# 音情報の多様な観測を活用した音響シーン分析の研究

井本　桂右

博士（情報学）

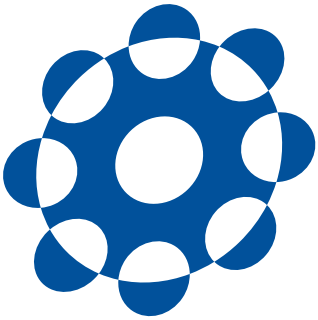
総合研究大学院大学
複合科学研究科
情報学専攻

平成２８（２０１６）年度

博士論文

# 音情報の多様な観測を活用した音響シーン分析の研究

井本桂右



総合研究大学院大学

SOKENDAI (The Graduate University for Advanced Studies)

2017年03月

本論文は総合研究大学院大学 複合科学研究科 情報学専攻に
博士(情報学)授与の要件として提出した博士論文である

# 審査委員

| | |
|---|---|
| Advisor | 小野順貴 (Nobutaka Ono) 准教授<br>国立情報学研究所 / 総合研究大学院大学 |
| Subadvisor | 山岸順一 (Junichi Yamagishi) 准教授<br>国立情報学研究所 / 総合研究大学院大学 |
| Subadvisor | 佐藤真一 (Shin'ichi Satoh) 教授<br>国立情報学研究所 |
| Examiner | 篠田浩一 (Koichi Shinoda) 教授<br>東京工業大学 |
| Examiner | 相原健郎 (Kenro Aihara) 准教授<br>国立情報学研究所 / 総合研究大学院大学 |

# 謝辞

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。

国立情報学研究所 小野順貴准教授には，本研究の方向性を決定づけるアイデアを多数頂くとともに，本論文の構成や内容について多大なる御指導を賜り深く感謝いたします。国立情報学研究所 山岸順一准教授には，音声や確率統計の観点から本研究の内容について有益なご意見を多数賜りましたことを深く感謝いたします。国立情報学研究所 佐藤真一教授には，パターン認識やメディア処理の観点から本研究の内容について有益なご意見を多数賜りましたことを深く感謝いたします。また，東京工業大学 篠田浩一教授，国立情報学研究所相原健郎准教授には本論文の執筆にあたり，有益なご意見を多数頂き，深く感謝いたします。

本研究を進めるにあたり，NTTコミュニケーション科学基礎研究所 大石康智博士には数多くの御指導と御助言を頂きました。総合研究大学院大学 北村大地様には日頃より本研究の内容についてご討論頂くとともに，多数のご支援を頂き感謝致します。国立情報学研究所 Le Trung Kien博士，越智景子博士，高木信二博士には日頃より本研究の内容についてご討論頂き，有益な助言を頂きました。国立情報学研究所 森元成子様には，研究活動を行う上で多数のご支援を賜りましたことを深く感謝致します。また，本研究を行うにあたり，辛抱強い支援により私を研究に没頭させてくれた家族に深く感謝します。

最後に，博士論文の執筆にあたりご支援頂いた全ての方に，私が尊敬する作曲家兼ピアノ奏者，オーケストラリーダーのDuke Ellington氏の言葉を借り，感謝の意を表します。

*Love you all madly*!!

# 概要

本論文では，様々な条件で観測された音情報を活用して，音が観測された場所や周囲の状況，周囲にいるユーザの行動などの音響シーンを分類する問題を扱う。近年，スマートホンやウェアラブルデバイス，IoT機器，街頭センサなど，気軽に利用可能なマイクロホンが急速に増加しており，これまでは活用されていなかった音声以外の様々な音を積極的に活用することを目指し，音響イベント検出や音響シーン分析といった研究が盛んに進められている。一方で，気軽に様々な音が収録可能になるにつれ，音の収録環境も多様化している。本論文では，これまでに音響シーン分類問題の観測条件として扱われていない，逐次的な観測や間欠的な欠損を有する観測，時間同期誤差を持つ多チャネル観測に対する音響シーン分類のための新たなアプローチを提案する。提案手法では，音のスパース性をモデルに導入でき，限られたデータからでも過学習することなく音響シーンをモデル化可能な音響トピックモデルに着目し，様々な観測条件に適用可能な新たな音響トピックモデルに基づくシーン分類を可能とする。

本論文ではまず，逐次的に新たな音データが得られる場合において，音響シーンをモデル化，分類する問題について検討を行う。逐次的に新たな音データが得られる場合，音データが得られる度に音響シーンモデルを再学習する必要があり計算コストが大きくなるという課題や，学習初期には限られた音データを用いて音響シーンをモデル化する必要があるため過学習に陥りやすいという課題がある。そこで本論文では，音響シーンに関する情報をパラメータとしてモデルに保存しておき，逐次的にこのパラメータを更新可能である点，また，音のスパース性をモデルに導入でき限られた音データから過学習することなく音響シーンをモデル化可能である点に着目し，音響トピックモデルに基づく逐次学習手法を提案する。提案手法では，従来の音響トピ

ックモデルの学習手法を逐次学習可能な手法として拡張することに加え，計算コストとパラメータ推定精度それぞれに利点を持つ，崩壊型変分ベイズ法の近似手法と崩壊型ギブスサンプリングを組合せたハイブリッド型パラメータ推定手法も併せて導入する。実環境収録音を用いた評価実験により，提案手法で学習されたモデルパラメータを用いることによりバッチ型学習法で学習した場合と同等の音響シーン分類が実現できることを示す。

続いて，観測の一部が欠損している音データから音響シーンを分類し，さらに欠損した観測を同時に推定する問題について検討を行う。本論文では，欠損した観測を潜在的な確率変数とみなすことで，欠損していない音響ワードの観測と欠損した観測を同一の生成モデルで扱うことができる点に着目し，音響トピックモデルを拡張した新たな音響シーン分類手法を提案する。提案モデルでは，音の時間連続性に基づき観測の時間遷移を教師あり音響トピックモデルに組み込むことで，欠損した観測を推定しながら音響シーンの分類を行うことを可能とする。具体的には，1) 音響ワードの時間遷移関係を単純マルコフモデルによりモデル化し，教師あり音響トピックモデルに組み込んだ，音響ワード遷移型教師あり音響トピックモデルと，2) 音響トピックの時間遷移関係を隠れマルコフモデルによりモデル化し，教師あり音響トピックモデルに組み込んだ，音響トピック遷移型教師あり音響トピックモデルを提案する。また，崩壊型ギブスサンプリングによる提案モデルのパラメータ推定方法も合わせて示す。実環境収録音を用いた性能評価実験により，提案手法では音響ワードの欠損率が50%程度になった場合でも，欠損がない場合と同程度の音響シーン分類性能が実現可能であることを示す。

さらに，時間同期誤差を持つ分散マイクロホンアレイによる多チャネル観測から音響シーンを分類する問題についても検討を行う。本論文では，音響トピックモデルにより空間情報を扱う方法について議論を行う。また，マイクロホン間の時間同期誤差に頑健な音響特徴抽出手法として，複数のマイクロホンで得られた振幅情報の対数に主成分分析を行うことで得られる，空間ケプストラムという新たな特徴抽出方法を提案する。空間ケプストラムは従来のマイクロホンアレイ処理と異なり，マイクロホンやスピーカの位置情報が不要であり，マイクロホン間の時間同期誤差にも頑健であるため，スマートホンなどが持つ複数のマイクロホンを用いた音響シーン分類に適していると言える。また，空間ケプストラムを一般化し，周波数情報と空間情報を同時に抽出可能な，一般化周波数-空間ケプストラムも併せて提案する。実環境

収録音を用いた音響シーン分類実験により，空間ケプストラムおよび一般化周波数-空間ケプストラムは周波数特徴量同様に音響シーン分類に効果的であることを示す。また，提案手法を用いることで，チャネル間に時間同期誤差がある場合でも頑健に音響シーン分類可能であることも併せて示す。

# 目次

# 図目次

# 表目次

# アルゴリズム目次

# 1 序論

## 1.1 研究背景

近年，スマートホンやウェアラブルデバイス，IoT機器，街頭センサなど気軽に利用可能なマイクロホンが急速に増加しており，これらのマイクロホンにより音声認識を利用して情報検索や機器の操作を行うことはすでに当たり前になりつつある。マイクロホンの増加に伴いあらゆる場面で音の収録が可能になったことを背景として，音声に限らないあらゆる音を積極的に活用する研究も盛んに行われており，発生音の種類を分析する音響イベント検出（AED: Acoustic event detection） [1, 2, 3, 4, 5, 6]や音が収録された場所や状況を分析する音響シーン分析（ASA: Acoustic scene analysis） [7, 8, 9]という分野に注目が集まっている。

音響イベント検出や音響シーン分析は，動画などのメディアへのタグ付け [10, 11, 6]や，自動監視システム [1, 12, 13, 14, 15]，高齢者や乳幼児の見守りシステム [3, 16]，自動ライフログシステム [17, 8]など様々なサービスに適用可能な技術として実用化が期待されているが解決すべき課題もまだ多い。本論文は，

表 1.1: 音響シーン分析で想定される観測条件と従来の音響シーン分析手法，従来手法の課題

| 観測条件 | 従来の音響シーン分析手法 | 従来法の課題 |
|---|---|---|
| 単一/多チャネル収録 | | |
| 1. 音の観測が逐次的 | 未知の音響シーンを既知のモデルからの外れ値として定義して分類[14, 18] | 新たな音響シーンをモデル化する場合は都度全てのデータを用いてモデル化処理が必要 |
| 2. 観測の一部が欠損 | – | – |
| 3. 符号化による観測信号の歪み | – | 従来の符号化方式を用いても信号に大きな歪みは生じない |
| 4. 高残響/雑音環境 | 前処理として残響除去/残響除去を適用 [19, 20, 21] | 残響除去/残響除去が困難な環境下では適用困難 |
| 多チャネル収録 | | |
| 5. マイクロホンの位置が不明 | チャネル毎に音響シーン分析した後，尤度に基づいて結果を統合[22] | 空間情報を十分に活用できていない |
| 6. チャネル間の時間同期が正確でない | チャネル毎に音響シーン分析した後，尤度に基づいて結果を統合[22] | 空間情報を十分に活用できていない |

その課題の一つである音情報の観測条件の多様性という観点から，音響シーン分析技術の一つである音響シーン分類問題に取り組む。

## 1.2 音響シーン分析における観測の多様性

スマートホンやIoT機器など気軽に利用可能なマイクロホンが増加することにより，あらゆる場面で音の収録が可能になる反面，理想的でない環境下で収録された音から音響シーンを分析する場面が増えつつある。音響シーン分析において課題となる観測の例と，それぞれの観測条件における音響シーン分析の従来手法の例を表1.1に示す。例えば，投稿型動画サイトのコンテンツの自動分類や日々の活動を記録し続けるライフログに音情報を利用する場合，音情報が逐次的に得られるため，全ての音情報を一度に用いた音響シーン分析，つまりバッチ処理による音響シーン分析を行うことが難しい。従来手法[14, 18]では，事前に得られた一部の音のみから音響シーンのモデルを作成し

ておき，既知のモデルから大きく外れた値を「未知の音響シーン」と定義することで音響シーン分析を実現している。しかしながら，これらの手法では新たな音響シーンをモデル化する場合，都度全てのデータを用いて音響シーンのモデル化処理が必要となる。投稿型動画サイトのコンテンツやライフログでは，蓄積される音データが膨大になるため全てのデータを用いて音響シーンのモデル化を繰り返し行う場合，計算コストや記憶容量が非常に大きな問題になる。そのため，繰り返しモデル化不要な手法の実現が投稿型動画サイトのコンテンツの自動分類や自動ライフログの実現には重要となる。

屋外で音を収録する場合には，背景雑音や音の過大な入力レベルにより観測に不要なノイズが混入したり部分的な欠損が起こることが考えられる。また，見守りや監視といった用途に音響シーン分析を適用する場合，プライバシーの観点から音の連続的な収録が困難である場合も多く，部分的に収録された音からシーン分析する状況も考えられる。高雑音環境下で収録された観測から音響シーンを分析する手法として[19, 20]が挙げられるが，これらの研究では音響シーン分類の前処理として雑音除去を行っているに過ぎず観測が部分的に欠損している場合には適用できない。一部が完全に欠損している観測から音響シーン分類を実現することは，音データの有効活用にもつながるため重要な課題である。

気軽に利用可能なマイクロホンが急速に増加していることで，複数のマイクロホンにより収録された多チャネル信号を用いて音響シーン分類を行うことも可能である。しかしながら，スマートホンやIoT機器などの複数マイクロホンは正確に時間同期されていることは稀であり，事前にマイクロホンの正確な位置を知ることも容易ではない。従来手法 [22]では，チャネル毎に音響シーン分類を行った後，各チャネルの音響シーンの尤度を用いて音響シーン分類結果を決定することでマイクロホンの位置やチャネル間の正確な時間同期を不要としている。しかしながら，従来手法は空間情報を十分に利用できていないため，より効果的に空間情報を利用する手法が期待される。

本論文では，様々な条件で観測された音から音響シーン分類を可能にする手法の実現を目指し，特に従来手法で実現困難である，1. 音の観測が逐次的，2. 音の観測の一部が欠損している，5. / 6. マイクロホンの位置が不明でチャネル間の時間同期が正確でない観測条件に対して適用可能な手法の提案を目指す。

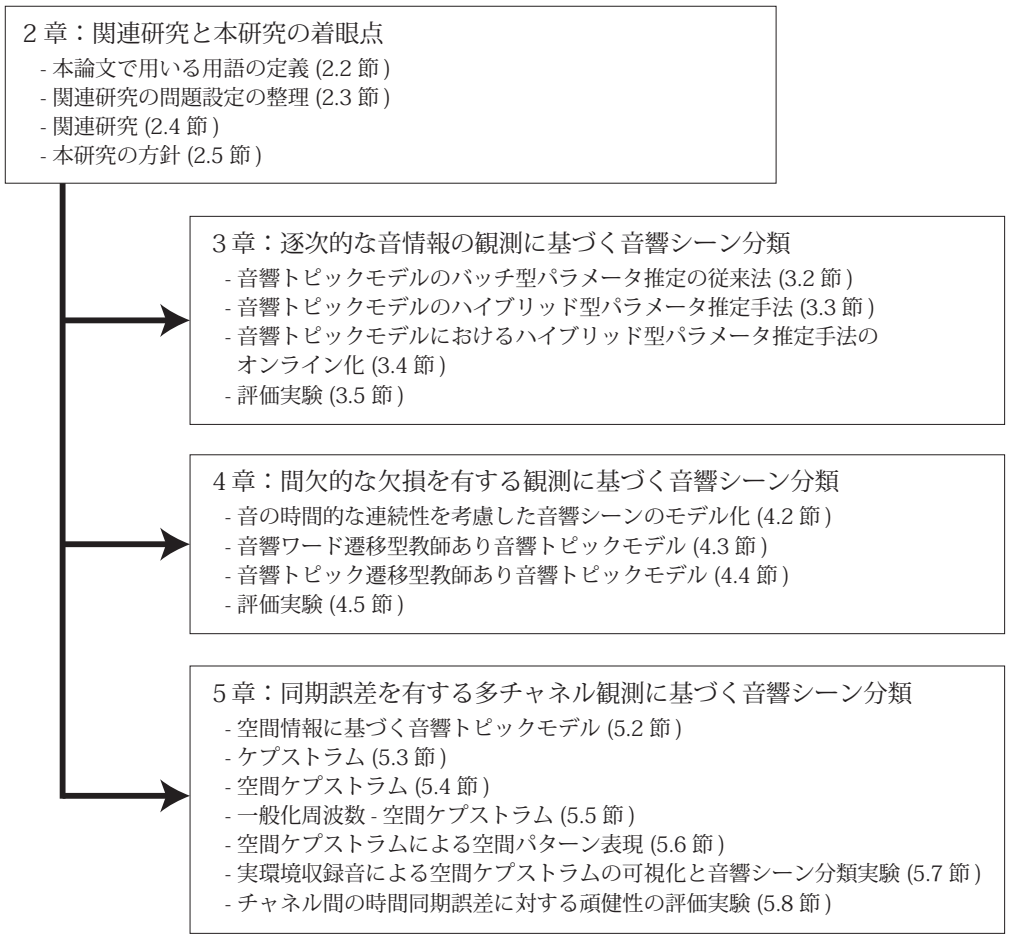



図 1.1: 本論文の構成

## 1.3 本論文の構成

本論文の構成を図1.1に示す。 2章では，まず音響イベント検出や音響シーン分析で用いられる用語と問題設定の整理を行う。その後，音響シーン分析の問題設定の一つであり本研究で取り扱う音響シーン分類の関連研究について述べ，従来研究の課題と本論文における音響シーン分類の着眼点および方針について述べる。 3章では，逐次的に得られた音情報を用いて音響シーン分析することが可能な音響トピックモデルを提案し，提案手法により高精度にモデルをオンライン学習できることを示す。 4章では，欠損を含む音の観測から音響シーンをモデル化/分類し，さらに欠損した観測を同時に推定可能な手法として，音の時間遷移をモデルに組み込んだ音響トピックモデルを提案し，欠

損を含む観測に対しても効果的に音響シーンを分類可能であることを示す。5章では，多数の分散配置されたマイクロホンで得られた観測から音響シーン分類を行う手法について検討する。まず，音響トピックモデルにより空間情報を扱う方法について議論を行い，その後，空間情報を用いた音響トピックモデルに利用可能な空間特徴として，空間ケプストラムと呼ばれる特徴量を提案する。空間ケプストラムはマイクロホンやスピーカの位置情報が不要であり，マイクロホン間の時間同期誤差にも頑健であるため，分散マイクロホンアレイを用いた音響シーン分類のための特徴量に適していることを評価実験により示す。最後に，6章で本論文の結論を述べる。

# 2 関連研究と本研究の着眼点

## 2.1 はじめに

本章ではまず，本論文で用いる用語の定義と関連研究の問題設定の整理を行う。その後，音響シーン分類に広く用いられている機械学習に基づく手法を概説し，従来研究の課題と本論文で提案する音響シーン分類手法の着眼点について述べる。 2.2節では音響イベント検出や音響シーン分析で用いられる用語の定義を行い，2.3節において詳細な問題設定の整理を行う。また2.4節では，音響シーン分類の従来手法の概説と，音響シーン分類問題において特に有効なアプローチである音響トピックモデルについて述べ，2.5節において，本論文で取り扱う課題においても音響トピックモデルが有効であることを述べる。

## 2.2 本論文で用いる用語の定義

音声や楽音に限らないあらゆる音情報から発生音の種類や音が収録された場

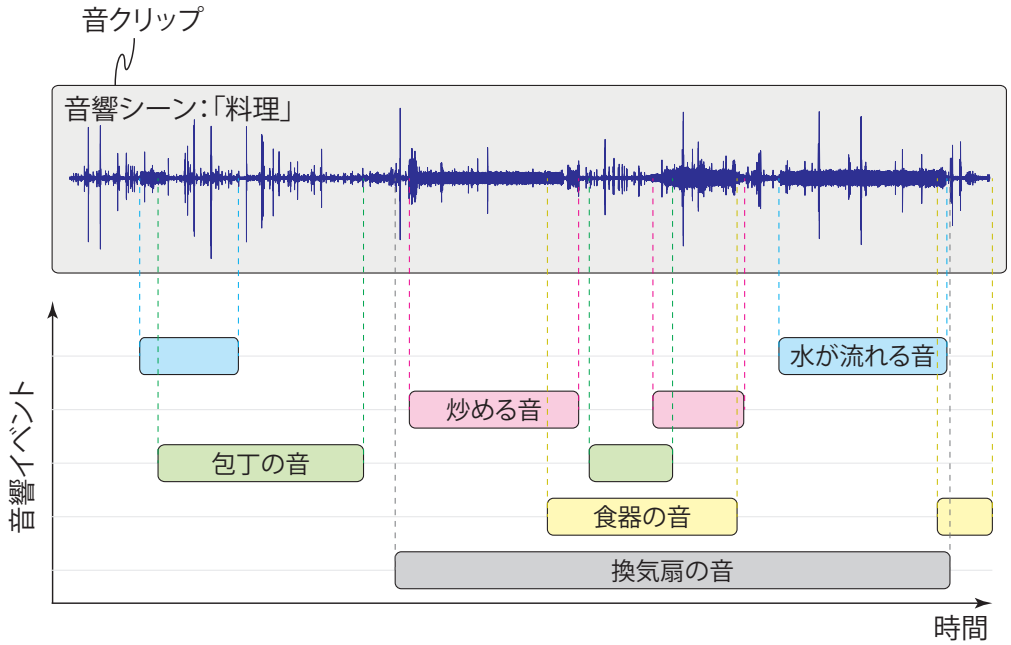



図 2.1: 本論文で用いる用語の定義

所や状況を分析する研究は音響イベント検出 [1, 2, 3, 4, 5, 6]や音響シーン分析 [7, 8, 9]などと呼ばれており，これまでにも複数の手法が提案されている。しかしながら，これらの研究分野は比較的新しいため，用語や問題設定は十分に整理されているとは言い難い。例えば近年では，大石が音響イベント検出や音響シーン分析の問題設定について概説している[23]。また，DCASE2016においてPlumbley [24]やRichard [25]らが音響イベント検出や音響シーン分析の用語や問題設定について概説しているが，これらの議論が統一的な見解となるにはまだ時間を要すると考えられる。そこで以下では，音響イベント検出や音響シーン分析で用いられる用語の定義と問題設定について，DCASE2016におけるPlumbley [24]やRichard [25]らの解説と図を参考に，あらためて整理を行う。

