能であったと確信しております．音環境情報処理グループの島内末廣博士，江村暁博士，栗原祥子氏，伊藤弘章氏，川瀬智子氏，矢澤櫻子博士，村田伸博士には，本研究を進めるにあたり様々なアドバイスをいただきました．島内博士からは 3, 4 章の内容をさらに改善するアイディアを多数いただいており，頂いたアイディアを活かしながら今後の研究を進めていくのが今から楽しみで仕方ありません．コミュニケーション科学基礎研究所の中谷智広上席特別研究員，荒木章子博士，木下慶介博士，情報通信研究機構の藤本雅清博士には，統計的音響信号処理の基礎を教えていただきました．

NTT データの北村唯夫氏，江口寛之氏，福島悠人氏，齋藤洋氏には，5 章の内容の実環境試験やデータ収集で大変お世話になりました．同氏らからは，実環境でのソフト/ハードウェア制約や現場のニーズなど，研究者とは違った視点からのコメントを多数いただきました．研究室で生まれた技術がいち早く世の中に出ていったのは，同氏らと交わした多角的な議論とご協力のおかげです．

本研究の成果の実環境運用にあたり，数々の関係者の皆様にお世話になりました．個々の御名前は省略させて頂きますが，NTT メディアインテリジェンス研究所音声言語メディアプロジェクトの皆様には数多くの御討論をさせて頂きました．学会を通じて，数多くの諸先輩方から温かい激励やコメントを頂きました．本研究の遂行は，そのような方々の熱心な御支援により成し得たものです．関係者の皆様方に感謝の意を表します．

最後に，研究を何不自由なく行えるよう支援してくれた親族と，常に陰から惜しみなく支え続けてくれた妻の桂子に心からの感謝の意を込めて本論文を捧げます．

関連論文

1. **Y. Koizumi**, K. Niwa, Y. Hioka, K. Kobayashi, and H. Ohmuro, "Informative Acoustic Feature Selection to Maximize Mutual Information for Collecting Target Sources," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol.25-4, pp.768–779, 2017. （3 章の内容に相当，研究業績リスト [J-1]）

2. **Y. Koizumi**, K. Niwa, Y. Hioka, K. Kobayashi, and Y. Haneda, "DNN-based Source Enhancement Self-Optimized by Reinforcement Learning using Sound Quality Measurements," *in Proc. of International Conference on Acoustics, Speech and Signal Processing* (ICASSP 2017), pp.81–85, 2017. （4 章の内容に相当，研究業績リスト [C-2]）

3. **Y. Koizumi**, S. Saito, H. Uematsu, and N. Harada, "Optimizing Acoustic Feature Extractor for Anomalous Sound Detection Based on Neyman-Pearson Lemma," *in Proc. of the 25th European Signal Processing Conference* (EUSIPCO 2017), pp.728–732, 2017. （5 章の内容に相当，研究業績リスト [C-1]）

研究業績リスト

受賞

[A-1] 日本音響学会 粟屋潔学術奨励賞 (2017)

[A-2] 日本電信電話株式会社 社長表彰 (2015)

[A-3] 情報処理学会 山下記念研究賞 (2015)

[A-4] 情報処理学会 船井ベストペーパー賞 (2013)

[A-5] 情報処理学会 第 75 回全国大会 学生奨励賞 (2013)

[A-6] 日本音響学会 第 6 回学生優秀発表賞 (2012)

[A-7] 文部科学省 第 1 回サイエンス・インカレ サイエンス・インカレ奨励表彰 (2012)

学術論文誌

[J-1] **Y. Koizumi**, K. Niwa, Y. Hioka, K. Kobayashi, and H. Ohmuro, "Informative Acoustic Feature Selection to Maximize Mutual Information for Collecting Target Sources," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol.25-4, pp.768–779, 2017.

[J-2] 小泉 悠馬，伊藤 克亘, "音量軌跡の遷移型状態空間表現に基づくダイナミックスとアーティキュレーションへの分解," 電子情報通信学会論文誌, Vol.J 98-D, No.3, 2015.

[J-3] 小泉 悠馬，伊藤 克亘, "連続励起振動楽器を対象としたノート内セグメンテーション," 電子情報通信学会論文誌, Vol.J 97-D, No.3, 2014.

[J-4] 小泉 悠馬, 伊藤 克亘, "擦弦楽器の意図表現合成のための奏法モデル," 情報処理学会論文誌，Vol.54，No.4，2013.

査読付き国際会議

[C-1] **Y. Koizumi**, S. Saito, H. Uematsu, and N. Harada, "Optimizing Acoustic Feature Extractor for Anomalous Sound Detection Based on Neyman-Pearson Lemma," *in Proc. of the 25th European Signal Processing Conference* (EUSIPCO 2017), 2017.