図2.1に本論文で用いる用語の概念図を示す。本論文では比較的長時間の収録音（数秒〜数十秒）の単位を音クリップ（Sound clip）と呼び，一つの音クリップは同一の収録環境下で収録されるものとする。各音クリップには音が収録された場所（e.g. 電車，車，公園，屋内）や状況（e.g. 会議中，非日常），周囲の人の行動（e.g. 料理，掃除，会話）を示すラベルが音クリップ単位で付与されており，これを音響シーン（Acoustic scene）と呼ぶ。つまり，それぞれ

表 2.1: 音響イベント検出/音響シーン分析の問題設定とその呼称

| | | 分析結果の出力 | |
|---|---|---|---|
| | | 単一のラベル | マルチラベル + タイムスタンプ |
| 分析対象 | 音響イベント | 音響イベント分類 | 音響イベント検出 |
| | 音響シーン | 音響シーン分類 | 音響シーン検出 |

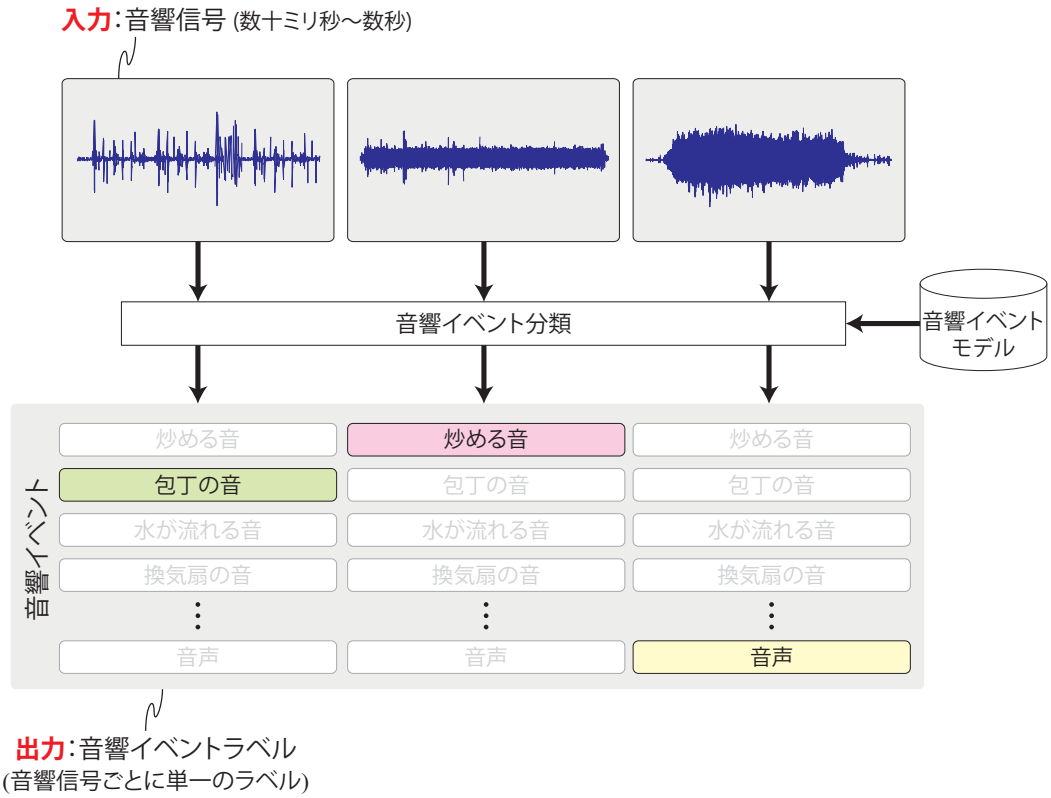



図 2.2: 音響イベント分類の問題設定

の音クリップには一つの音響シーンが重複なく含まれているものとする。また，各音クリップには，様々な時間間隔を持つ様々な種類の音（e.g. 水が流れる音，足音，換気扇の音，音声，音楽）が含まれており，これらの発生音の種類のことを本論文では音響イベント（Acoustic event）と呼ぶこととする。

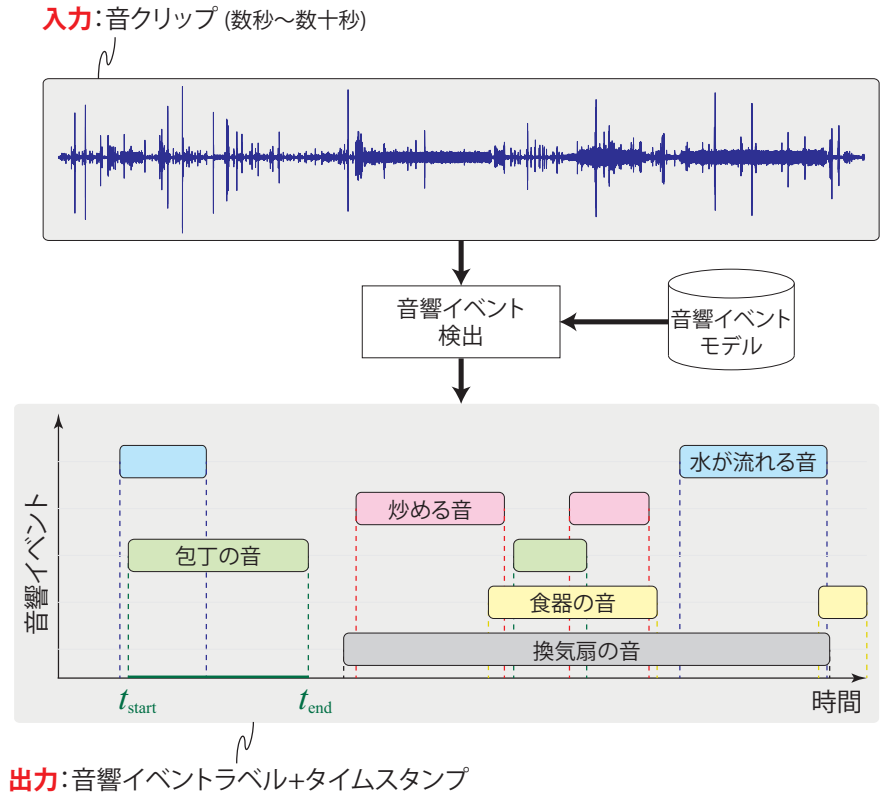



図 2.3: 音響イベント検出の問題設定

## 2.3 関連研究の問題設定の整理

次に音響イベント検出や音響シーン分析における問題設定の例と，その適用例ついて述べる。本論文では，音響イベント検出や音響シーン分析において頻繁に取り扱われる問題を，分析対象と分析結果の出力という観点で表2.1，図2.2，図2.3，図2.4のように整理する。まず，最もよく取り扱われる問題として，図2.2に示すように，音響イベントを分析し単一の音響イベントラベルを出力とする音響イベント分類が挙げられる。音響イベント分類では，単一の音響イベントを含む比較的短時間（数十ミリ秒〜数秒程度）の音響信号を入力として，音響信号を最もよく表す音響イベントのラベルを出力する。また，図2.3に示すように，数秒〜数十秒程度の比較的長時間の音クリップを入力として，音クリップ中に発生した音響イベントのラベルとイベントの発生区間を出力とする音響イベント検出に関する研究も行われている。音響イベント

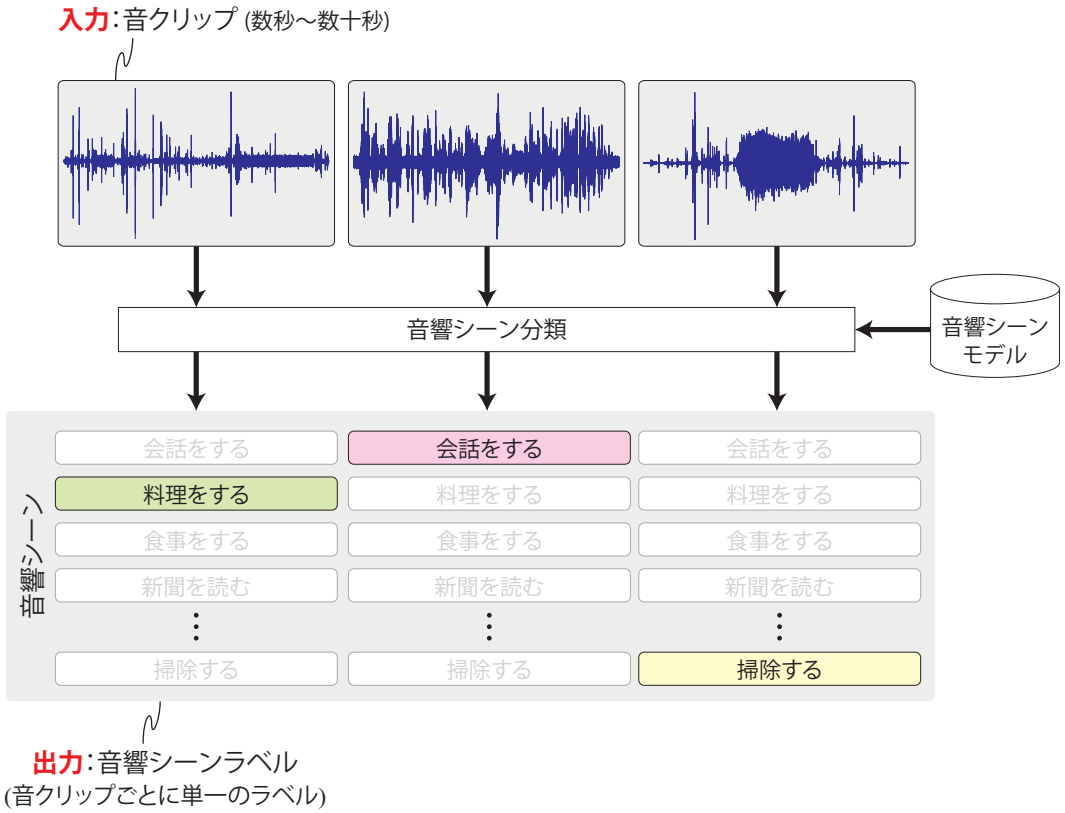



図 2.4: 音響シーン分類の問題設定

分類や音響イベント検出は動画などのメディアの検索[11, 6]や自動監視システムにおける異常音の検出[12, 13, 15]に重要な技術として盛んに取り組まれている。

次に音響シーンを扱う問題として最もよく取り扱われる，音響シーン分類の問題設定を図2.4に示す。音響シーン分類では，音クリップを入力とし，音クリップを収録した場所や状況，周囲の人の行動などの音響シーンラベルを出力とする。出力される音響シーンラベルはあらかじめ定められたカテゴリを扱う場合が主であるが，本論文では，特定のカテゴリラベルが付与されておらず，定められたカテゴリ数に音響シーンを分類する問題も音響シーン分類と呼ぶこととする。音響シーン分類は見守りや自動ライフログシステム[17, 8]，動画の自動分類システム[10]など幅広い用途に適用することが可能な技術として注目を集めている。本論文では音響シーン分類を対象として，様々な観測に対して精度良く音響シーンを分類可能な手法について検討を行

う。

## 2.4 関連研究

### 2.4.1 音響シーン分類の従来研究

音響シーンの分類手法としてはこれまでにも多数の手法が提案されている。例えばEronenらは，音声認識で広く用いられているメル周波数ケプストラム係数（MFCCs: Mel-frequency cepstral coefficients）と隠れマルコフモデル（HMM: Hidden Markov model）を利用して音響シーンを分類する手法を提案している[7]。また，Geigerらは，短時間のフレーム毎にMFCCsやMPEG-7 standard feature [26]と線形カーネルを用いたサポートベクターマシーン（SVM: Support vector machine） [27, 28]を適用して各フレーム毎に音響シーンを分類し，Majority voteにより音クリップ単位での音響シーンの出力結果を決定する手法を提案している[29]。これらの手法では，音クリップを短時間のフレーム毎に分割して音響特徴を抽出した後，各フレーム毎に音響シーン分類を行い，それらを結果統合することで最終的な音響シーンの出力結果を決定している。同様の手法はChumら[30]によっても提案されている。一方，Bisotらは音クリップから算出されたスペクトログラムに対し，スパース制約付き非不値行列因子分解（Sparse NMF: Sparse non-negative matrix factorization）や畳み込みNMF（Convolutive NMF）を用いて教師なし学習により音響シーンの特徴量を学習して音響シーン分類に利用する手法を提案している[31]。

近年では深層学習による音響シーン分類手法も多数提案されている。例えばXuらは，メルフィルタバンク特徴量を入力として，屋内，屋外などのより広範な意味を持つ音響シーンと，公園，バス，台所などより詳細な音響シーンの二段階で深層ニューラルネットワーク（DNN: Deep neural network）を学習し，音響シーンを分類する手法を提案している[32]。 Baeら[33]やValenti[34]らは，スペクトルやケプストラムを入力として畳み込みニューラルネットワーク（CNN: Conventional neural network）により音響シーンをモデル化する手法を提案している。その他にも，リカレントニューラルネットワーク（RNN: Recurrent neural network）と深層型線形判別分析（Deep LDA: Deep linear discriminant analysis）を組み合わせた手法[35]や深層型パーセプトロン（Deep multilayer perceptron neural network）に基づく音響シーン分類手法[36]なども提

案されている。

音響シーンは「水が流れる音」という情報のみでは一意に特定することは困難であり，多くの音響シーンは複数の音響イベントの組み合わせ情報により特徴付けられる点に着目した手法も提案されている。例えば，Guoら[37]やHeittolaら[38]，Elizaldeら[39]は「料理」という音響シーンが「水が流れる音」「包丁の音」「フライパンを熱する音」など複数の音響イベントの情報を組合せることで効果的に分類可能であることに着目し，音クリップに含まれる音響イベントのヒストグラムを特徴量としてSVMにより分類する方法を提案している。また，Leeらは音響シーンと音響イベントヒストグラムの関係を表す確率モデルのパラメータを最大事後確率推定（MAP estimation: Maximum a posteriori estimation）により推定することで音響シーン分類を行う方法を提案している [40]。しかしながら，音響シーンや音響イベントの種類，また音響シーンと音響イベントの組み合わせは非常に多岐にわたるため，音響イベントの組み合わせにより直接音響シーンを特徴付ける手法では，大規模な学習データを用意しなければモデルが学習データに過剰適合してしまうという課題がある。

一方，音響シーン分析の研究に利用可能なデータセットとして，DCASE2016 Dataset [41]やCHiME-Home Dataset [42]，DIRHA Simulated Corpus [43]などがあるが，これらのデータセットは十分な量のデータが含まれているとは言えない。今後，より大規模なデータセットが整備されていくと考えられるが，あらゆる音響シーンや音響ワードを十分に含むデータを収録することは困難であるため，限られたデータから音響シーンを分析する手法を実現することは重要である。

### 2.4.2 音響トピックモデル

音響シーンを音響イベントの組み合わせにより特徴付ける際，特に音響シーンや音響イベントの種類が多い場合は大規模な学習データを用意しなければモデルが学習データに過剰適合してしまう。他方，音響シーンと音響イベントの関係に注目すると，ある音響シーン関連した音クリップには限られた種類の音響イベントのみ含まれている場合が多く，音響イベントのヒストグラムは疎になっている。この音響イベントのスパース性に着目することで，限られたデータセットからでも頑健に音響シーンの分類が可能な手法が提案さ

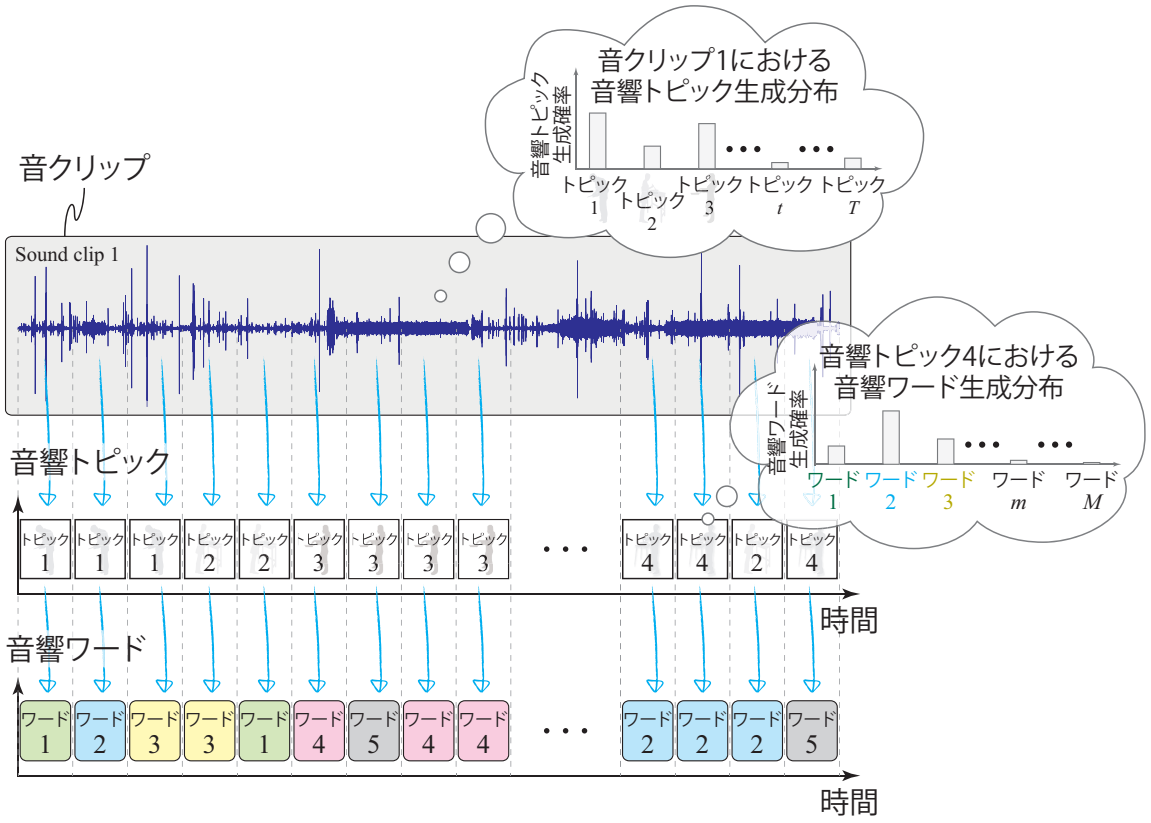



図 2.5: 音クリップから音響ワードの系列が生成される過程

れており[44, 8, 45, 46, 47]，音響トピックモデル（ATM: Acoustic topic model）と呼ばれている。

音響トピックモデルでは，図2.5, 2.6に示すように，音クリップから音響ワードの系列が生成される過程をベイズモデルの枠組みによりモデル化し，音響ワード系列が持つ潜在的な構造である音響トピックを利用して音響シーンの分類を行う。ここで，音響トピックモデルでは発生区間がばらばらで重複のある音響イベントの代わりに，一定の時間区間で重複を持たない音響ワード（ラベル）の生成過程を考える。つまり，音響トピックモデルでは，音響イベントの組み合わせの代わりに，(音響イベントの種類数$M$)>(音響トピックの種類数$T$)となるような音響トピックの組み合わせにより音響シーンを特徴づけることにより，従来法で課題となる過剰適合の問題を回避し，限られたデータセットからでも頑健に音響シーンをモデル化/分類することを可能として

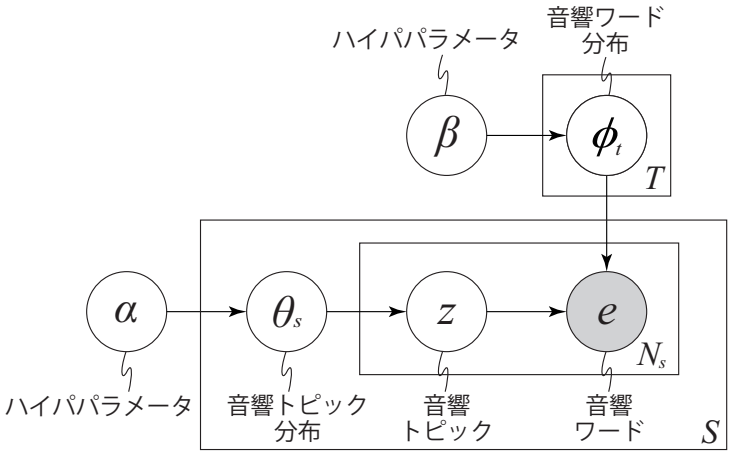



図 2.6: 音響トピックモデルのグラフィカルモデル

いる。以降では，音響トピックモデルによる音響ワード系列の生成過程のモデル化とモデルパラメータ推定，音響トピックモデルを用いた音響シーン分類方法について述べる。

### 音響ワード系列の生成過程のモデル化

図2.5, 2.6に示すように，音響トピックモデルでは音クリップから音響ワードの系列が生成される過程を，音響トピックと呼ばれる潜在変数と音響ワードの階層的な生成過程によりモデル化する。なお，音響トピックモデルという呼称は，自然言語処理分野において文書から単語が生成される過程をモデル化して文書中の話題（トピック）を分析するトピックモデルに準えてこのように命名されている。以下では，音響トピックモデルにおいて音クリップから音響ワード系列が生成される過程の詳細と，その定式化について述べる。なお，音響ワード系列は，音響信号や音響特徴量が混合ガウスモデル（GMM: Gaussian mixuture model）やSVM [27, 28]などの手法により，時間フレーム毎に離散化されたラベル系列として表現する。また，音響トピックとは音響シーンと音響ワードを関係付ける潜在的な構造であり，異なるトピック毎に異なる音響ワードの生成分布を有しているものとする。

音響トピックモデルの生成過程によると，各音クリップは音響トピックの生成を決める分布を有しており，その分布に従って音響トピックが生成される。また，各音響トピックも音響ワードの生成に関する分布を有しており，

表 2.2: 音響トピックモデルで用いる変数の定義

| Symbol | Definition |
|---|---|
| $S$ | 音クリップ（音響ワード系列）の総数 |
| $A$ | 音響シーンのクラス数 |
| $T$ | 音響トピックのクラス数 |
| $M$ | 音響ワードのクラス数 |
| $N_{\boldsymbol{e}_s}$ | 音響ワード系列$\boldsymbol{e}_s$に含まれる音響ワードの数 |
| $s$ | 音クリップのインデックス |
| $a$ | 音響シーンのクラスインデックス |
| $t$ | 音響トピックのクラスインデックス |
| $m$ | 音響ワードのクラスインデックス |
| $i$ | 音クリップ中の音響ワードのインデックス |
| $\mathcal{S}$ | 音響ワード系列の集合 |
| $\mathbf{a}_s$ | 音クリップ$s$の音響シーンの候補 |
| $a_s$ | 音クリップ$s$の音響シーン |
| $\boldsymbol{z}$ | 音響トピック表す潜在変数 |
| $\boldsymbol{e}$ | 音響ワードを表す変数 |
| $z_{s,i}, e_{s,i}$ | 音クリップ$s$の$i$番目の音響トピックおよび音響ワード |
| $\boldsymbol{\theta}_s, \boldsymbol{\theta}_a$ | 音クリップ$s$または音響シーン$a$における音響トピック生成分布のパラメータ |
| $\theta_{a,t}$ | 音響シーン$a$における音響トピック$t$の生成確率 |
| $\boldsymbol{\phi}_t$ | 音響トピック$t$における音響ワード出現分布のパラメータ |
| $\phi_{t,m}$ | 音響トピック$t$における音響ワード$m$の生成確率 |
| $\alpha, \beta$ | Dirichlet分布の超パラメータ |
| Dir$(\cdot)$ | Dirichlet分布 |
| Categ$(\cdot)$ | Categorical分布 |
| Uni$(\cdot)$ | Uniform分布 |
| $\Gamma(\cdot)$ | Gamma関数 |
| KL$(\cdot,\cdot)$ | KLダイバージェンス |
| $n_t^s$ | 音クリップ$s$において音響トピック$t$に割り当てられた音響ワードの数 |
| $n_m^t$ | 音響トピック$t$において音響ワード$m$に割り当てられた音響ワードの数 |

**Algorithm 2.1** 音響トピックモデルにおける音響ワード系列の生成過程

**for** $t = 1$ to $T$ **do**
  Choose $\boldsymbol{\phi}_t$ $\sim$ Dirichlet($\beta$)
**end for**
**for** $s = 1$ to $S$ **do**
  Choose $\boldsymbol{\theta}_s$ $\sim$ Dirichlet($\alpha$)
  **for** $i = 1$ to $N_{\boldsymbol{e}_s}$ **do**
    Choose $z_{s,i} \mid \boldsymbol{\theta}_s$ $\sim$ Categorical($\boldsymbol{\theta}_s$)
    Choose $e_{s,i} \mid \boldsymbol{\phi}_{z_{s,i}}, z_{s,i}$ $\sim$ Categorical($\boldsymbol{\phi}_{z_{s,i}}$)
  **end for**
**end for**

音響トピックが決まるとこの音響ワード分布に基づいて音響ワードが決定される。この生成過程は各時間フレーム毎に繰り返され，最終的に音響ワードの系列が生成される。ここで，音響トピックモデルでは音響ワードの時間的な位置関係は考慮せずに交換可能とする "Bag of acoustic word" 表現 [48]を用いて音響ワード系列の生成過程をモデル化する。これは，例えば「料理」という音響シーンにおいて「包丁の音」と「水が流れる音」という音響イベントの時間関係が入れ替わり得るように，音響ワードの出現頻度のみに着目することで音響シーンの特徴を捉え易くするためである。上記の階層的な生成過程を基にそれぞれの生成確率に対して事前分布を導入して，音響トピックモデルの生成過程はAlgorithm 2.1のように表現される。なお，各変数の定義は表2.2 に示す通りである。 Algorithm 2.1では，音響トピックの生成分布のパラメータ$\boldsymbol{\theta}_s$および音響ワードの生成分布のパラメータ$\boldsymbol{\phi}_{z_{s,i}}$がDirichlet事前分布を持つと仮定することで，各音クリップに含まれる音響トピック数や各音響トピックに含まれる音響ワード数のスパースさを制御することができる。

音響ワードの時間的な位置関係は考慮しないことを踏まえると，音響ワード系列の生成確率は以下で表現される。

$$
\begin{aligned}
& p(\boldsymbol{e}_S|\alpha,\beta) \\
&\quad = \sum_{\boldsymbol{z}} p(\boldsymbol{e}_S|\boldsymbol{z},\alpha,\beta)p(\boldsymbol{z}|\alpha,\beta) \\
&\quad = \prod_{s=1}^{S} \int \mathrm{Dir}(\boldsymbol{\theta}_s|\alpha) \left\{ \prod_{i=1}^{N_{\boldsymbol{e}_s}} \sum_{z_{s,i}} \mathrm{Categ}(z_{s,i}|\boldsymbol{\theta}_s) \int \mathrm{Dir}(\boldsymbol{\phi}_t|\beta)\mathrm{Categ}(e_{s,i}|\boldsymbol{\phi}_t, z_{s,i}) d\boldsymbol{\phi}_t \right\} d\boldsymbol{\theta}_s
\end{aligned}
\tag{2.1}
$$

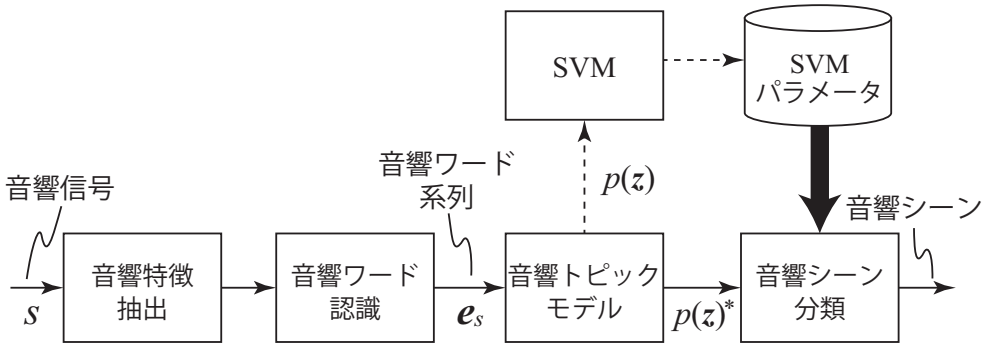



図 2.7: 音響トピックモデルを利用した音響シーン分類システムの構成例

但し，$\theta_{st}, \phi_{tm}$は各時刻において音クリップ$s$から音響トピック$t$が生成される確率と，音響トピック$t$から音響ワード$m$が生成される確率をそれぞれ表す。このとき，各音響トピックや音響ワード毎の出現回数$n_t^s, n_m^t$だけ，生成確率$\theta_{st}, \phi_{tm}$を掛け合わせた形でCategorical分布を表現すると，音響ワード系列の生成確率は具体的に以下のように表現可能である。

$$p(\boldsymbol{e}_S|\alpha,\beta) = \prod_{s=1}^{S}\int \frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\prod_{t=1}^{T}\theta_{st}^{\alpha-1}\left\{\sum_{z_{s,i}}\theta_{st}^{n_t^s}\int\frac{\Gamma(M\beta)}{\Gamma(\beta)^M}\prod_{m=1}^{M}\phi_{tm}^{\beta-1+n_m^t}d\boldsymbol{\phi}_t\right\}d\boldsymbol{\theta}_s \quad (2.2)$$

なお，上記の音響ワード系列の生成確率では，Dirichlet事前分布として全ての超パラメータが同一の値となる対称Dirichlet分布を用いている。また，音響トピックモデルのパラメータ推定方法として，変分ベイズ法 (VB: Variational Bayes) [49, 50]に基づく手法を付録のA.1に示す。

### 音響シーンの分類

音響トピックモデルでは，音響ワードの組み合わせ情報を用いて音響シーンの分類を行う代わりに，音響トピックの生成分布を用いて音響シーンの分類を行う。例えば，Kim [44]らは音響シーンを分類するため，図2.7に示すように，音響トピックモデルを用いて音響ワード系列から音響トピックの生成分布$p(z)$を学習した後，各音響ワード系列に対応する音響シーンラベルと音響トピックの生成分布$p(z)$を入力とした多クラスSVMを用いて音響シーン分類を行っている。

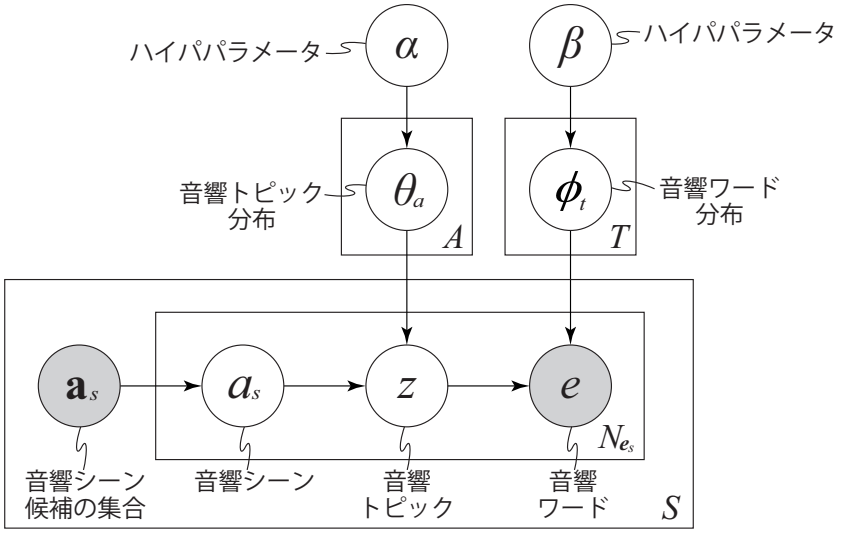



図 2.8: 教師あり音響トピックモデルのグラフィカルモデル

### 2.4.3 教師あり音響トピックモデル

#### 音響ワード系列の生成過程のモデル化

音響トピックモデルでは，音響ワード系列が音響トピックと呼ばれる潜在的な構造を持ち，音響トピックにより音響シーンが特徴付けられると仮定している。一方で，音響シーンと音響ワード系列の関係を明示的にモデル化し音響シーン分類を行う手法として，教師あり音響トピックモデル [8, 47]が提案されている。文献 [8, 47]では，音響シーンの教師ラベルを用いた音響トピックモデルを音響シーン-サブトピックモデル (ASTM: Acoustic scene and sub-topic model)と称しているが，本稿ではこのモデルを教師あり音響トピックモデル (sATM: Supervised acoustic topic model)と称する。教師あり音響トピックモデルでは，音響トピックの生成分布により音響シーンを特徴づける点は音響トピックモデルと同じであるが，音響シーンラベルを用いて明示的に音響シーンと音響ワード系列の関係をモデル化するため，音響シーンの分類問題により適したアプローチとなっている。

教師あり音響トピックモデルでは，各音響ワード系列に対し（複数の）音響シーンの候補をラベルとして与え，その内のいずれかの音響シーンが音響ワード系列で1つ現れると仮定する。また，各音響シーンはそれぞれ異なる音響トピックの生成分布を持ち，時間フレーム毎に異なる音響トピックが生成

**Algorithm 2.2** 教師あり音響トピックモデルにおける音響ワード系列の生成過程

**A set of possible acoustic scenes** $\mathrm{a}_s$ **is given,**
**for** $a = 1$ **to** $A$ **do**
  **Choose** $\boldsymbol{\theta}_a$ $\sim$ Dirichlet($\alpha$)
**end for**
**for** $t = 1$ **to** $T$ **do**
  **Choose** $\boldsymbol{\phi}_t$ $\sim$ Dirichlet($\beta$)
**end for**
**for** $s = 1$ **to** $S$ **do**
  **Choose** $a_s$ $\sim$ Uniform($\mathbf{a}_s$)
  **for** $i = 1$ **to** $N_{\boldsymbol{e}_s}$ **do**
    **Choose** $z_{s,i} \mid \boldsymbol{\theta}_{a_s}, a_s$ $\sim$ Categorical($\boldsymbol{\theta}_{a_s}$)
    **Choose** $e_{s,i} \mid \boldsymbol{\phi}_{z_{s,i}}, z_{s,i}$ $\sim$ Categorical($\boldsymbol{\phi}_{z_{s,i}}$)
  **end for**
**end for**

されるものとする。また，その他の生成過程は音響トピックモデルと同様である。具体的には，教師あり音響トピックモデルの生成過程はAlgorithm 2.2のように表すことが可能である。

教師あり音響トピックモデルにおける音響ワード系列の生成確率は以下で表現される。

$$p(\boldsymbol{e}_{\mathcal{S}}|\alpha,\beta,\boldsymbol{a}_s)=\prod_{s=1}^{S}\frac{1}{A_s}\sum_{a\in\boldsymbol{a}_s}\int\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\prod_{t=1}^{T}\theta_{at}^{\alpha-1}\left\{\sum_{z_{s,i}}\theta_{at}^{n_t^a}\int\frac{\Gamma(M\beta)}{\Gamma(\beta)^M}\prod_{m=1}^{M}\phi_{tm}^{\beta-1+n_m^t}d\boldsymbol{\phi}_t\right\}d\boldsymbol{\theta}_a \tag{2.3}$$

また，教師あり音響トピックモデルのパラメータ推定方法について，変分ベイズ法 (VB: Variational Bayes) [49, 50]に基づく手法を付録のA.2に示す。

### 音響シーンの分類

教師あり音響トピックモデルでは，音響トピックモデル同様，音響ワードの組み合わせ情報を用いて音響シーンの分類を行う代わりに音響トピックの生成分布を用いて音響シーンの分類を行う。具体的には図2.7に示すように，ま

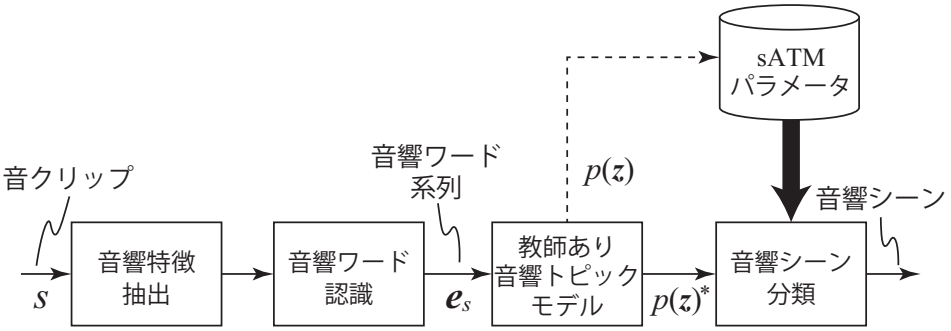



図 2.9: 教師あり音響トピックモデルを利用した音響シーン分類システムの構成

ず学習データから音響トピック分布$\boldsymbol{\theta}_a$および音響ワード分布$\boldsymbol{\phi}_t$を学習する。その後，音響シーンを分類したい音響ワード系列に対して，以下の事後確率が最も高くなる音響シーンを推定することにより音響シーン分類が実現できる。

$$\arg\max_{a} p(a|\boldsymbol{\theta}_a, \boldsymbol{\phi}_t, \boldsymbol{e}_s, \alpha, \beta) \tag{2.4}$$

## 2.5 本研究の方針

本論文では，1.2節で述べた観測のうち，従来研究で扱われていない，1) 未知の音響シーンを含む音情報が逐次的に得られる場合における音響シーン分類，2) 観測の一部が欠損している場合における音響シーン分類，3) 正確に時間同期されておらず，マイクロホン位置も不明な多チャネル観測に基づく音響シーン分類手法を提案することを目標とする。

提案手法では，音響シーン分類に共通する課題である，大規模な学習データを用意しなければモデルが学習データに過剰適合してしまうという問題に対処するため，音響トピックモデルを拡張したシーン分類手法を提案する。また，音響トピックモデルでは音響シーンや音響ワードに関連する情報を，確率変数という抽象的なパラメータで表現する。これは例えば周波数情報と空間情報や，観測された情報と欠損した情報といった異なる性質を持つ情報を，同一の枠組みで扱うことが可能となる。このように，観測条件の観点からも音響トピックモデルは本研究の課題解決に対して有効なアプローチであ

ることを示している。

# 3 逐次的な音情報の観測に基づく音響シーン分類

## 3.1 はじめに

投稿型動画サイトのコンテンツや日々の活動を記録し続けるライフログのように音情報が逐次的に得られる場合に，音響シーンをモデル化し分類する問題を考える。従来手法を用いて音響シーンのモデル化や分類を行うには，新たなデータが得られる度にモデルを再学習する必要があり，計算コストが大きな問題となる。さらに，データ量に比例してモデル学習に必要となるメモリ量が増加するため，投稿型動画サイトのコンテンツのように膨大な量のデータに従来法を適用することは難しい。

事前に全ての学習データが得られない場合における音響イベント分類手法は従来にも複数提案されている。例えばLecomteらは，比較的学習データが得やすい通常時の音を事前に収集し，等分割したフィルタバンクのエネルギーとOne-class SVMを用いて通常時の音を学習しておき，学習した通常音に当て

はまらない音響イベントを未知の音響イベントとして検出する手法を提案している[14]。また，MarchiらはBidirectional long short-term memory-recurrent neural networks (BLSTM-RNN)を用いたDenoising autoencoderによる未知の音響イベント分類手法を提案している[18]。文献[18]では，すでに多数のサンプルが得られてる音響イベントはDenoising autoencoderにより正確にモデル化できるが，未知の音響イベントはサンプル数が多く得られていないためモデル化を行う際に誤差を多く含むことに着目して，未知の音響イベント検出する手法を提案している。

しかしながら従来の手法では，音響イベントや音響シーンを逐次的にモデル化するには，依然として新たなデータが得られる度に全てのデータを用いてモデルを再学習する必要がある。この問題を解決するため，本章では逐次的に得られたデータのみを用いて適応的にモデル学習する手法について検討する。ここで，投稿型動画サイトのコンテンツなどでは多くの場合，音クリップ（音響ワード列）単位の観測が逐次的に得られることから，本論文では図3.1に示すように，逐次的に音クリップ単位の観測が得られた場合に音響シーンを分析する問題を考える。また，投稿型動画サイトのコンテンツや日々の活動を記録し続けるライフログのような逐次観測を想定した場合，音情報の観測のみが得られ，音響シーンの種類（ラベル）が同時に得られることは稀であることから，本論文では教師なし学習により逐次的にモデル学習する問題を取り扱う。

過去に得られたデータを保存せず，逐次的に得られたデータからモデルを更新するためには，音響シーンに関する情報をパラメータとしてモデルに保存しておき，逐次的にこのパラメータを更新する方法が考えられる。機械学習を用いた音響シーン分類の多くは音響シーンに関する情報をパラメータとしてモデルに保存しており，音響シーンモデルを音響トピックの生成分布や音響ワードの生成分布のパラメータで表現する音響トピックモデルもその一つである。また，逐次的に得られたデータから適応的に音響シーンモデルを更新する際の別の課題として，学習初期には限られたデータを用いて音響シーンをモデル化しなければならず，過学習に陥りやすいという問題がある。音響トピックモデルでは，音のスパース性をモデルに導入できるため，限られたデータからでも過学習することなく音響シーンをモデル化可能であるという点から，本論文では，音響トピックモデルを逐次的に学習可能な手法に拡張する。