[C-2] **Y. Koizumi,** K. Niwa, Y. Hioka, K. Kobayashi, and Y. Haneda, "DNN-based Source Enhancement Self-Optimized by Reinforcement Learning using Sound Quality Measurements," *in Proc. of International Conference on Acoustics, Speech and Signal Processing* (ICASSP 2017), 2017.

[C-3] S. Shimauchi, S. Kudo, **Y. Koizumi**, and K. Furuya, "On Relationships between Amplitude and Phase of Short-Time Fourier Transform," *in Proc. of International Conference on Acoustics, Speech and Signal Processing* (ICASSP 2017), 2017.

[C-4] K. Niwa, **Y. Koizumi,** T. Kawase, K. Kobayashi, and Y. Hioka, "Supervised Source Enhancement Composed of Nonnegative Auto-Encoders and Complementarity Subtraction," *in Proc. of International Conference on Acoustics, Speech and Signal Processing* (ICASSP 2017), 2017.

[C-5] **Y. Koizumi,** K. Niwa, Y. Hioka, K. Kobayashi, and H. Ohmuro, "Integrated Approach of Feature Extraction and Sound Source Enhancement based on Maximization of Mutual Information," *in Proc. of International Conference on Acoustics, Speech and Signal Processing* (ICASSP 2016), 2016.

[C-6] K. Niwa, **Y. Koizumi**, T. Kawase, K. Kobayashi, and Y. Hioka, "Pinpoint Extraction of Distant Sound Source based on DNN Mapping from Multiple Beamforming Outputs to Prior SNR," *in Proc. of International Conference on Acoustics, Speech and Signal Processing* (ICASSP 2016), 2016.

[C-7] K. Niwa, **Y. Koizumi**, K. Kobayashi, and H. Uematsu, "Binaural Sound Generation Corresponding to Omnidirectional Video View using Angular Region-wise Source Enhancement," *in Proc. of International Conference on Acoustics, Speech and Signal Processing* (ICASSP 2016), 2016.

[C-8] **Y. Koizumi,** K. Niwa, Y. Hioka, K. Kobayashi, and H. Ohmuro, "Informative Acoustic Feature Selection on Microphone Array Wiener Filtering for Collecting Target Source on Sports Ground," *in Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (WASPAA 2015), 2015.

[C-9] **Y. Koizumi**, K. Itou, "Intra-note Segmentation via Sticky HMM with DP Emission," *in Proc. of International Conference on Acoustics, Speech and Signal Processing* (ICASSP 2014), 2014.

[C-10] **Y. Koizumi**, K. Itou, "Expressive Oriented Time-Scale Adjustment for Misplayed Musical Signals based on Tempo Curve Estimation," *in Proc. of the 16th International Conference on Digital Audio Effects Conference* (DAFx-16), 2013.

[C-11] **Y. Koizumi**, K. Itou, "Performance Expression Synthesis for Bowed-String Instruments using "Expression Mark Functions"," *Proceedings of Meetings on Acoustics* (POMA), 2012.

研究会・全国大会等

[D-1] 小泉 悠馬，丹羽 健太, 小林 和則, 羽田 陽一, " 聴感評点を向上させるための DNN 音源強調関数のブラックボックス最適化," 日本音響学会 2017 年秋季研究発表会, 2017.

[D-2] 小泉 悠馬，齊藤 翔一郎, 小林 和則, 島内 末廣，羽田 陽一, "広域に分散配置したマイクロホンを連携させる遠方雑音抑圧法の検討," 日本音響学会 2017 年秋季研究発表会, 2017.

[D-3] 河内 祐太，小泉 悠馬，原田 登, "L_p ノルム回帰を用いた異常音検知の検討," 日本音響学会 2017 年秋季研究発表会, 2017.

[D-4] 小泉 悠馬，齊藤 翔一郎, 植松 尚, "深層学習を用いた機器動作音の異常音検知," 日本音響学会 2017 年春季研究発表会, 2017.

[D-5] 島内 末廣，工藤 晋也，小泉 悠馬，古家 賢一, "短時間フーリエ変換の振幅と位相の依存性に関する考察," 日本音響学会 2017 年春季研究発表会, 2017.

[D-6] 小泉 悠馬，丹羽 健太, 小林 和則, 大室 仲, 羽田 陽一, " 聴感評点を最大化するための強化学習に基づく音源強調の検討 ," 日本音響学会 2016 年秋季研究発表会, 2016.

[D-7] 小泉 悠馬，丹羽 健太, 齊藤 翔一郎, 植松 尚, "機器動作音の異常音検知のための音響特徴量自動設計," 日本音響学会 2016 年秋季研究発表会, 2016.

[D-8] 丹羽 健太, 小泉 悠馬，川瀬 智子, 小林 和則, 日岡 裕輔, "対称性マイクロホンアレイを用いた目的音／雑音 PSD の推定," 日本音響学会 2016 年秋季研究発表会, 2016.

[D-9] 小泉 悠馬，丹羽 健太, 小林 和則, 大室 仲, “ガウシアンカーネルを用いた相互情報量最大化に基づく特徴選択; 音源強調を事例として,” 日本音響学会 2016 年春季研究発表会, 2016.