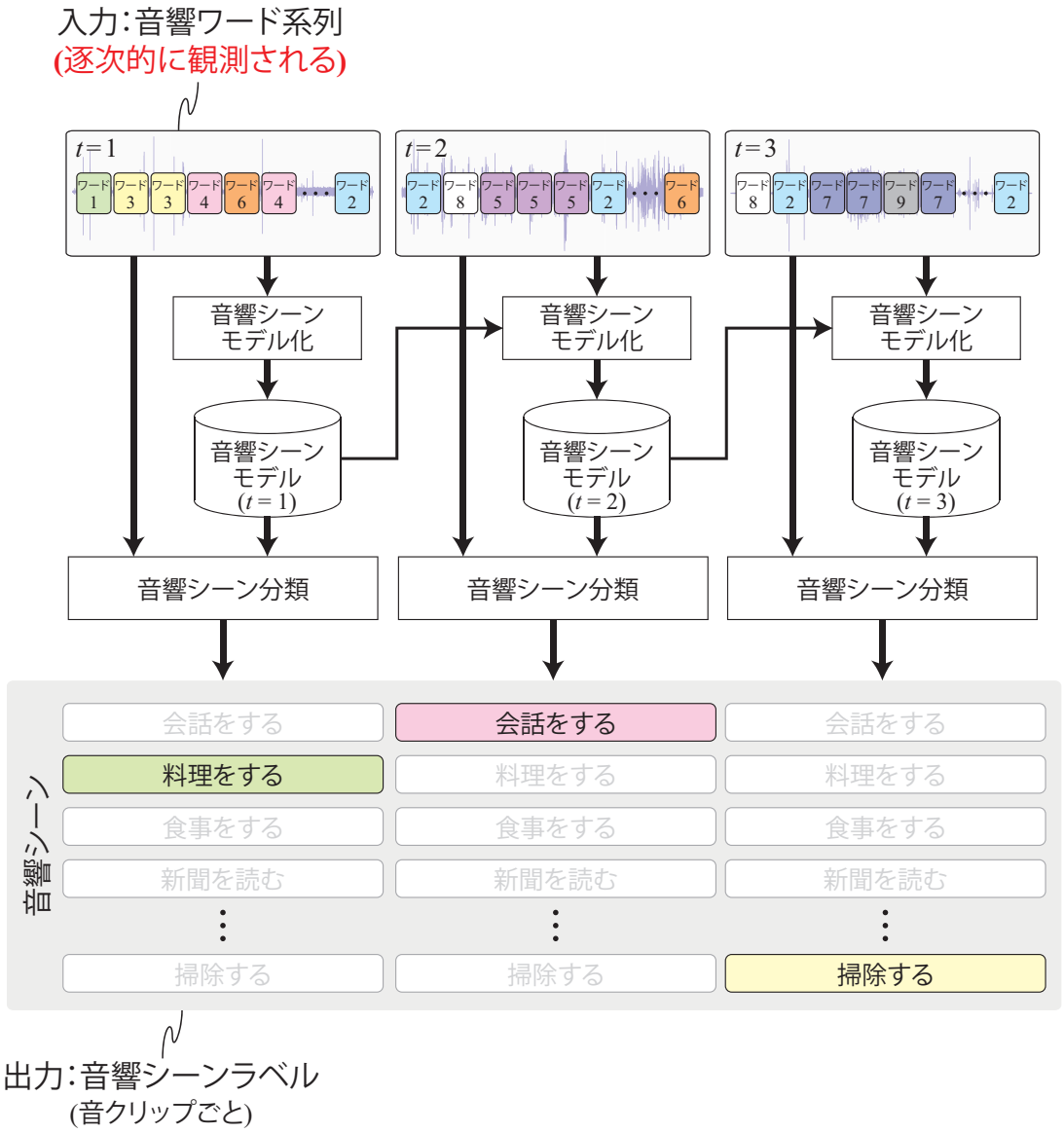



図 3.1: 3章で扱う音響シーン分類問題

加えて本章では，逐次学習を行う際に重要となる，高速な計算時間と限られたデータから頑健にモデル学習することを両立させるため，計算コストとパラメータ推定精度それぞれに利点を持つ崩壊型変分ベイズ法 (CVB: Collapsed variational Bayes) [50, 49]と崩壊型ギブスサンプリング (CGS: Collapsed Gibbs sampling) [51, 52]を組合せたハイブリッド型パラメータ推定手法も音響トピックモデルに導入する。

本章の構成は以下のとおりである。 3.2節では，音響トピックモデルにおけるパラメータ推定手法とその課題について述べる。 3.3節では，CVB法とCGS法のハイブリッド型手法を適用する場合におけるデータセットの分割方法について述べ，その後，CVB法とCGS法それぞれのパラメータ推定手法について述べる。 3.4節では，音響トピックモデルのハイブリッド型パラメータ推定手法のオンライン化について述べる。 3.5節において，実環境収録音を用いた提案手法の性能評価実験について議論し，3.6節で本章のまとめを述べる。

## 3.2 音響トピックモデルのバッチ型パラメータ推定の従来法

音響トピックモデルのパラメータ$\boldsymbol{\theta}_s, \boldsymbol{\phi}_t, \boldsymbol{z}$の推定には変分ベイズ (VB: Variational Bayes) [50, 49]や崩壊型ギブスサンプリング (CGS: Collapsed Gibbs sampling) [51, 52]，期待値伝播 (EP: Expectation propagation) [53]などの手法が利用可能である。 VB法に基づく手法では比較的少ない演算量でパラメータの推定が可能であり，また，オンライン学習にも適用しやすい利点がある。しかしながら，各パラメータに対して平均場近似と呼ばれる独立性を仮定するため推定結果が局所解に陥りやすく，CGS法と比較してしばしば推定精度が低下する。CGS法では，事後確率に対する潜在変数のサンプリングを無限回繰り返すと理論的に大域的最適解を推定することが可能であるが，精度の良い推定を実現するためには多数回のサンプリングを行う必要があり，計算コストが非常に大きくなるという問題がある。

これらの問題に対処するためHoffman等は，VB法における平均場近似の独立性を緩和することでパラメータの推定結果が局所解に陥りにくくする手法[54]を提案している。またPaisley等は，推定したいパラメータに対する周辺対数尤度の下限値を，確率的探索アルゴリズムにより直接推定する手法[55]を

提案している。しかしながら，これらの手法はVB法と比較して計算コストが大きくなるという課題がある。一方でWelling等は，パラメータの推定に用いる音響ワード系列を，局所解に収束しやすいパラメータに関係するデータセット$\mathcal{S}^{GS}$とその他のデータセット$\mathcal{S}^{VB}$に分割し，$\mathcal{S}^{GS}$に関係するパラメータの推定にはCGS法を適用し，$\mathcal{S}^{VB}$に関係するパラメータにはCVB法を適用するハイブリッド手法[56]を提案している。 CGSとCVB法のハイブリッド手法では，CGSの利点である大域的な最適解に近いパラメータの推定を可能にしつつも，CVB法の利点である計算コストの大幅な低減が可能となる。しかしながら従来のハイブリッド法では，CVB法を適用する際にパラメータの二次統計量を推定する必要があるため，依然として計算コストの大きさが問題となる。そこで本論文では，より計算コストを低減しながらも，CVB法と同程度の精度でパラメータ推定が可能とされている，崩壊型変分ベイズの0次近似に基づく手法(CVB0：Collapsed variational Bayes with 0$th$-order approximation) [57]とCGS法のハイブリッド手法を用いる。

以降ではまず，少ない計算コストで精度良くパラメータ推定が可能な，CVB0法とCGS法のハイブリッド型推定手法について述べ，その後，オンライン推定手法への拡張を行う。

## 3.3 音響トピックモデルのハイブリッド型パラメータ推定手法

### 3.3.1 ハイブリッド型パラメータ推定のためのデータ分割

音響トピックモデルのパラメータは，音クリップに含まれる音響ワードの種類毎のサンプル数$n_{sm}$に依存しており，多数のサンプルが得られている音響ワードに関連するパラメータは大域的な最適解を得ることが比較的容易である一方で，小数のサンプルのみ得られている音響ワードに関連するパラメータは局所解に陥りやすいことが知られている[56]。このことを踏まえ，本論文では多数のサンプルが得られている音響ワードに関連するパラメータの推定においては，計算コストにおいて利点の大きいCVB0法を導入し，小数のサンプルのみ得られている音響ワードに関連するパラメータの推定にはCGS法を導入することで，音クリップ$s$全体として効率的にパラメータ推定を行う。

具体的には，ハイブリッド型パラメータ推定のためのデータセットの分割は以下のように行う。なお，本章で用いる変数の定義を表3.1に示す。

$$\mathcal{S}^{VB} = \{s, i | n_{se_{s,i}} > d\},\ \mathcal{S}^{GS} = \{s, i | n_{se_{s,i}} \leq d\} \tag{3.1}$$

但し，$d$および$n_{se_{s,i}}$はそれぞれ，$0 \leq d \leq N_{\boldsymbol{e}_s}$を満たす任意の整数値，音クリップ$s$に含まれる音響ワード$e_{s,i}$の数とする。以降では，CVB0法とCGS法によるパラメータ推定問題の定式化と，繰り返し最適化によるパラメータ更新式の導出を行う。

### 3.3.2 CVB0法によるモデルパラメータ推定

CVB0法 [57]では，推定したい潜在変数や生成分布をパラメータとして持つ変分事後分布$q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi})$を定義し，Jensenの不等式と平均場近似を用いて，変分事後分布を繰り返し真の事後分布$p(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{e})$に近づけることによりパラメータの推定を行う。まず，Jensenの不等式より，全ての未知量に対する周辺対数尤度の下限値$\mathcal{F}[q]$を計算すると以下のようになる。

$$\begin{aligned}
\mathcal{L}(\boldsymbol{e}) &\triangleq \log p(\boldsymbol{e}|\alpha, \beta) \\
&= \iint \sum_{\boldsymbol{z}} \log p(\boldsymbol{e}, \boldsymbol{z}, \boldsymbol{\phi}, \boldsymbol{\theta}|\alpha, \beta) d\boldsymbol{\phi} d\boldsymbol{\theta} \\
&\geq \iint \sum_{\boldsymbol{z}} q(\boldsymbol{z}, \boldsymbol{\phi}, \boldsymbol{\theta}) \log \frac{p(\boldsymbol{e}, \boldsymbol{z}, \boldsymbol{\phi}, \boldsymbol{\theta}|\alpha, \beta)}{q(\boldsymbol{z}, \boldsymbol{\phi}, \boldsymbol{\theta})} d\boldsymbol{\phi} d\boldsymbol{\theta} \\
&\triangleq \mathcal{F}[q]
\end{aligned} \tag{3.2}$$

また，$\mathcal{L}(\boldsymbol{e})$と$\mathcal{F}[q]$はKLダイバージェンスを用いて以下のように表現する事も可能である。

$$\begin{aligned}
\mathcal{L}(\boldsymbol{e}) - \mathcal{F}[q] &= \iint \sum_{\boldsymbol{z}} q(\boldsymbol{z}, \boldsymbol{\phi}, \boldsymbol{\theta}) \log \frac{q(\boldsymbol{z}, \boldsymbol{\phi}, \boldsymbol{\theta})}{p(\boldsymbol{z}, \boldsymbol{\phi}, \boldsymbol{\theta}|\boldsymbol{e})} d\boldsymbol{\phi} d\boldsymbol{\theta} \\
&= \mathrm{KL}\Big(q(\boldsymbol{z}, \boldsymbol{\phi}, \boldsymbol{\theta}), p(\boldsymbol{z}, \boldsymbol{\phi}, \boldsymbol{\theta}|\boldsymbol{e})\Big)
\end{aligned} \tag{3.3}$$

Jensenの不等式より下限値$\mathcal{F}[q]$を最大化すれば，変分事後分布$q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi})$は$p(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{e})$の最良近似となることが分かる。また，これは(3.3)よりKLダイバージェンス基準

表 3.1: 3章で用いる変数の定義

| Symbol | Definition |
|---|---|
| $S$ | 音クリップ (音響ワード系列)の総数 |
| $S'$ | ミニバッチで処理する音クリップ数 |
| $\mathcal{S}^{VB}$ | 変分ベイズによりパラメータの更新を行う音響ワードの集合 |
| $\mathcal{S}^{GS}$ | ギブスサンプリングによりパラメータの更新を行う音響ワードの集合 |
| $T$ | 音響トピックのクラス数 |
| $M$ | 音響ワードのクラス数 |
| $N_s$ | 音クリップ$s$に含まれる音響ワードの数 |
| $s, s'$ | 音クリップのインデックス |
| $t$ | 音響トピックのクラスインデックス |
| $m$ | 音響ワードのクラスインデックス |
| $i$ | 音響ワード系列中の音響ワードの順序インデックス |
| $\backslash s, i$ | 音クリップ$s$の$i$番目の音響ワードを除くことを示す |
| $\boldsymbol{z}$ | 音響トピックを表す潜在変数 |
| $\boldsymbol{e}$ | 音響ワード系列を表す変数 |
| $z_{s,i}, e_{s,i}$ | 音クリップ$s$の$i$番目の音響トピックおよび音響ワード |
| $\boldsymbol{\theta}_s$ | 音クリップ$s$における音響トピック分布のパラメータ |
| $\boldsymbol{\phi}_t$ | 音響トピック$t$における音響ワード分布のパラメータ |
| $\alpha, \beta$ | Dirichlet分布の超パラメータ |
| $\mathrm{Dir}(\cdot)$ | Dirichlet分布 |
| $\mathrm{Categ}(\cdot)$ | Categorical分布 |
| $\Gamma(\cdot)$ | Gamma関数 |
| $\mathrm{KL}(\cdot,\cdot)$ | KLダイバージェンス |
| $n_t^s$ | 音クリップ$s$において音響トピック$t$に割り当てられた音響ワードの数 |
| $n_m^t$ | 音響トピック$t$において音響ワード$m$に割り当てられた音響ワードの数 |

において変分事後分布を真の分布に近似していると解釈する事が可能である。
ここで，以下の平均場近似を仮定する。

$$q(\boldsymbol{z}, \boldsymbol{\phi}, \boldsymbol{\theta}) = q(\boldsymbol{\phi}, \boldsymbol{\theta}|\boldsymbol{z})q(\boldsymbol{z}) = p(\boldsymbol{\phi}, \boldsymbol{\theta}|\boldsymbol{e}, \boldsymbol{z}, \alpha, \beta)q(\boldsymbol{z}) \tag{3.4}$$

$q(\boldsymbol{\phi}, \boldsymbol{\theta}|\boldsymbol{z}) = p(\boldsymbol{\phi}, \boldsymbol{\theta}|\boldsymbol{e}, \boldsymbol{z}, \alpha, \beta)$を仮定することは，平均場近似の一部を推定したい真の分布そのものと仮定することを意味し，これは以下に示すように，$\boldsymbol{\phi}, \boldsymbol{\theta}$の周辺化に相当する事が知られている[58]。

$$\begin{aligned}\mathcal{F}[q] &= \iint \sum_{\boldsymbol{z}} q(\boldsymbol{z}, \boldsymbol{\phi}, \boldsymbol{\theta}) \log \frac{p(\boldsymbol{e}, \boldsymbol{z}, \boldsymbol{\phi}, \boldsymbol{\theta}|\alpha, \beta)}{p(\boldsymbol{\phi}, \boldsymbol{\theta}|\boldsymbol{e}, \boldsymbol{z}, \alpha, \beta) q(\boldsymbol{z})} d\boldsymbol{\phi} d\boldsymbol{\theta} \\ &= \iint \sum_{\boldsymbol{z}} q(\boldsymbol{z}, \boldsymbol{\phi}, \boldsymbol{\theta}) \log \frac{p(\boldsymbol{e}, \boldsymbol{z}|\alpha, \beta)}{q(\boldsymbol{z})} d\boldsymbol{\phi} d\boldsymbol{\theta} \\ &= \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log p(\boldsymbol{e}, \boldsymbol{z}|\alpha, \beta) - \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log q(\boldsymbol{z}) \end{aligned} \tag{3.5}$$

ここで，$q(z_{s,i} = t) = \gamma_{sit} (i \in \mathcal{S}^{VB})$を音響ワード系列$s$の$i$番目の音響ワードに対する音響トピック$t$の事後分布とし，$i \in \mathcal{S}^{VB}$となる$i$について$\gamma_{sit}$の最大値を考える。$\mathcal{F}[\gamma_{sit}]$は各$\gamma_{sit}$について凸関数であるので，$\partial \mathcal{F}[\gamma_{sit}] / \partial \gamma_{sit} = 0$を解くことで以下を得る。なお，(3.6)の導出の詳細を付録のB.1に示す。

$$\gamma_{sit} \propto \exp \left[ E_{q(\boldsymbol{z}_{\backslash s,i})} \left\{ \log(n^{s}_{(\backslash s,i),t} + \alpha) + \log(n^{t}_{(\backslash s,i),m} + \beta) - \log(n^{t}_{(\backslash s,i),\cdot} + M\beta) \right\} \right] \tag{3.6}$$

但し，$q(\boldsymbol{z}_{\backslash s,i}) = \gamma_{11t} \gamma_{12t} \gamma_{13t} \cdots \gamma_{si-1t} \gamma_{si+1t} \cdots \gamma_{SN_{\boldsymbol{e}_S} t}$であり，また，「$\cdot$」は対応する変数について和を取ることを意味する。ここで，(3.6)では$q(\boldsymbol{z}_{\backslash s,i})$についての期待値を考えているが，後述するように$n^{s}_{(\backslash s,i),t}$は，

$$n^{s}_{(\backslash s,i),t} = n^{sVB}_{(\backslash s,i),t} + n^{sGS}_{t} \approx \sum_{i \in \mathcal{S}^{VB}, (\backslash s,i)} \gamma_{sit} + n^{sGS}_{t} \tag{3.7}$$

と近似できるため，(3.6)の右辺の$n_{(\backslash s,i),t}$は$\gamma_{sit}$（もしくは$q(\boldsymbol{z}_{\backslash s,i})$）に依存する形となっていることに注意する。また，$n^{sVB}_{t}$と$n^{sGS}_{t}$はそれぞれ，$s$番目の音響ワード系列の中で$\mathcal{S}^{VB}$および$\mathcal{S}^{GS}$に割り当てられた音響ワードのうち，音響トピック$t$に割り当てられた音響ワードの数を表す。

$E_{q(\boldsymbol{z}_{\backslash s,i})}\left[ \log(n^{s}_{(\backslash s,i),t} + \alpha) \right]$などを厳密に計算するには大きな演算コストを要するため，CVB0法[57]では，$\log(n^{s}_{(\backslash s,i),t} + \alpha) = \log(n^{sVB}_{(\backslash s,i),t} + n^{sGS}_{t} + \alpha)$を$n^{sVB}_{(\backslash s,i),t}$の関数

として，その期待値$E_{q(z_{\backslash s,i})}\left[n^{sVB}_{(\backslash s,i),t}\right]$周りで以下のようにテイラー展開し，

$$E_{q(z_{\backslash s,i})}\Big[\log(n^{sVB}_{(\backslash s,i),t}+n^{sGS}_{t}+\alpha)\Big] = E_{q(z_{\backslash s,i})}\Bigg[\log(E_{q(z_{\backslash s,i})}[n^{sVB}_{(\backslash s,i),t}]+n^{sGS}_{t}+\alpha) + \frac{E_{q(z_{\backslash s,i})}[n^{sVB}_{(\backslash s,i),t}]-n^{sVB}_{(\backslash s,i),t}}{E_{q(z_{\backslash s,i})}[n^{sVB}_{(\backslash s,i),t}]+n^{sGS}_{t}+\alpha} - \frac{\mathrm{Var}_{q(z_{\backslash s,i})}[n^{sVB}_{(\backslash s,i),t}]}{2(E_{q(z_{\backslash s,i})}[n^{sVB}_{(\backslash s,i),t}]+n^{sGS}_{t}+\alpha)^2}+\cdots\Bigg] \tag{3.8}$$

さらに，0次項のみで近似することにより以下の近似式を得ている。

$$\begin{aligned} E_{q(z_{\backslash s,i})}\Big[\log(n^{sVB}_{(\backslash s,i),t}+n^{sGS}_{t}+\alpha)\Big] &\approx E_{q(z_{\backslash s,i})}\Big\{\log(E_{q(z_{\backslash s,i})}[n^{sVB}_{(\backslash s,i),t}]+n^{sGS}_{t}+\alpha)\Big\} \\ &= \log(E_{q(z_{\backslash s,i})}[n^{sVB}_{(\backslash s,i),t}]+n^{sGS}_{t}+\alpha) \\ &= \log(\textstyle\sum_{i\in\mathcal{S}^{VB},(\backslash s,i)}\gamma_{sit}+n^{sGS}_{t}+\alpha) \\ &= \log(n^{sVB}_{(\backslash s,i),t}+n^{sGS}_{t}+\alpha) \end{aligned} \tag{3.9}$$

但し，$E_{q(z_{\backslash s,i})}[n^{sVB}_{(\backslash s,i),t}]$は音クリップ$s$のうち音響トピック$t$を生成する音響ワードの数の期待値である。また，音響トピックモデルの生成過程より音クリップから音響トピックの生成はCategorical分布に従うため，$n^{sVB}_{(\backslash s,i),t}\gg 0$のとき，$E_{q(z_{\backslash s,i})}[n^{sVB}_{(\backslash s,i),t}]\approx\sum_{i\in\mathcal{S}^{VB},\backslash s,i}\gamma_{sit}=n^{sVB}_{(\backslash s,i),t}$となる。これを(3.6)に代入すると，$\mathcal{S}^{VB}$に対する更新式を以下ように得る。

$$\gamma_{sit} \propto (n^{sVB}_{(\backslash s,i),t}+n^{sGS}_{t}+\alpha)(n^{tVB}_{(\backslash s,i),m}+n^{tGS}_{m}+\beta)(n^{tVB}_{(\backslash s,i),\cdot}+n^{tGS}_{\cdot}+M\beta)^{-1} \tag{3.10}$$

最終的に，下限値$\mathcal{F}(q)$が所与の収束条件を満たすまで，(3.10)を繰り返し計算することで最適な変分事後分布が得られる。また，$\boldsymbol{\phi},\boldsymbol{\theta}$の事後分布$q(\boldsymbol{\phi}),q(\boldsymbol{\theta})$を，$n^{s}_{t}$などをパラメータを用いたMAP推定量として，以下のように推定することも可能である[58]。

$$q(\boldsymbol{\phi};n^{tVB}_{m},n^{tGS}_{m}) = \frac{n^{tVB}_{m}+n^{tGS}_{m}+\beta}{n^{tVB}_{\cdot}+n^{tGS}_{\cdot}+M\beta} \tag{3.11}$$

$$q(\boldsymbol{\theta};n^{sVB}_{t},n^{sGS}_{t}) = \frac{n^{sVB}_{t}+n^{sGS}_{t}+\alpha}{n^{sVB}_{\cdot}+n^{sGS}_{\cdot}+T\alpha} \tag{3.12}$$

なお，バッチ型推定手法では，事前に全ての音響ワード列$\mathcal{S}^{VB}$を用意しておき，(3.10)を$s,i,t$に対して繰り返し計算することにより変分事後分布を推定

し，その後(3.11), (3.12)により$q(\boldsymbol{\phi}), q(\boldsymbol{\theta})$のMAP推定量を算出する。

### 3.3.3 CGS法によるモデルパラメータ推定

ギブスサンプリングでは，$s$番目の音響ワード系列の$i$番目の音響ワードに対応する音響トピック$z_{s,i}^{GS}$を除いた音響トピックの集合（$\boldsymbol{z}_{\backslash s,i}^{GS}$と表す）が与えられたときの，$z_{s,i}^{GS}$の事後分布$p(z_{s,i}^{GS}|\boldsymbol{z}_{\backslash s,i}^{GS}, \boldsymbol{z}^{VB}, \boldsymbol{e})$に従い，新たな$z_{s,i}$をサンプリングすることを繰り返し，モデルパラメータの推定を実現する。ここで事後分布$p(z_{s,i}^{GS}|\boldsymbol{z}_{\backslash s,i}^{GS}, \boldsymbol{z}^{VB}, \boldsymbol{e})$は以下のように書き下すことが可能である。

$$\begin{aligned} p(z_{s,i}^{GS}|\boldsymbol{z}_{\backslash s,i}^{GS}, \boldsymbol{z}^{VB}, \boldsymbol{e}) &= \frac{p(z_{s,i}^{GS}, \boldsymbol{z}_{\backslash s,i}^{GS}, \boldsymbol{z}^{VB}|\boldsymbol{e})}{p(\boldsymbol{z}_{\backslash s,i}^{GS}, \boldsymbol{z}^{VB}|\boldsymbol{e})} \\ &\propto \frac{p(\boldsymbol{e}|\boldsymbol{z}^{GS}, \boldsymbol{z}^{VB})}{p(\boldsymbol{e}_{\backslash s,i}|\boldsymbol{z}_{\backslash s,i}^{GS}, \boldsymbol{z}^{VB})} \cdot \frac{p(\boldsymbol{z}^{GS}, \boldsymbol{z}^{VB})}{p(\boldsymbol{z}_{\backslash s,i}^{GS}, \boldsymbol{z}^{VB})} \end{aligned} \tag{3.13}$$

また，(3.13)の右辺各項は，付録のB.2に示すパラメータの周辺積分を利用すると，以下のように計算可能である。

$$\frac{p(\boldsymbol{e}|\boldsymbol{z}^{GS}, \boldsymbol{z}^{VB})}{p(\boldsymbol{e}_{\backslash s,i}|\boldsymbol{z}_{\backslash s,i}^{GS}, \boldsymbol{z}^{VB})} = \frac{n_{(\backslash s,i),m}^{tGS} + \sum_{m=1}^{M}\sum_{i=1}^{n_t^{VB}} \delta_{sim}\gamma_{sit} + \beta}{n_{(\backslash s,i),\cdot}^{tGS} + \sum_{m=1}^{M}\sum_{i=1}^{n_t^{VB}} \gamma_{sit} + M\beta} \tag{3.14}$$

$$\frac{p(\boldsymbol{z}^{GS}, \boldsymbol{z}^{VB})}{p(\boldsymbol{z}_{\backslash s,i}^{GS}, \boldsymbol{z}^{VB})} = \frac{n_{(\backslash s,i),t}^{sGS} + \sum_{t=1}^{T}\sum_{i=1}^{N_{\boldsymbol{e}_s}^{VB}} \delta_{sit}\gamma_{sit} + \alpha}{n_{(\backslash s,i),\cdot}^{sGS} + \sum_{t=1}^{T}\sum_{i=1}^{N_{\boldsymbol{e}_s}^{VB}} \gamma_{sit} + T\alpha} \tag{3.15}$$

但し，$\delta_{sim}, \delta_{sit}$はクロネッカーのデルタ関数であり，$s$番目の音響ワード系列の$i$番目の音響ワードに割り当てられた音響ワードおよび音響トピックが$m$や$t$の場合に1となりその他の場合に0となる。

(3.14), (3.15)を(3.13)に代入すると，$\mathcal{S}^{GS}$に対する以下の更新式が得られる。なお，各更新において音響ワード系列$s$毎に正規化を行うため，(3.15)の分母は省略している。

$$p(z_{si}^{GS}|\boldsymbol{z}_{\backslash s,i}^{GS}, \boldsymbol{z}^{VB}, \boldsymbol{e}) \propto \frac{n_{(\backslash s,i),m}^{tGS} + \sum_{m=1}^{M}\sum_{i=1}^{n_t^{VB}} \delta_{sim}\gamma_{sit} + \beta}{n_{(\backslash s,i),\cdot}^{tGS} + \sum_{m=1}^{M}\sum_{i=1}^{n_t^{VB}} \gamma_{sit} + M\beta} (n_{(\backslash s,i),t}^{sGS} + \sum_{t=1}^{T}\sum_{i=1}^{N_{\boldsymbol{e}_s}^{VB}} \delta_{sit}\gamma_{sit} + \alpha) \tag{3.16}$$

CGS法を利用した潜在変数の割り当ての更新では，対象となる全ての潜在変数の更新を繰り返し行い，一定回数以上繰り返し更新した後に複数の潜在変数を独立にサンプルし（例えば1000回目，1100回目，1200回目，... の更新後の潜在変数の割り当てをサンプルする），各時刻のサンプルに対する潜在変数の分布を潜在変数の事後確率とすればよい。また，$\boldsymbol{\phi}, \boldsymbol{\theta}$の推定は(3.11), (3.12)により行うことができる。なお，従来のバッチ型推定手法では，CVB0法同様，事前に全ての音響ワード列$\mathcal{S}^{GS}$を用意しておき，(3.16)を$s, i$に対して繰り返しサンプリングすることによりパラメータを推定する。

## 3.4 音響トピックモデルにおけるハイブリッド型パラメータ推定手法のオンライン化

3.3.2, 3.3.3節で導出したモデルパラメータの推定法をオンライン手法へと拡張する。具体的には，音響ワード系列が得られる度に，(3.1)より各音響ワードが$\mathcal{S}^{VB}, \mathcal{S}^{GS}$のどちらに属するか判定し，以下で導出する$\mathcal{S}^{VB}, \mathcal{S}^{GS}$それぞれの更新式を適用しパラメータの推定を行えるようにする。

まず，$\mathcal{S}^{VB}$に対するオンライン推定の方法を考える。逐次的にデータが得られる場合のパラメータ推定手法として，事前にバッチ型アルゴリズムによりパラメータを仮推定し，その後オンライン推定により修正を加える手法[59]や，確率的勾配降下法 [60, 61]を利用した手法が提案されている。本稿では，事前学習が不要で過学習にも頑健とされる確率的勾配降下法に基づく手法を採用する。

(3.5), (3.6)より，全てのパラメータに対する周辺対数尤度の下限値$\mathcal{F}[q]$は$n_t^{sVB}, n_m^{tVB}, n_\cdot^{tVB}$に依存することが分かる。 $n_t^{sVB}$は各音響ワード系列で独立であるが，$n_m^{tVB}, n_\cdot^{tVB}$は全ての音響ワード系列でパラメータを共有しているため，そのままでは逐次的に更新するためにはできない。そこで，$n_t^{sVB}$は(3.10)により更新を行い， $n_m^{tVB}, n_\cdot^{tVB}$は，音響ワード系列$s$が得られる度に確率的勾配降下法により逐次的に更新を行うことにより，最終的に$\mathcal{F}[q]$を最大化することを目指す。具体的には，逐次的に得られた音響ワード系列に対して，まず$n_m^{tVB}, n_\cdot^{tVB}, n_m^{tGS}, n_\cdot^{tGS}, n_t^{sGS}$を固定したまま，$\gamma_{se_{s,i}t}$および$n_t^{sVB}$を(3.10)を用いて繰り返し更新し，その後，最適化された$\gamma_{se_{s,i}t}, n_t^{sVB}$を用いて$n_m^{tVB}, n_\cdot^{tVB}$を更新する。$n_m^{tVB}, n_\cdot^{tVB}$の更新においては，時間変数$k$および時間シフト係数$\tau_0$, 減衰係数$\kappa$を

表 3.2: 音響トピックモデルのためのCVB0/CGSハイブリッド型オンラインアルゴリズム

**Set** $\alpha$, $\beta$, $\kappa$, $\tau_0$, $d$
**Initialize** $n_m^{tVB(0)}, n_\cdot^{tVB(0)}, n_m^{tGS(0)}, n_\cdot^{tGS(0)}, \rho^{(0)}$
**Iterate** $k \leftarrow 1$ **to** $\lceil \frac{S}{S'} \rceil$
  **Initialize** $n_t^{sVB(0)}, n_t^{sGS(0)}, z_{se_{s,i}}, \gamma_{se_{s,i}t}$
  **Iterate over** $s, m, t$ **until convergence**
    **If** $n_{sm} > d$
      $\gamma_{smt}^{(k)} \propto (n_t^{sGS} + n_{(\backslash s,i),t}^{sVB} + \alpha)(n_m^{tGS} + n_{(\backslash s,i),m}^{tVB} + \beta)(n_\cdot^{tGS} + n_{(\backslash s,i),\cdot}^{tVB} + M\beta)^{-1}$
      $\gamma_{smt}^{(k)} = \gamma_{smt}^{(k)} \Big/ \sum_t \gamma_{smt}^{(k)}$
      $n_t^{sVB(k)} = n_t^{sVB(k-1)} + \sum_m \left\{ n_{sm}\gamma_{smt}^{(k)} - n_{sm}\gamma_{smt}^{(k-1)} \right\}$
    **Elseif** $n_{sm} \leq d$
      **Sample** $z_{s,i}^{GS(k)}$ **from**
        $$p(z_{s,i}^{GS} | \boldsymbol{z}_{\backslash s,i}, \boldsymbol{e}) \propto \frac{n_{(\backslash s,i),m}^{tGS} + n_m^{tVB} + \beta}{n_{(\backslash s,i),\cdot}^{tGS} + n_\cdot^{tVB} + M\beta} \cdot (n_{(\backslash s,i),t}^{sGS} + n_t^{sVB} + \alpha)$$
      $n_t^{sGS(k)} = n_t^{sGS(k-1)} + \sum_t \left\{ \delta_{sit} z_{s,i}^{GS(k)} - \delta_{sit} z_{s,i}^{GS(k-1)} \right\}$
    **End**
  **End**
  **If** $n_{sm} > d$
    $n_m^{tVB(k)} = (1 - \rho^{(k)}) n_m^{tVB(k-1)} + \rho^{(k)} \frac{S}{S'} \sum_{S'} n_{s'm} \gamma_{s'mt}^{(k)}$
    $n_\cdot^{tVB(k)} = \sum_m n_m^{tVB(k)}$
  **Elseif** $n_{sm} \leq d$
    $n_m^{tGS(k)} = (1 - \rho^{(k)}) n_m^{tGS(k-1)} + \rho^{(k)} \frac{S}{S'} \sum_{S', i' \in \mathcal{S}^{GS}} z_{i'}^{(k)}$
    $n_\cdot^{tGS(k)} = \sum_m n_m^{tGS(k)}$
  **End**
  **Set** $k \leftarrow k + 1$, $\rho^{(k)} = (k + \tau_0)^{-\kappa}$
**End**

用いて，スケジューリングを$\rho^{(k)} = (k + \tau_0)^{-\kappa}$と設定し，以下のように更新を行う。

$$n_m^{tVB(k)} = (1 - \rho^{(k)}) n_m^{tVB(k-1)} + \rho^{(k)} S n_m^{sVB(k)} \gamma_{smt}^{(k)} \tag{3.17}$$

$$n_\cdot^{tVB(k)} = \sum_m n_m^{tVB(k)} \tag{3.18}$$

次に，$\mathcal{S}^{GS}$に対するオンライン推定の方法を考える。 CGS法の場合においてもCVB0法と同様の考え方に基づき，確率的勾配降下法を利用する。 CGS法においても音響ワード系列が得られる度に$p(z_{s,i}^{GS} | \boldsymbol{z}_{\backslash s,i}^{GS}, \boldsymbol{z}^{VB}, \boldsymbol{e})$を用いて逐次的にサンプリングを繰り返す。つまり，逐次的に得られた音響ワード系列に対して$n_m^{tGS}, n_\cdot^{tGS}, n_m^{tVB}, n_\cdot^{tVB}, n_t^{sVB}$を固定したまま，(3.16)に従って$z_{s,i}$のサンプリングおよび$n_t^{sGS}$の更新を繰り返し行い，その後，最適化された$z_{s,i}$，もしくはその分布を用いて$n_m^{tGS}, n_\cdot^{tGS}$を更新する。加えて本論文では，逐次的に得られた$S'(\leq S)$個の音響ワード系列をまとめて繰り返し最適化を行うミニバッチ法[61]を導入する。ミニバッチ法を用いることにより，単一の音響ワード系列毎にパラメータを推定する場合よりも，確率的勾配降下法の各更新における勾配の統計的な分散を小さくする事が可能であり，安定したパラメータ推定が期待できる。ミニバッチ法を用いた場合の最終的な更新アルゴリズムは表3.2のようになる。

## 3.5 評価実験

### 3.5.1 実験条件

評価実験に先立ち，ユーザ行動収録音データセット（Activity :「会話をする」「料理をする」「食事をする」「PCを操作する」「新聞を読む」「掃除する」「移動する」「皿を洗う」「TVを見る」の9種の行動を含む収録音）と屋外環境収録音データセット（Situation :「自転車」「バス」「自動車」「会議」「オフィス」「公園」「街路」「電車」の8種の収録環境）の2種類の実環境データセットを収録した。ユーザ行動データセットは各データが16秒の長さからなる11,105の音クリップで構成されており，9,802の音クリップをモデル学習用に，1,303の音クリップを音響シーン分類の評価用に利用した。また，屋外環

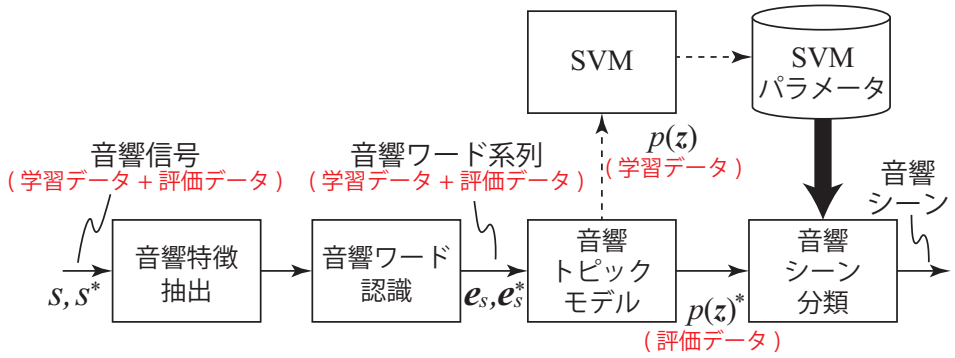



図 3.2: 音響トピックモデルを利用した音響シーン評価システムの構成

境データセットは各データが16秒の長さからなる25,277の音クリップで構成されており，21,461の音クリップをモデル学習用に，3,816の音クリップを評価用に利用した。

提案手法により学習されたモデルの性能を評価するため，提案手法とバッチ型手法により得られた音響トピックモデルのパラメータを用いてそれぞれ音響シーン分類を行い，分類性能を比較した。本章で取り扱う問題設定では，事前に音響シーンラベルが得られないことを想定しており，逐次的に音響シーンそのものをモデル化/分類することは難しい。そこで，図3.2に示すように先に全てのデータを用いて音響トピックモデルのパラメータを逐次学習し，パラメータ学習後に音響シーンのモデル化/分類を別途SVMにより行うことにより，モデルの性能を評価した。評価システムではまず，全ての音響信号に対してフレーム毎に音響特徴量（12次のMFCCs）を算出し，GMMクラスタリングを用いて音響ワードのモデル化と認識を行う。なお，音響ワードのモデル化には学習用のデータセットを用い，GMMクラスタリングにより推定されたガウス分布の各混合要素を1つの音響ワードとして定義した。次に，GMMにより認識された音響ワードの系列をモデル学習用，評価用の順に逐次的に提案アルゴリズムに入力し，モデルパラメータの推定を行う。ここで，モデル学習用データセットと評価用データセットにおいて，推定された音響トピックの変分事後分布をそれぞれ$p(z)$および$p(z)^*$とする。同一の音響シーンに含まれる音響トピックの分布は類似しているため，パラメータ推定用と評価用の事後分布$p(z)$および$p(z)^*$を比較することで音響シーンの推定が可能である[44]。本稿では，各音響ワード系列に対応する音響シーンラベルと$p(z)$を学習

表 3.3: 3章の実験に用いたパラメータ

| | |
|---|---|
| サンプリング周波数 | 16 kHz |
| 量子化ビット数 | 16 bits |
| フレームサイズ | 512 |
| フレームシフト | 256 |
| 音響ワードのクラス数 | 8–512 |
| ハイパパラメータ $\alpha$ | 3.33 |
| ハイパパラメータ $\beta$ | 0.1 |
| $\tau_0$ | 5.0 |
| $\kappa$ | 0.7 |
| ミニバッチサイズ $S'$ | 20 |

データとして，RBFカーネルを用いた多クラスSVMにより音響シーン識別器を構成し，$p(z)^*$を評価データとして音響シーンの分類を行った。なお，提案手法と同一のデータで，バッチ型ハイブリッド手法 (Hybrid (CVB0+CGS)およびHybrid (CVB+CGS))，オンライン型ハイブリッド手法 (oHybrid (CVB+CGS))，バッチ型CGS (CGS)，バッチ型CVB0 (CVB0)，オンライン型CVB0 (oCVB0)，によりパラメータ推定を行った場合の音響シーン分類も行った。また，その他の実験条件を表3.3に示す。

### 3.5.2 音響シーン分類性能を利用した学習モデルの評価結果

図3.3, 3.4にそれぞれユーザ行動データセット，屋外環境データセットにおける音響シーンの分類結果 (F-score)を示す。本実験では，$\boldsymbol{z}^{VB}, \boldsymbol{z}^{GS}$の初期値および入力する音響ワード系列の順序をランダムに選択して各手法で15回ずつ評価を行い，図3.3, 3.4にはF-scoreの平均値と95%の信頼区間を示している。また，CVB0法とCGS法のデータセットの分割は，$\mathcal{S}^{GS}$が各音響ワード系列で50%以下となるように$d = 2$とした。実験結果より，ユーザ行動データセット，屋外環境データセットの双方でバッチ型学習手法によるシーン分類結果がオンライン型学習よりも良いことが分かる。また，バッチ型学習と比較してオンライン学習ではF-scoreのばらつきが大きいことも分かる。この理由として，全ての音響ワード系列を用いて繰り返しパラメータを最適化するバッチ型手法と比較して，逐次的にパラメータを更新するオンライン学習手法