[D-10] 丹羽 健太, 小泉 悠馬，川瀬 智子, 小林 和則, 日岡 裕輔, “相互情報量増大型受音と DNN 回帰に基づく遠距離収音技術に関する検討,” 日本音響学会 2016 年春季研究発表会, 2016.

[D-11] 丹羽 健太，小泉 悠馬，小林 和則，越智 大介，亀田 明男，鎌本 優，守谷 健弘, “スマートホンを用いた全天球映像音声配信・視聴系の実装,” 日本音響学会 2016 年春季研究発表会, 2016.

[D-12] 小泉 悠馬，丹羽 健太, 小林 和則, 大室 仲, “ 競技音を抽出するための特徴選択と音源強調の統合的アプローチの検討,” 日本音響学会 2015 年秋季研究発表会, 2015.

[D-13] 丹羽 健太, 小泉 悠馬，小林 和則, “ 全天球映像に対応したバイノーラル音を生成するための方向別収音に関する検討,” 電子情報通信学会 技術研究報告, 2015.

[D-14] 小泉 悠馬，丹羽 健太, 小林 和則, 大室 仲, “ 複数領域を区別して収音するためのウィーナーフィルタ設計技術,” 日本音響学会 2015 年春季研究発表会, 2015.

[D-15] 小泉 悠馬，伊藤 克亘, “ディリクレ過程を出力する Nest 型 HMM を用いた音符内状態推定,” 日本音響学会 2014 年春季研究発表会, 2014.

[D-16] 小泉 悠馬，伊藤 克亘, “連続励起振動楽器を対象とした音量軌跡のダイナキクスとアーティキュレーションへの分解法,” 情報処理学会研究報告, SIGMUS-102, 2014.

[D-17] 安田 沙弥香, 小泉 悠馬，伊藤 克亘, “ラジオ放送話者ダイアライゼーション,” 情報処理学会 第 76 回全国大会, 2014.

[D-18] 塩出 萌子, 小泉 悠馬，伊藤 克亘, “中間話者コーパスを用いたアニメーション演技音声のための話者変換,” 情報処理学会 第 76 回全国大会, 2014.

[D-19] 小泉 悠馬，伊藤 克亘, “奏者の意図したテンポ変動の推定に基づく演奏録音の自動伸縮修正法 ,” FIT2013 第 12 回情報科学技術フォーラム, 2013.

[D-20] 小泉 悠馬，伊藤 克亘, “連続励起振動楽器のためのパワーに基づく音符内状態推定,” 日本音響学会 2013 年秋季研究発表会, 2013.

[D-21] 小泉 悠馬，伊藤 克亘, “音楽表現の生成モデリングの検討 ～熟練度に依存しない演奏表現の解析技術を目指して～,” 情報処理学会研究報告, 2013-MUS-99-58, 2013.

[D-22] 小泉 悠馬，伊藤 克亘, “演奏音の音量時系列からの奏者の意図表現成分の推定,” 情報処理学会 第 75 回全国大会, 2013.

[D-23] 上野 涼平, 小泉 悠馬，伊藤 克亘, “音楽知識を利用したハーモナイザー,” 情報処理学会 第 75 回全国大会, 2013.

[D-24] 森田 花野, 小泉 悠馬，伊藤 克亘, “教則本を利用したギターフレーズの難易度推定,” 情報処理学会 第 75 回全国大会, 2013.

[D-25] 小泉 悠馬，伊藤 克亘, “演奏意図関数に基づく表現力を反映させた音響信号の伸縮修正,” 情報処理学会研究報告, 2012-MUS-97-02, 2012.

[D-26] 小泉 悠馬，伊藤 克亘, “意図表現における非周期擦弦振動を考慮した楽音合成手法の検討,” 日本音響学会 2012 年秋季研究発表会, 2012.

[D-27] 小泉 悠馬，伊藤 克亘, “擦弦時の奏法行動を考慮した意図表現の合成手法:VIOCODER,” 情報処理学会研究報告, 2012-MUS-95-02, 2012.

[D-28] **Y. Koizumi**，K. Itou, “Synthesis of Performance Expression of Bowed String Instruments using “Expression Mark Functions”,” The Acoustics 2012 Hong Kong Conference and Exhibition, 2012.

[D-29] 小泉 悠馬，伊藤 克亘, “合成音への表現力付与のための擦弦楽器の発想伝達関数の推定,” 情報処理学会 第 74 回全国大会, 2012.

著者略歴

小泉 悠馬 (Yuma Koizumi)

2012 年 法政大学 情報科学部卒，2014 年 同大学院 情報科学研究科修了．同年 日本電信電話株式会社 NTT メディアインテリジェンス研究所入所．音声・音響信号処理技術の研究開発に従事．2016 年 電気通信大学大学院 情報理工学研究科入学．IEEE，電子情報通信学会，日本音響学会各会員．