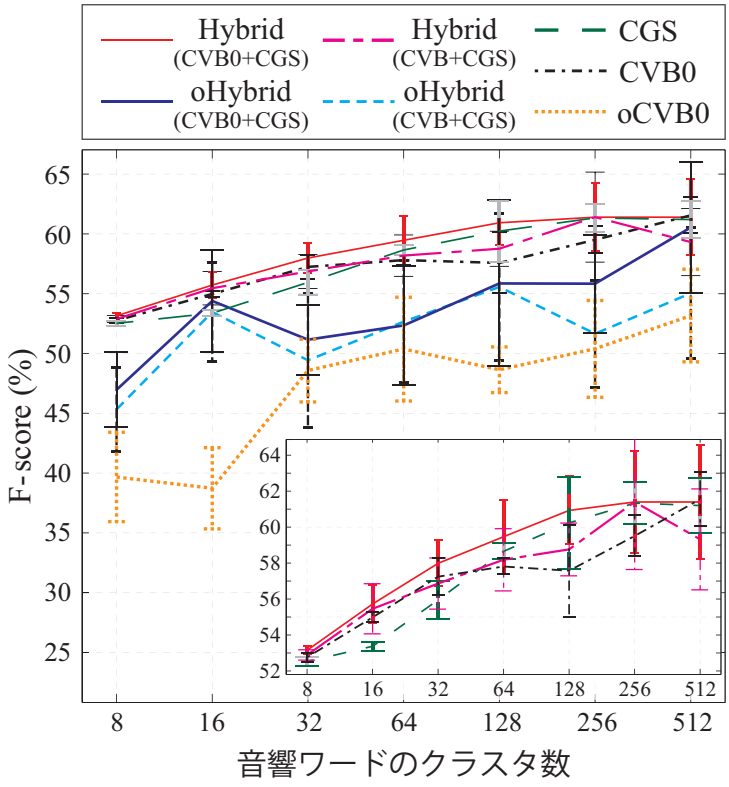



図 3.3: ユーザ行動データセットの音響シーン分類結果

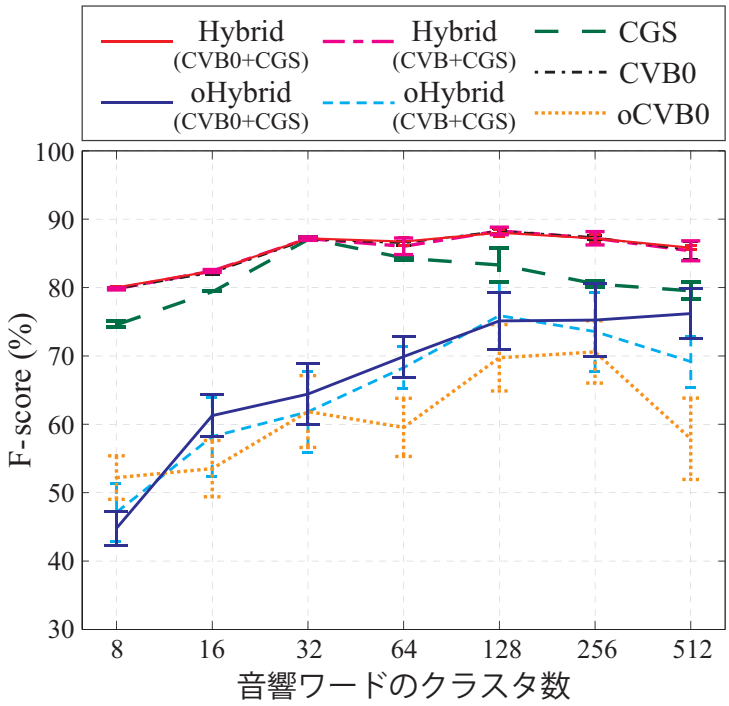



図 3.4: 屋外環境データセットの音響シーン分類結果

表 3.4: 提案手法によるCVB0法とCGS法のデータセットの分割閾値ごとの音響シーン分類結果 (F-score)およびモデルのパラメータ推定に要した時間 (秒)

| | d=0 (oCVB0) | d=1 | d=2 | d=4 | d=8 |
|---|---|---|---|---|---|
| Activity | 50.4% | 53.8% | 55.8% | 56.1% | 56.8% |
| | 483.2s | 526.7s | 585.9s | 697.8s | 983.3s |
| Situation | 70.6% | 74.0% | 75.2% | 74.8% | 76.3% |
| | 1247.9s | 1476.4s | 1579.3s | 1878.2s | 2525.9s |

では，入力される音響ワード系列の順序によりパラメータの推定が不安定になるためと考えられる。一方，オンライン型学習間でのF-scoreを比較すると，Hybrid (CVB+CGS)やoCVB0に比べ提案手法ではほぼ全ての条件において結果が向上していることが分かる。特にユーザ行動データセットでは，Hybrid (CVB0+CGS)によるF-scoreが61.4% (音響ワード数512の場合の平均値)であるのに対し，提案手法では60.5% (音響ワード数512の場合の平均値)とバッチ型の手法と遜色ない精度を実現出来ている。つまり，提案手法により，逐次的に得られた観測からでもバッチ型の手法により学習された音響トピックモデルのパラメータと類似したパラメータが学習できていると推測される。また，本実験では半数以上の音響ワードがCVB0法によってパラメータ推定されるように$d$を選択しているにも関わらず，提案手法はoCVB0よりも音響シーンの推定性能が大幅に向上しており，効率的にCVB0法とCGS法の切り替えが行えていることも分かる。

また，図3.3, 3.4より，ユーザ行動データセット，屋外環境データセットともに音響ワードのクラスタ数が増加するにつれて音響シーンの分類結果は向上していることが分かる。他方，いずれのデータセットにおいても音響ワードのクラスタ数が大きくなるにつれて，音響シーン分類性能の向上率は小さくなっており，音響ワードのクラスタ数は512程度あれば良いことも分かる。

表3.4に提案手法における，データセットの分割閾値$d$を変化させた際の音響シーン分類結果（F-score）とモデルのパラメータ推定に要した平均時間（秒）を示す。本実験では音響ワードの種類の数として256を用い，その他の実験条件は先の実験と同様とした。実験結果より，$d$を大きくするにつれて音響シーンの分類性能が向上するが，$d$が大きくなるにつれて分類性能は向上しに

くくなることが分かる。この理由として，多数のサンプルが観測されている音響ワードに関連するパラメータでは，CVB0法とCGS法により推定される解が近いためと考えられる。一方で，$d$が大きくなるにつれてパラメータの推定に要する時間は大幅に増大する事が分かる。これらの結果より，データセットの分割閾値$d$を2〜4程度とすれば効率的に音響シーンが分類できると考えられる。

### 3.5.3 モデルの汎化性能

提案モデルの汎化性能を評価するため，各データベースに対するパープレキシティを算出した。ここで，パープレキシティは音響ワード系列における音響ワードの平均的な接続分岐数を表す。パープレキシティが小さくなることは各音響トピックを特徴付ける音響ワードを，過剰適合することなく学習できていることを表す。図3.5, 3.6にそれぞれのデータベースの評価用データに対するパープレキシティを示す。実験結果より，シーン分類性能同様，オンライン型手法よりもバッチ型手法の方が優れた汎化性を示していることが分かる。また，オンライン型手法を比較すると提案手法はoHybrid (CVB+CGS)よりも優れた汎化性能を示している事が分かる。一方，提案手法とoCVB0で大きな差は見られなかった。この理由として，パープレキシティでは出現頻度の少ない音響ワードに対して汎化性能が反映されづらい事が考えられる。つまり，提案手法とoCVB0の違いは出現頻度の少ない音響ワードに対するパラメータ推定手法の違いであるため，パープレキシティに大きな差が見られなかったと推測される。音響シーンの分類性能では提案手法が優れていたことと併せて考察すると，出現頻度の少ない音響ワードが提案手法とoCVB0の音響シーンの分類性能に大きな影響を与えていると考えられ，出現頻度の少ない音響ワードに対してCGS法を適用することにより音響シーン分類の向上が実現できたと考えられる。

### 3.5.4 パラメータ推定の計算コスト

オンライン学習による音響シーンのモデル化の利点として学習データを事前に全て用意しておく必要がないことを先に述べたが，モデル学習時間の削減効果もまた利点として挙げられる。表3.5に各手法により音響シーンをモデル化した場合の平均学習時間を示す。なお，本実験ではパラメータ推

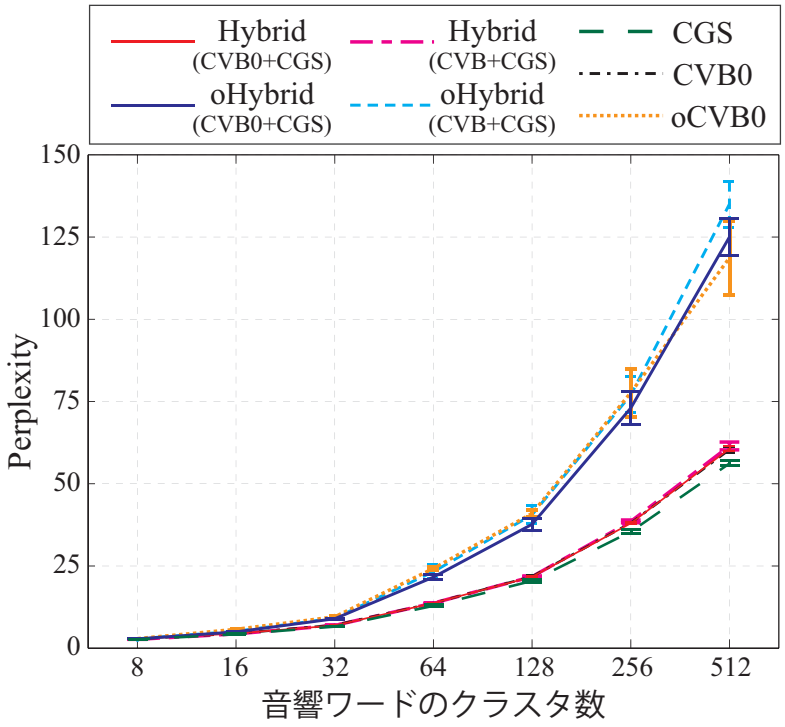



図 3.5: ユーザ行動データセットの評価用データに対するパープレキシティ

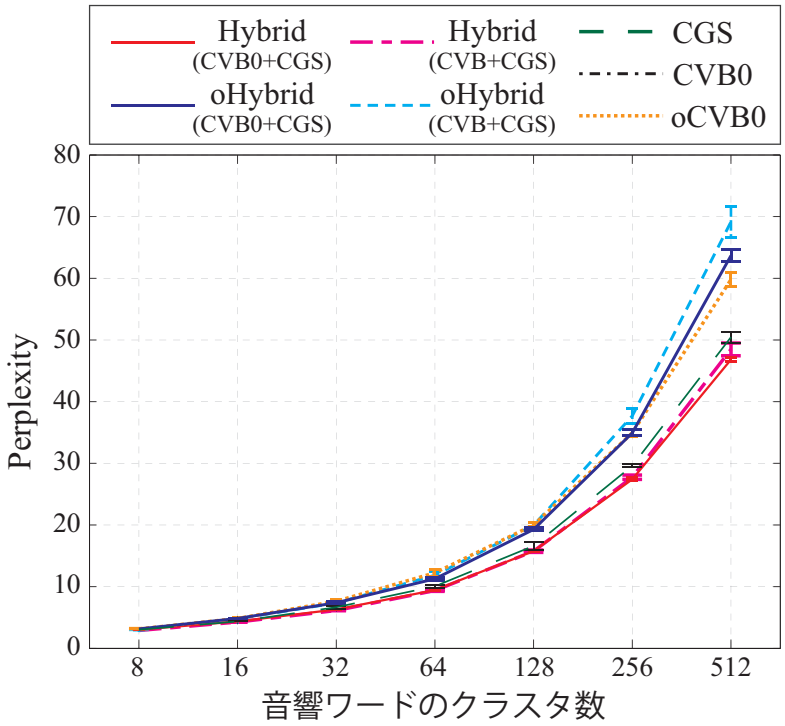



図 3.6: 屋外環境データセットの評価用データに対するパープレキシティ

表 3.5: 音響トピックモデルのパラメータ推定に要した時間 (秒)

| | Activity | Situation |
|---|---|---|
| Hybrid (CVB0+CGS) | 1,556.1 | 4,435.6 |
| Hybrid (CVB+CGS) | 1,921.5 | 5,477.7 |
| oHybrid (CVB0+CGS) | 585.9 | 1,579.3 |
| oHybrid (CVB+CGS) | 799.3 | 2,020.5 |
| CGS | 21,089.4 | 69,549.6 |
| CVB0 | 1,577.2 | 3,862.4 |
| oCVB0 | 483.2 | 1,247.9 |

定手法の差分による計算コストを評価するため，各学習手法のパラメータの推定に要した平均時間を示している。また，学習はIntel Core™ i7-3820QM (2.70GHz)のCPUおよびDDR3 SDRAM (16GB, 1600MHz)の主記憶装置を搭載したPCを用いた。

実験結果より，オンライン型手法はバッチ型手法と比較してパラメータ推定に要する時間が1/3程度まで削減出来ていることが分かる。その理由としてオンライン型の音響トピックモデルでは，$\theta_s^{VB}, \theta_s^{GS}$は各時刻毎に入力された音響信号毎に学習されるため，バッチ型学習と比較して各時刻の演算量が$S'/S$になることや，各時刻毎に繰り返し最適化するパラメータの数がバッチ型学習と比較して少ないため，繰り返し最適化の収束速度が向上することが挙げられる。また，提案手法とoHybrid (CVB+CGS)を比較すると，提案手法ではパラメータ推定に要する時間が3/4程度まで削減出来ていることが分かり，音響シーンの推定性能および計算コストの双方で提案手法が優れている事が分かる。

## 3.6 3章のまとめ

本章では，音情報が逐次的に得られる場合に，音響シーンをモデル化し分類する手法として，崩壊型変分ベイズの0次近似手法と崩壊型ギブスサンプリングのハイブリッド法に基づくオンライン型音響トピックモデルを提案した。提案モデルでは，推定するパラメータを局所解に収束しやすいものとその他に分割し，それぞれに崩壊型ギブスサンプリングと崩壊型変分ベイズの0次近

似手法に基づくオンライン型パラメータ手法を適用することで，逐次的に得られた音響ワード系列からモデルパラメータの推定が可能になるとともに，パラメータ推定精度の向上も可能となる。実環境音を用いた音響シーンの分類実験を行った結果，提案手法では，音響シーン分類を従来のバッチ型学習法と同程度の精度で実現でき，音響シーンのモデル学習が従来のバッチ手法に対して1/3程度の演算時間で実現できることが明らかになった。

実際の動画投稿サイトやオンラインストレージでは，本実験で評価したデータベースよりも大規模なデータが日々蓄積され続けており，今後はそれらの実データを利用した評価実験を進められる必要がある。また，本章の実験では，学習後のパラメータを用いて音響シーン分類性能を評価したが，パラメータ学習中のモデルを用いた場合の音響シーン分類性能についてもより詳細に検討される必要がある。

# 4

# 間欠的な欠損を有する観測に基づく音響シーン分類

## 4.1 はじめに

従来の音響シーン分類手法では，観測された音情報に欠損は含まれないことを仮定している。しかしながら実際の環境下では，背景雑音や音の過大な入力レベルにより観測に不要なノイズが混入したり部分的な欠損が起こる。また，ネットワーク越しに収録した音を伝送する場合には，パケットロスによる情報の欠損もしばしば起こり，音の観測の数%から数十%の情報が欠損する場合もある。他方，見守りや監視といった用途に音響シーン分類を適用する際，プライバシーの観点から音の連続的な収録が困難である場合も多く，部分的に収録された音からシーンを分類する状況も考えられる。プライバシーの観点から部分的に音を収録する場合には，非常に限られた観測（全体の数%のみ観測など）から音響シーンを分析する場合も想定されるため，部分的な観測から音響シーンを分類する技術の実現は重要である。

これまでにも，雑音が混入した観測から音響シーン分類や音響イベント分類を行う手法が提案されており，例えばWaldoらは，雑音が混入した観測から音響イベントを分類するための方法として，Minimum staticsに基づくスペクトル減算[62]を前処理として適用した後，MFCCsとSVMを適用する手法を提案している[19]。また，Schröderらは，雑音のパワースペクトル密度推定とDesision directed法によって事前に雑音抑圧を行った上で，特徴抽出と音響イベント分類を行う方法を提案している[20]。 Lopatkaらは，複数のマイクロホンアレイにより雑音が混入した観測から音響イベント分類を行う手法を提案している[63]。文献[63]ではMPEG-7に基づく特徴量[26]とSVMを用いた音響イベント分類と，音響インテンシティを利用した音源定位を組み合わせ，雑音でない観測に対する音響イベント分類を実現している。文献[19]，[20]のように，スペクトル減算[64, 65, 62]やウィーナフィルタ[66]，最小平均二乗誤差規範短時間振幅スペクトル推定器[67]など従来の雑音抑圧手法を前処理に適用した後に音響シーン分類や音響イベント分類手法を適用する方法もあるが，これらの手法は主に音声の強調を主な目的としたものである。しかしながら，音響シーン分類問題に従来の雑音抑圧手法を適用する際は，「音響シーンに関連する音」と「雑音」の違いを明確にして雑音抑圧処理を行うことができないため，従来法では十分な音響シーン分類性能を実現することが困難である。

一方で，雑音抑圧が困難になる程雑音が混入した観測やパケットロスを含む観測，プライバシーの観点から部分的に収録された観測，つまり，一部が欠損している観測に対する音響シーン分類手法や音響イベント分類手法については過去に例がない。そこで本章では，観測の一部に欠損を有する音クリップから音響シーンを分類する問題を扱う。欠損を含む音の観測から音響シーンを分類するための最も単純なアプローチとして，欠損した音情報を無視し，観測された音情報のみを用いて音響シーン分類を行う方法が考えられる。しかしながら，欠損した音情報を無視して音響シーン分類を行う場合，音の欠損率が高くなるにつれて音響シーンを特徴づける重要な情報が失われてしまい音響シーン分類性能が低下する。また，欠損した観測を補完した後に音響シーン分類を行う手法も考えられる。欠損した部分を補完するため手法としては，例えば，隠れマルコフモデル (HMM: Hidden Markov Model) [68]を利用した手法が考えられる。具体的には，近い時間に発生する音は強く関連しているという事実に基づき，隠れマルコフモデルにより音の遷移をモデル化し欠損した音の復元に利用する。しかしながら，欠損した音響信号や音響特徴

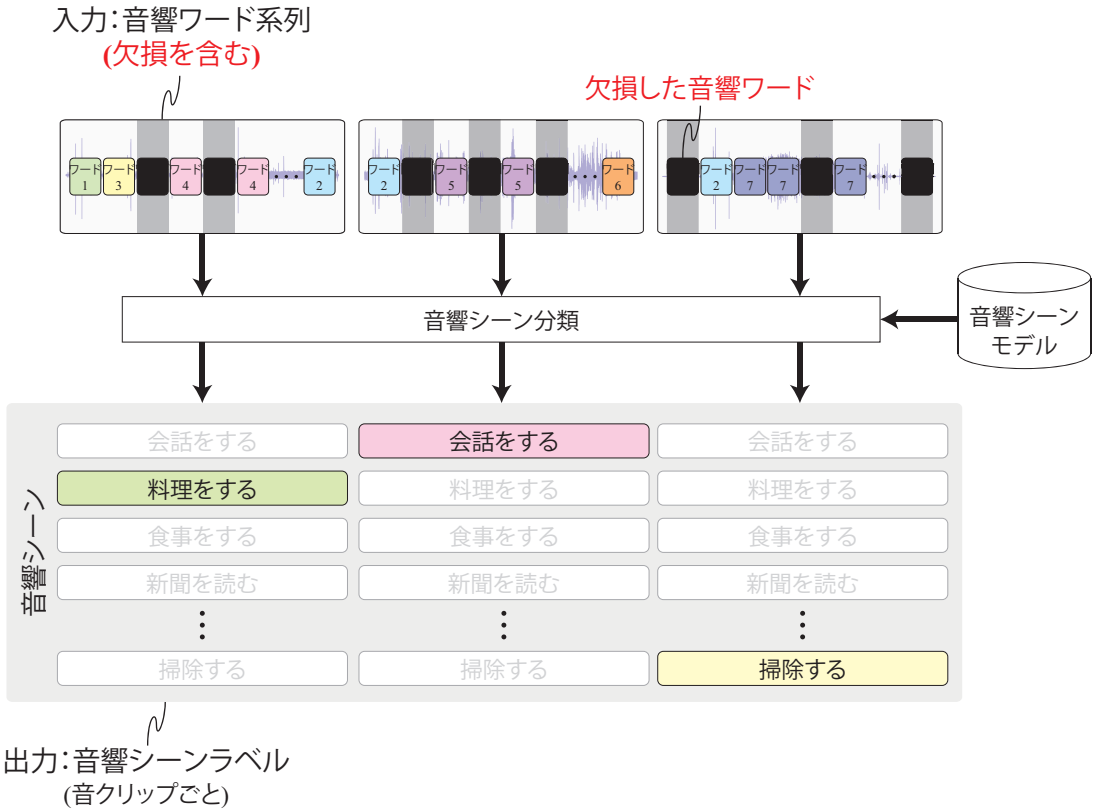



図 4.1: 4章で扱う音響シーン分類問題

量そのものを復元するのは容易ではないため，隠れマルコフモデルのような手法を音響信号に直接適用することは難しい。

他方，ある音響シーンから生成される観測は音響シーンが持つ潜在的構造に従い確率的に決定されると考えると，欠損した観測を潜在的な（観測できない）確率変数とみなすことで容易に扱えるようになる。これは，音響トピックモデルで論じた確率的な生成モデルと非常に親和性が高く，観測された音響ワード系列と欠損した観測の生成過程を同一の枠組みで扱えることを示している。そこで，本論文では図4.1に示すように，音情報の観測の欠損を音響ワードの欠損と捉え，欠損を有する音響ワード系列から欠損した音響ワードを復元して同時に音響シーンを分析する問題を考える。なお，欠損は音響ワード単位で発生するものとし，音響ワードが欠損しているか否かは既知である問題を考える。音のクリップの検出やウインドノイズの検出手法など，欠損が観測しているか否かを検出する手法は数多く検討されていることや，

パケットロスやプライバシー保護を目的とした観測の場合，音響ワードが欠損している箇所は多くの場合において既知であることを踏まえると，本章の問題設定は妥当と考えられる。

提案手法では，欠損した音響ワードを推定するための方法として，近い時間に発生する音響ワードは関連しているという事実を利用する。つまり，音響ワードが生成される際には時間的に前後の音響ワードに影響を受けると仮定し，音響ワードの時間遷移を考慮した新たな音響トピックモデルを提案する。

本章の構成は以下のとおりである。 4.2節では，間欠的に欠損を有する音クリップから，欠損した音情報を推定しながら音響シーンを分類する手法として，音の時間連続性を考慮した新たな音響トピックモデルの基本的なアイデアについて述べる。 4.3節では，音の時間連続性を考慮した新たな音響トピックモデルの実現方法として，音響ワードの時間遷移をマフコフモデルによりモデル化した教師あり音響トピックモデルについて述べる。また，崩壊型ギブスサンプリングによる提案モデルのパラメータ推定手法と，提案モデルを用いた音響シーンの分類方法についても述べる。 4.4節では，音の時間連続性を考慮した音響トピックモデルの別の実現方法として，音響トピックの時間遷移を考慮した教師あり音響トピックモデルについて述べる。併せて，崩壊型ギブスサンプリングによる提案モデルのパラメータ推定手法と，提案モデルを用いた音響シーンの分類方法についても述べる。 4.5節において，実環境収録音を用いた提案手法の性能評価実験について議論し，4.6節で本章のまとめを述べる。

## 4.2 音の時間的な連続性を考慮した音響シーンのモデル化

音響ワードが生成される際には時間的に前後の音響ワードに影響を受けると仮定し，音響ワードの時間遷移を考慮した新たな音響トピックモデルを提案する。ここで，音響ワードの時間遷移を考慮した音響トピックモデルは以下の2通り考えられる。

- 音響ワードの時間遷移を考慮した音響トピックモデル
- 音響トピックの時間遷移を考慮した音響トピックモデル

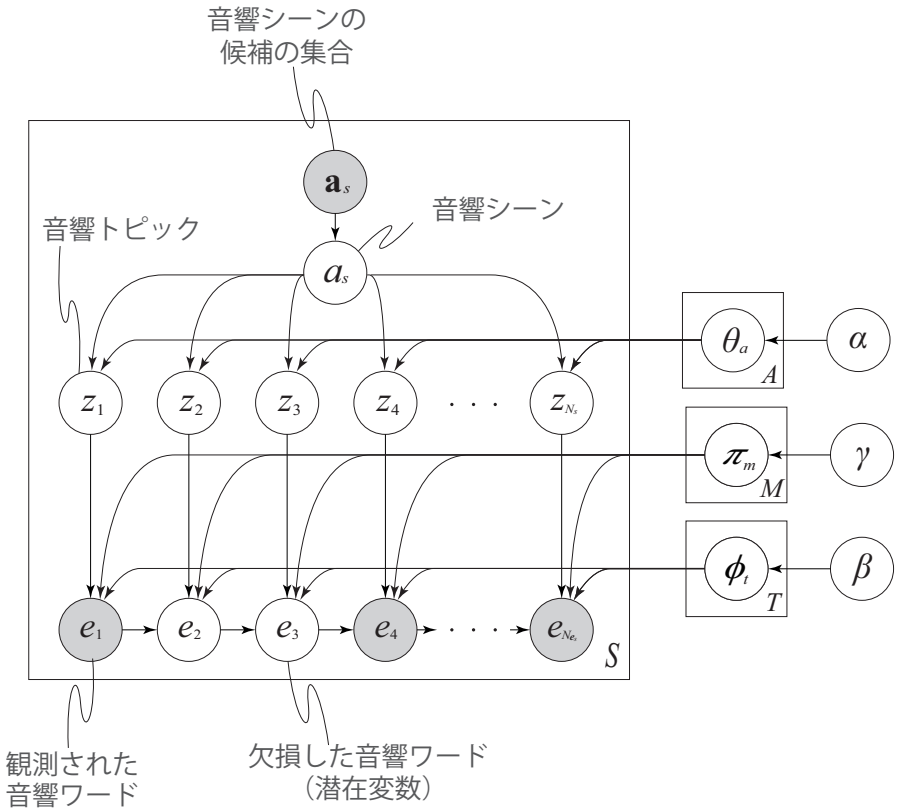



図 4.2: 音響ワードの時間遷移を考慮した教師あり音響トピックモデルのグラフィカルモデル

前者は音響ワードの遷移関係を直接モデル化する手法であり，後者は隠れマルコフモデルのように音響ワードの潜在的な構造の遷移をモデル化する手法である。以降では，それぞれのモデル化の方法について述べる。

## 4.3 音響ワード遷移型教師あり音響トピックモデル

### 4.3.1 音響ワード系列の生成過程のモデル化

図4.2に示すように，音響ワードの遷移を考慮した音響ワード系列の生成過程を考える。本モデルは，教師あり音響トピックモデルのように，音響シーンと音響トピック，音響ワードの階層的な生成過程であるが，音響ワードの生成が音響トピックにのみ依存しているのではなく，一時刻前に生成された音響ワードにも依存している点が従来のモデルと異なる。提案手法では，音響

ワードの遷移は時間遷移関係を確率的にモデル化する単純マルコフ過程 (SMP: Simple Markov process)を用いる。

提案モデルの生成過程をAlgorithm 4.1に示す。なお，本論文では音響ワードの遷移を考慮した教師あり音響トピックモデルを，音響ワード遷移型教師あり音響トピックモデル (Word-transition sATM: Acoustic word transition-based supervised acoustic topic model)と称する。また，本章で用いる変数の定義を表4.1に示す。 音響ワード遷移型教師あり音響トピックモデルでは欠損した観測を潜在変数とし，観測された音響ワードを観測変数とする。欠損した観測を潜在変数として表現することにより，モデルパラメータと同時に欠損した音響ワードを推定可能という利点がある。 Algorithm 4.1より，提案モデルにおいて音響ワードの生成は，音響ワードの生成確率$\boldsymbol{\phi}_{z_{s,i}}$と音響ワードの前時刻からの遷移確率$\boldsymbol{\pi}_{e_{s,i-1}}$の結合確率によって表される。また，音響ワードの生成確率$\boldsymbol{\phi}_{z_{s,i}}$と音響ワードの前時刻からの遷移確率$\boldsymbol{\pi}_{e_{s,i}}$はそれぞれCategorical分布に従い，それぞれのCategoriacl分布はDirichlet事前分布に従うものとする。このとき，全ての音響ワード系列$\mathcal{S}$の生成確率$p(\mathcal{S}|\alpha,\beta,\gamma,\mathcal{A})$は以下のように表すことができる。

$$
\begin{aligned}
&p(\mathcal{S}|\alpha,\beta,\gamma,\mathcal{A}) \\
&\quad= \prod_{s=1}^{S}\prod_{i=1}^{N_{e_s}}\sum_{\boldsymbol{a}}\sum_{\boldsymbol{z}}\sum_{\boldsymbol{m}} p(e_{s,i}|e_{s,i-1},z_{s,i},\alpha,\beta,\gamma,a_s)p(z_{s,i}|a_s,\alpha)p(a_s|\mathbf{a}_s) \\
&\quad= \prod_{s=1}^{S}\Bigg[\mathrm{Uni}(a_s|\mathbf{a}_s)\sum_{\boldsymbol{a}}\int \mathrm{Categ}(\boldsymbol{\theta}_a|a_s,\alpha)\prod_{i=1}^{N_{e_s}}\Bigg\{\sum_{\boldsymbol{z}}\mathrm{Categ}(z_{s,i}|\boldsymbol{\theta}_a)\int \mathrm{Dir}(\boldsymbol{\phi}_t|\beta) \\
&\qquad\cdot\int\sum_{\boldsymbol{m}}\mathrm{Dir}(\boldsymbol{\pi}_m|\gamma)\mathrm{Categ}(e_{s,i}|\boldsymbol{\phi}_t,\boldsymbol{\pi}_m,e_{s,i-1},z_{s,i})d\boldsymbol{\pi}_m d\boldsymbol{\phi}_t\Bigg\}d\boldsymbol{\theta}_a\Bigg] \\
&\quad= \frac{1}{A}\prod_{s=1}^{S}\Bigg[\int\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\prod_{i=1}^{N_{e_s}}\Bigg\{\prod_{t=1}^{T}\theta_{a,t}^{\alpha-1+n_t^a}\int\frac{\Gamma(M\beta)}{\Gamma(\beta)^M}\prod_{m=1}^{M}\phi_{t,m}^{\beta-1+n_m^t} \\
&\qquad\cdot\int\frac{\Gamma(M\gamma)}{\Gamma(\gamma)^M}\prod_{m^+=1}^{M}\pi_{m^-,m^+}^{\gamma-1+n_{m^+}^{m^-}}d\boldsymbol{\pi}_m d\boldsymbol{\phi}_t\Bigg\}d\boldsymbol{\theta}_a\Bigg]
\end{aligned}
\tag{4.1}
$$

音響ワード遷移型教師あり音響トピックモデルを用いて音響シーンの分類を行うためには，事前に学習データから事後確率を最大とするモデルパラメ

表 4.1: 4章で用いる変数の定義

| Symbol | Definition |
|---|---|
| $S$ | 音クリップ（音響ワード系列）の総数 |
| $A$ | 音響シーンのクラス数 |
| $T$ | 音響トピックのクラス数 |
| $M$ | 音響ワードのクラス数 |
| $N_{\boldsymbol{e}_s}$ | 音響ワード系列$\boldsymbol{e}_s$に含まれる音響ワードの数 |
| $s$ | 音クリップのインデックス |
| $a$ | 音響シーンのクラスインデックス |
| $t$ | 音響トピックのクラスインデックス |
| $m$ | 音響ワードのクラスインデックス |
| $i$ | 音響ワード系列中の音響ワードの順序インデックス |
| $\mathcal{S}$ | 音響ワード系列の集合 |
| $\mathcal{A}$ | 全ての音クリップの音響シーンの集合の集合 |
| $\mathbf{a}_s$ | 音クリップ$s$の音響シーンの候補 |
| $a_s$ | 音クリップ$s$の音響シーン |
| $\boldsymbol{z}$ | 音響トピックを表す潜在変数 |
| $\boldsymbol{e}_s$ | $s$番目の音響ワード系列 |
| $\boldsymbol{\theta}_s, \boldsymbol{\theta}_a$ | 音クリップ$s$または音響シーン$a$における音響トピック生成分布のパラメータ |
| $\theta_{a,t}$ | 音響シーン$a$における音響トピック$t$の生成確率 |
| $\boldsymbol{\phi}_t$ | 音響トピック$t$における音響ワード生成分布のパラメータ |
| $\phi_{t,m}$ | 音響トピック$t$における音響ワード$m$の生成確率 |
| $\boldsymbol{\pi}_m, \boldsymbol{\pi}_t$ | 音響ワードおよび音響トピックの遷移分布のパラメータ |
| $\pi_{m^-,m^+}$ | 音響ワード$m^-$から$m^+$への遷移確率 |
| $\pi_{t^-,t^+}$ | 音響トピック$t^-$から$t^+$への遷移確率 |
| $\alpha, \beta, \gamma$ | Dirichlet分布のパラメータ |
| $n_t^a, n_m^t, n_{m^+}^{m^-}$ | 音響シーン$a$において音響トピック$t$に割り当てられた音響ワードの数, etc. |
| $n_\cdot^a, n_\cdot^t, n_\cdot^{m^-}$ | 音響シーン$a$に割り当てられた音響ワードの数, etc. |
| $\backslash s, i$ | 音クリップ$s$の$i$番目の音響ワードを除くことを示す |
| Dir$(\cdot)$ | Dirichlet分布 |
| Categ$(\cdot)$ | Categorical分布 |
| Uni$(\cdot)$ | Uniform分布 |
| $\Gamma(\cdot)$ | Gamma関数 |

**Algorithm 4.1** 音響ワード遷移型教師あり音響トピックモデルにおける音響ワード系列の生成過程

**A set of possible acoustic scenes** $\mathbf{a}_s$ **is given,**
**for** $a = 1$ **to** $A$ **do**
  **Choose** $\boldsymbol{\theta}_a$ $\sim$ Dirichlet($\alpha$)
**end for**
**for** $t = 1$ **to** $T$ **do**
  **Choose** $\boldsymbol{\phi}_t$ $\sim$ Dirichlet($\beta$)
**end for**
**for** $m = 1$ **to** $M$ **do**
  **Choose** $\boldsymbol{\pi}_m$ $\sim$ Dirichlet($\gamma$)
**end for**
**for** $s = 1$ **to** $S$ **do**
  **Choose** $a_s$ $\sim$ Uniform($\mathbf{a}_s$)
  **for** $i = 1$ **to** $N_{\boldsymbol{e}_s}$ **do**
    **Choose** $z_{s,i} \mid \boldsymbol{\theta}_{a_s}, a_s$ $\sim$ Categorical($\boldsymbol{\theta}_{a_s}$)
    **Choose** $e_{s,i} \mid \boldsymbol{\phi}_{z_{s,i}}, \boldsymbol{\pi}_{e_{s,i-1}}, z_{s,i}, e_{s,i-1}$
      $\sim$ Categorical($\boldsymbol{\phi}_{z_{s,i}}$), Categorical($\boldsymbol{\pi}_{e_{s,i-1}}$)
  **end for**
**end for**

ータを推定し，その後，教師あり音響トピックモデルと同様の方法により音響クリップ毎の音響シーンと，欠損した音響ワードを推定する。

$$\arg\max_{a_s, e_{s,i}} p(a_s, e_{s,i} | \boldsymbol{\theta}_a, \boldsymbol{\phi}_t, \boldsymbol{\pi}_m, \boldsymbol{e}_s, \alpha, \beta, \gamma) \tag{4.2}$$

本論文ではモデルパラメータの推定方法として，次節に示す崩壊型ギブスサンプリングによる方法を用いる。

### 4.3.2 CGS法によるモデルパラメータ推定

提案する音響ワード遷移型教師あり音響トピックモデルや音響トピック遷移型教師あり音響トピックモデルを用いて，音響シーンの分類と欠損した音響ワードの推定を行うためには，事後確率が最大となるようなモデルパラメータを推定する必要がある。しかしながら，音響ワード遷移型教師あり音響トピックモデルや音響トピック遷移型教師あり音響トピックモデルにおいてモ

デルパラメータを解析的に推定することは容易ではない。そこで本論文では，ベイズ推定，特に崩壊型ギブスサンプリング (CGS: Collapsed Gibbs sampling) [51, 52]に基づく繰り返し最適化手法によりモデルパラメータを推定する。ベイズ推定に基づくモデルパラメータの推定法としては他にも変分ベイズ法 (VB: Variational Bayes) [50, 49]や期待値伝搬法 (EP: Expectation propagation) [53]などが挙げられるが，本論文では初期値や学習データによるバイアスの影響を受けにくいとされる崩壊型ギブスサンプリングを採用する。

崩壊型ギブスサンプリングによるパラメータ推定では，音響シーンや音響トピック，欠損した音響ワードに対応する潜在変数を，学習用の音響ワード系列に対する条件付き事後確率に基づいて繰り返しサンプリングする。ここで，音響ワード遷移型教師あり音響トピックモデルにおいて音響トピック$z_{s,i}$および音響ワード$e_{s,i}$は各時間フレーム$s,i$毎に生成され，音響シーン$a_s$は各音響ワード系列$s$毎に1つ生成されるものと仮定している。そのため，崩壊型ギブスサンプリングを用いて提案モデルのパラメータを推定するためには，1) $z_{s,i}, e_{s,i}$ と 2) $a_s$を別々にサンプリングし，パラメータを推定する必要がある。以下では，1) $z_{s,i}, e_{s,i}$，2) $a_s$それぞれのサンプリングに対する条件付き事後確率の導出について述べる。

### 音響トピックおよび音響ワードの事後確率

音響トピックと音響ワードの同時事後確率$p(e_{s,i}, z_{s,i}|\boldsymbol{e}_{\backslash s,i}, \boldsymbol{z}_{\backslash s,i}, \mathcal{A}, \alpha, \beta, \gamma)$を考える。まず，$p(e_{s,i}, z_{s,i}|\boldsymbol{e}_{\backslash s,i}, \boldsymbol{z}_{\backslash s,i}, \mathcal{A}, \alpha, \beta, \gamma)$は以下のように変形できる。

$$
\begin{aligned}
&p(e_{s,i}, z_{s,i}|\boldsymbol{e}_{\backslash s,i}, \boldsymbol{z}_{\backslash s,i}, \mathcal{A}, \alpha, \beta, \gamma) \\
&\quad = \frac{p(\boldsymbol{e}|\boldsymbol{z}, \mathcal{A}, \alpha, \beta, \gamma)}{p(\boldsymbol{e}_{\backslash s,i}|\boldsymbol{z}_{\backslash s,i}, \mathcal{A}, \alpha, \beta, \gamma)} \cdot \frac{p(\boldsymbol{z}|\mathcal{A}, \alpha, \beta, \gamma)}{p(\boldsymbol{z}_{\backslash s,i}|\mathcal{A}, \alpha, \beta, \gamma)} \\
&\quad = \frac{p(\boldsymbol{e}|\boldsymbol{z}, \beta, \gamma)}{p(\boldsymbol{e}_{\backslash s,i}|\boldsymbol{z}_{\backslash s,i}, \beta, \gamma)} \cdot \frac{p(\boldsymbol{z}|\mathcal{A}, \alpha)}{p(\boldsymbol{z}_{\backslash s,i}|\mathcal{A}, \alpha)}
\end{aligned} \tag{4.3}
$$

ここで，(4.3)において$p(z_{s,i}|\boldsymbol{\theta}_a, a_s), p(e_{s,i}|\boldsymbol{\phi}_{z_{s,i}}, z_{s,i}), p(e_{s,i}|\boldsymbol{\pi}_{e_{s,i-1}}, e_{s,i-1})$はCategorical分布に従うこと，および，$p(\boldsymbol{\theta}|\alpha), p(\boldsymbol{\phi}|\beta), p(\boldsymbol{\pi}|\gamma)$はDirichlet分布に従うことを考慮

すると，$p(\boldsymbol{z}|\mathcal{A},\alpha)$と$p(\boldsymbol{e}|\boldsymbol{z},\beta,\gamma)$は以下のように表現可能である。

$$\begin{aligned} p(\boldsymbol{z}|\mathcal{A},\alpha) &= \int p(\boldsymbol{z},\boldsymbol{\theta}|\mathcal{A},\alpha)d\boldsymbol{\theta} \\ &= \left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\right)^A \prod_{a=1}^{A} \frac{\prod_{t=1}^{T}\Gamma(n_t^a+\alpha)}{\Gamma(n_\cdot^a+T\alpha)} \end{aligned} \tag{4.4}$$

$$\begin{aligned} p(\boldsymbol{e}|\boldsymbol{z},\beta,\gamma) &= \int\int p(\boldsymbol{e},\boldsymbol{\phi},\boldsymbol{\pi}|\boldsymbol{z},\beta,\gamma)d\boldsymbol{\phi}d\boldsymbol{\pi} \\ &= \prod_{s=1}^{S}\prod_{i=1}^{N_{e_s}} \int\int p(e_{s,i},\boldsymbol{\phi}|\boldsymbol{z},\beta)p(e_{s,i},\boldsymbol{\pi}|e_{s,i-1},\gamma)d\boldsymbol{\phi}d\boldsymbol{\pi} \end{aligned} \tag{4.5}$$

ただし，

$$\prod_{s=1}^{S}\prod_{i=1}^{N_{e_s}} \int p(e_{s,i},\boldsymbol{\phi}|\boldsymbol{z},\beta)d\boldsymbol{\phi} = \left(\frac{\Gamma(M\beta)}{\Gamma(\beta)^M}\right)^T \prod_{t=1}^{T} \frac{\prod_{m=1}^{M}\Gamma(n_m^t+\beta)}{\Gamma(n_\cdot^t+M\beta)} \tag{4.6}$$

$$\prod_{s=1}^{S}\prod_{i=1}^{N_{e_s}} \int p(e_{s,i},\boldsymbol{\pi}|e_{s,i-1},\gamma)d\boldsymbol{\pi} = \left(\frac{\Gamma(M\gamma)}{\Gamma(\gamma)^M}\right)^M \prod_{m^-=1}^{M} \frac{\prod_{m^+=1}^{M}\Gamma(n_{m^+}^{m^-}+\gamma)}{\Gamma(n_\cdot^{m^-}+M\gamma)} \tag{4.7}$$

ガンマ関数の性質$\Gamma(x+1)/\Gamma(x)=x$を用いて(4.4)–(4.7)を変形し，(4.3)に代入することで，音響トピックと音響ワードの事後確率（崩壊型ギブスサンプリングの更新式）は以下のように得られる。

$$\begin{aligned} p(e_{s,i},z_{s,i}|\boldsymbol{e}_{\backslash s,i},\boldsymbol{z}_{\backslash s,i},\mathcal{A},\alpha,\beta,\gamma) \propto{}& (n_{(\backslash s,i),t}^{a}+\alpha)\cdot\frac{n_{(\backslash s,i),m}^{t}+\beta}{n_{(\backslash s,i),\cdot}^{t}+M\beta} \\ &\cdot\frac{(n_{(\backslash s,i),e_{s,i}}^{e_{s,i-1}}+\gamma)\left\{n_{(\backslash s,i),e_{s,i+1}}^{e_{s,i}}+\delta_{e_{s,i-1},e_{s,i}}\cdot\delta_{e_{s,i},e_{s,i+1}}+\gamma\right\}}{n_{(\backslash s,i),\cdot}^{e_{s,i}}+\delta_{e_{s,i-1},e_{s,i}}+M\gamma} \end{aligned} \tag{4.8}$$

ただし，$n_{(\backslash s,i),e_{s,i}}^{e_{s,i-1}}$は音響ワード系列$\boldsymbol{e}_s$から$e_{s,i}$を除いたもののうち，$e_{s,i-1}$から$e_{s,i}$に遷移した音響ワードの数を表す。また，$\delta_{e_{s,i-1},e_{s,i}}$はクロネッカーのデルタ関数を表し，$e_{s,i-1}=e_{s,i}$の場合1となりその他の場合0となる。

音響ワード$e_{s,i}$が欠損していない場合には，それぞれの更新においては音響トピック$z_{s,i}$のみをサンプリングすれば良く，音響トピック$z_{s,i}$の事後確率

は(4.9)で与えられる。

$$p(z_{s,i}|\boldsymbol{e}_{\backslash s,i}, \boldsymbol{z}_{\backslash s,i}, \mathcal{A}, \alpha, \beta, \gamma) \propto (n^{a}_{(\backslash s,i),t} + \alpha) \cdot \frac{n^{t}_{(\backslash s,i),m} + \beta}{n^{t}_{(\backslash s,i),\cdot} + M\beta} \tag{4.9}$$

なお，音響ワード遷移型教師あり音響トピックモデルにおける音響トピックおよび音響ワードの事後確率の導出の詳細を付録のC.1に示す。

### 音響シーンの事後確率

音響シーンの事後確率$p(a_s|\boldsymbol{e}, \boldsymbol{z}, \boldsymbol{a}_{\backslash s}, \alpha, \beta, \gamma)$を考える。 $p(\boldsymbol{a}_s|\alpha)$がUniform分布に従うこと，および，$p(\boldsymbol{e})$が$\boldsymbol{a}_s$に依存していないことを踏まえると，音響シーンの事後確率は以下で与えられる。

$$\begin{aligned} &p(a_s|\boldsymbol{e}, \boldsymbol{z}, \boldsymbol{a}_{\backslash s}, \alpha, \beta, \gamma) \\ &\quad = \frac{p(\boldsymbol{e}|\boldsymbol{z}, \mathcal{A}, \beta, \gamma)}{p(\boldsymbol{e}|\boldsymbol{z}, \boldsymbol{a}_{\backslash s}, \beta, \gamma)} \cdot \frac{p(\boldsymbol{z}|\mathcal{A}, \alpha)}{p(\boldsymbol{z}|\boldsymbol{a}_{\backslash s}, \alpha)} \cdot \frac{p(\mathcal{A}|\alpha)}{p(\boldsymbol{a}_{\backslash s}|\alpha)} \\ &\quad \propto \frac{p(\boldsymbol{z}|\mathcal{A}, \alpha)}{p(\boldsymbol{z}|\boldsymbol{a}_{\backslash s}, \alpha)} \end{aligned} \tag{4.10}$$

音響トピックや音響ワードの事後分布と同様の導出を行うことで，音響シーン$a_s$の事後分布は以下のように得られる。

$$p(a_s|\boldsymbol{e}, \boldsymbol{z}, \boldsymbol{a}_{\backslash s}, \alpha, \beta, \gamma) \propto \frac{n^{a}_{(\backslash s),t} + \alpha}{n^{a}_{(\backslash s),\cdot} + T\alpha} \tag{4.11}$$

なお，提案手法では音響シーンはそれぞれの音響ワード系列毎に1つ含まれると仮定しているため，音響ワード系列毎に1回サンプリングすれば良い。

### 事後分布の更新

音響トピックや音響ワード，音響シーンの事後確率が与えられたとき，音響トピックや音響ワード，また音響ワード遷移の事後分布は，(4.8), (4.9), (4.11)を用いて十分に更新された後の潜在変数の割り当てを用いて推定することができる。具体的には，事後分布のパラメータは十分大きい数のサンプルの分布

の平均を用いて以下のように推定できる。

$$\overline{\theta}_{a,t} = \frac{1}{N_G}\sum_{j=1}^{N_G}\left\{\frac{\sum_{N_{e_s}}\hat{z}_{s,i,t,j}+\alpha}{\sum_{N_{e_s}}\sum_{t}\hat{z}_{s,i,t,j}+T\alpha}\right\} \tag{4.12}$$

$$\overline{\phi}_{t,m} = \frac{1}{N_G}\sum_{j=1}^{N_G}\left\{\frac{\sum_{s}\sum_{N_{e_s}}\hat{z}_{s,i,t,j}\,\hat{e}_{s,i,m,j}+\beta}{\sum_{s}\sum_{N_{e_s}}\sum_{m}\hat{z}_{s,i,t,j}\hat{e}_{s,i,m,j}+M\beta}\right\} \tag{4.13}$$

$$\overline{\pi}_{m^-,m^+} = \frac{1}{N_G}\sum_{j=1}^{N_G}\left\{\frac{\sum_{s}\sum_{N_{e_s}}\hat{e}_{s,i,m^-,j}\cdot\hat{e}_{s,i+1,m^+,j}+\gamma}{\sum_{s}\sum_{N_{e_s}}\sum_{m^+}\hat{e}_{s,i,m^-,j}\cdot\hat{e}_{s,i+1,m^+,j}+M\gamma}\right\} \tag{4.14}$$

ここで，$N_G$はサンプリングの回数を表す。ただし，サンプリングを開始した初期の期間（バーンイン区間）は上記の計算に利用しない。また，$\hat{z}_{s,i,t,j}$と$\hat{\boldsymbol{e}}_{s,i,t,j}$は$j$番目にサンプリングされた音響トピックと音響ワードをそれぞれ表し，もし，音響トピックや音響ワードのインデックスが$t$や$m$の場合は1となり，その他の場合は0となる。

## 4.4 音響トピック遷移型教師あり音響トピックモデル

### 4.4.1 音響ワード系列の生成過程のモデル化

音響ワードの時間遷移を考慮した別のモデル化として，図4.3に示すように音響トピックの時間遷移を利用する方法が考えられる。つまり，音響ワードの時間関係を，隠れマルコフモデルのように潜在変数（音響トピック）の時間遷移によってモデルする。本モデルの生成過程は具体的にはAlgorithm 4.2のように表現可能である。なお，本論文ではこのモデルをトピック遷移型教師あり音響トピックモデル (Topic-transition sATM: Acoustic topic transition-based supervised acoustic topic model)と称する。音響トピック遷移型教師あり音響トピックモデルでは，音響ワード遷移型教師あり音響トピックモデル同様に欠損した観測を潜在変数として表現する。また，音響トピック遷移型教師あり音響トピックモデルでは，それぞれの音響トピックの生成は音響トピックの生成確率$\boldsymbol{\theta}_{a_s}$と，音響トピックの前時刻からの遷移確率$\boldsymbol{\pi}_{t_{s,i-1}}$の結合確率によって表される。また，音響トピックの生成確率$\boldsymbol{\theta}_{a_s}$と音響トピックの前時刻から

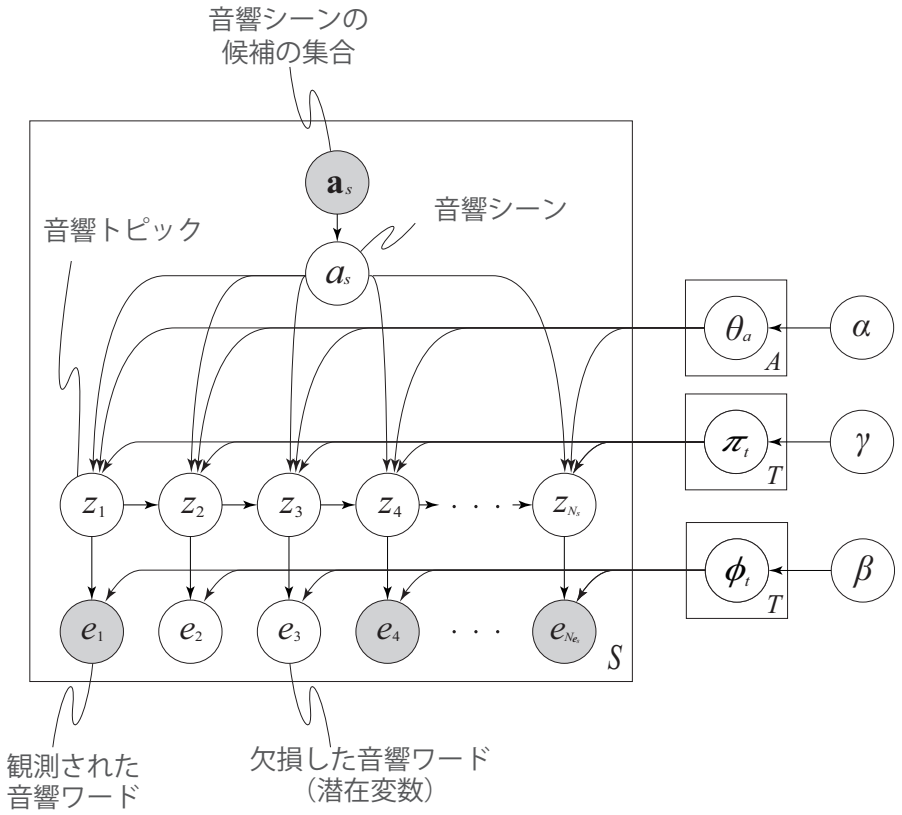



図 4.3: 音響トピックの時間遷移を考慮した教師あり音響トピックモデルのグラフィカルモデル

の遷移確率$\boldsymbol{\phi}_{t_{s,i-1}}$はそれぞれCategorical分布に従い，それぞれのCategoriacl分布はDirichlet事前分布に従うものとする。このとき，全ての音響ワード系列$\mathcal{S}$の生成確率$p(\mathcal{S}|\alpha,\beta,\gamma,\mathcal{A})$は以下のように表すことができる。

$$
\begin{aligned}
&p(\mathcal{S}|\alpha,\beta,\gamma,\mathcal{A}) \\
&\quad= \prod_{s=1}^{S}\prod_{i=1}^{N_{\boldsymbol{e}_s}}\sum_{\boldsymbol{a}}\sum_{\boldsymbol{z}} p(e_{s,i}|z_{s,i},\beta)p(z_{s,i}|a_s,z_{s,i-1},\alpha,\gamma)p(a_s|\mathbf{a}_s) \\
&\quad= \prod_{s=1}^{S}\left[\mathrm{Uni}(a_s|\mathbf{a}_s)\sum_{\boldsymbol{a}}\int \mathrm{Categ}(\boldsymbol{\theta}_a|a_s,\alpha)\prod_{i=1}^{N_{\boldsymbol{e}_s}}\left\{\sum_{\boldsymbol{z}}\int \mathrm{Dir}(\boldsymbol{\pi}_z|\gamma)\right.\right. \\
&\qquad\left.\left.\cdot\,\mathrm{Categ}(z_{s,i}|\boldsymbol{\theta}_a,\boldsymbol{\pi}_t,z_{s,i-1})\int \mathrm{Dir}(\boldsymbol{\phi}_t|\beta)\mathrm{Categ}(e_{s,i}|\boldsymbol{\phi}_t,z_{s,i})d\boldsymbol{\phi}_t d\boldsymbol{\pi}_t\right\}d\boldsymbol{\theta}_a\right]
\end{aligned}
$$

**Algorithm 4.2** 音響トピック遷移型教師あり音響トピックモデルにおける音響ワード系列の生成過程

**A set of possible acoustic scenes** $\mathrm{a}_s$ **is given,**
**for** $a = 1$ **to** $A$ **do**
  **Choose** $\boldsymbol{\theta}_a$ $\sim$ Dirichlet($\alpha$)
**end for**
**for** $t = 1$ **to** $T$ **do**
  **Choose** $\boldsymbol{\phi}_t$ $\sim$ Dirichlet($\beta$)
  **Choose** $\boldsymbol{\pi}_t$ $\sim$ Dirichlet($\gamma$)
**end for**
**for** $s = 1$ **to** $S$ **do**
  **Choose** $a_s$ $\sim$ Uniform($\mathrm{a}_s$)
  **for** $i = 1$ **to** $N_{\boldsymbol{e}_s}$ **do**
    **Choose** $z_{s,i} \mid \boldsymbol{\theta}_{a_s}, \boldsymbol{\pi}_{z_{s,i-1}}, a_s, z_{s,i-1}$
      $\sim$ Categorical($\boldsymbol{\theta}_{a_s}$), Categorical($\boldsymbol{\pi}_{z_{s,i-1}}$)
    **Choose** $e_{s,i} \mid \boldsymbol{\phi}_{z_{s,i}}, z_{s,i}$ $\sim$ Categorical($\boldsymbol{\phi}_{z_{s,i}}$)
  **end for**
**end for**

$$= \frac{1}{A}\prod_{s=1}^{S}\left[\int \frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\prod_{i=1}^{N_{\boldsymbol{e}_s}}\left\{\prod_{t=1}^{T}\theta_{a,t}^{\alpha-1+n_t^a}\int \frac{\Gamma(T\gamma)}{\Gamma(\gamma)^T}\prod_{t^+=1}^{T}\pi_{t^-,t^+}^{\gamma-1+n_{t^+}^{t^-}}\right.\right.$$
$$\left.\left.\cdot\int \frac{\Gamma(M\beta)}{\Gamma(\beta)^M}\prod_{m=1}^{M}\phi_{t,m}^{\beta-1+n_m^t}d\boldsymbol{\phi}_t d\boldsymbol{\pi}_t\right\}d\boldsymbol{\theta}_a\right] \tag{4.15}$$

なお，音響トピック遷移型教師あり音響トピックモデルを用いて音響シーンを分類するためには音響ワード遷移型教師あり音響トピックモデルと同様の方法を用いれば良い。

### 4.4.2 CGS法によるモデルパラメータ推定

音響トピック遷移型教師あり音響トピックモデルにおける音響トピック$z_{s,i}$，音響ワード$e_{s,i}$および音響シーン$a_s$の事後確率は音響ワード遷移型教師あり音響トピックモデルの場合と同様の方法で導出することが可能である。以下では，音響ワード遷移型教師あり音響トピックモデルの場合と同様，1) $z_{s,i}, e_{s,i}$，2) $a_s$それぞれのサンプリングに対する条件付き事後確率の導出について議論する。

### 音響トピックおよび音響ワードの事後確率

まず，音響トピックと音響ワードの同時事後確率$p(e_{s,i}, z_{s,i}|\boldsymbol{e}_{\backslash s,i}, \boldsymbol{z}_{\backslash s,i}, \mathcal{A}, \alpha, \beta, \gamma)$を以下のように変形する。

$$\begin{aligned} p(e_{s,i}, z_{s,i}|\boldsymbol{e}_{\backslash s,i}, \boldsymbol{z}_{\backslash s,i}, \mathcal{A}, \alpha, \beta, \gamma) &= \frac{p(\boldsymbol{e}|\boldsymbol{z}, \mathcal{A}, \alpha, \beta, \gamma)}{p(\boldsymbol{e}_{\backslash s,i}|\boldsymbol{z}_{\backslash s,i}, \mathcal{A}, \alpha, \beta, \gamma)} \cdot \frac{p(\boldsymbol{z}|\mathcal{A}, \alpha, \beta, \gamma)}{p(\boldsymbol{z}_{\backslash s,i}|\mathcal{A}, \alpha, \beta, \gamma)} \\ &= \frac{p(\boldsymbol{e}|\boldsymbol{z}, \beta)}{p(\boldsymbol{e}_{\backslash s,i}|\boldsymbol{z}_{\backslash s,i}, \beta)} \cdot \frac{p(\boldsymbol{z}|\mathcal{A}, \alpha, \gamma)}{p(\boldsymbol{z}_{\backslash s,i}|\mathcal{A}, \alpha, \gamma)} \end{aligned} \tag{4.16}$$

$p(z_{s,i}|\boldsymbol{\theta}_a, a_s)$，$p(e_{s,i}|\boldsymbol{\phi}_{z_{s,i}}, z_{s,i})$，$p(z_{s,i}|\boldsymbol{\pi}_{z_{s,i-1}}, z_{s,i-1})$がCategorical分布に従うこと，また，$p(\boldsymbol{\theta}|\alpha)$，$p(\boldsymbol{\phi}|\beta)$，$p(\boldsymbol{\pi}|\gamma)$がDirichlet分布に従うことを踏まえると，$p(\boldsymbol{z}|\mathcal{A}, \alpha, \gamma)$および$p(\boldsymbol{e}|\boldsymbol{z}, \beta)$は以下のように記述できる。

$$\begin{aligned} p(\boldsymbol{z}|\mathcal{A}, \alpha, \gamma) &= \iint p(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\pi}|\mathcal{A}, \alpha, \gamma) d\boldsymbol{\theta} d\boldsymbol{\pi} \\ &= \prod_{s=1}^{S} \prod_{i=1}^{N_{es}} \iint p(z_{s,i}, \boldsymbol{\theta}|\mathcal{A}, \alpha) p(z_{s,i}, \boldsymbol{\pi}|z_{s,i-1}, \gamma) d\boldsymbol{\theta} d\boldsymbol{\pi} \end{aligned} \tag{4.17}$$

$$\begin{aligned} p(\boldsymbol{e}|\boldsymbol{z}, \beta) &= \int p(\boldsymbol{e}, \boldsymbol{\phi}|\boldsymbol{z}, \beta) d\boldsymbol{\phi} \\ &= \left( \frac{\Gamma(M\beta)}{\Gamma(\beta)^M} \right)^T \prod_{t=1}^{T} \frac{\prod_{m=1}^{M} \Gamma(n_m^t + \beta)}{\Gamma(n_\cdot^t + M\beta)} \end{aligned} \tag{4.18}$$

ただし，

$$\prod_{s=1}^{S} \prod_{i=1}^{N_{es}} \int p(z_{s,i}, \boldsymbol{\theta}|\boldsymbol{a}, \alpha) d\boldsymbol{\theta} = \left( \frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T} \right)^A \prod_{a=1}^{A} \frac{\prod_{t=1}^{T} \Gamma(n_t^a + \alpha)}{\Gamma(n_\cdot^a + T\alpha)} \tag{4.19}$$

$$\prod_{s=1}^{S} \prod_{i=1}^{N_{es}} \int p(z_{s,i}, \boldsymbol{\pi}|z_{s,i-1}, \gamma) d\boldsymbol{\pi} = \left( \frac{\Gamma(T\gamma)}{\Gamma(\gamma)^T} \right)^T \prod_{t^-=1}^{T} \frac{\prod_{t^+=1}^{T} \Gamma(n_{t^+}^{t^-} + \gamma)}{\Gamma(n_\cdot^{t^-} + T\gamma)} \tag{4.20}$$

ここで(4.17)–(4.20)を(4.16)に代入し，ガンマ関数の性質$\Gamma(x+1)/\Gamma(x) = x$を考慮すると以下を得る。

$$\frac{p(\boldsymbol{z}|\mathcal{A}, \alpha, \gamma)}{p(\boldsymbol{z}_{\backslash s,i}|\mathcal{A}, \alpha, \gamma)} = \frac{n_{(\backslash s,i),t}^a + \alpha}{n_{(\backslash s,i),\cdot}^a + T\alpha} \cdot \frac{n_{(\backslash s,i),t_{s,i}}^{t_{s,i-1}} + \gamma}{n_{(\backslash s,i),\cdot}^{t_{s,i-1}} + T\gamma} \cdot \frac{n_{(\backslash s,i),t_{s,i+1}}^{t_{s,i}} + \delta_{t_{s,i-1},t_{s,i}} \cdot \delta_{t_{s,i},t_{s,i+1}} + \gamma}{n_{(\backslash s,i),\cdot}^{t_{s,i}} + \delta_{t_{s,i-1},t_{s,i}} + T\gamma} \tag{4.21}$$

表 4.2: 提案モデルにおけるパラメータの事後確率

| 変数 | 音響ワード遷移型<br>教師あり音響トピックモデル | 音響トピック遷移型<br>教師あり音響トピックモデル |
|---|---|---|
| 音響トピック&ワード<br>(音響ワードは欠損)<br>$p(e_{s,i}, z_{s,i} \mid \boldsymbol{e}_{\backslash s,i}, \boldsymbol{z}_{\backslash s,i}, \mathcal{A}, \alpha, \beta, \gamma)$ | $\propto (n^{a}_{(\backslash s,i),t}+\alpha)\frac{n^{t}_{(\backslash s,i),m}+\beta}{n^{t}_{(\backslash s,i),\cdot}+M\beta}(n^{e_{s,i-1}}_{(\backslash s,i),e_{s,i}}+\gamma)$<br>$\cdot\frac{n^{e_{s,i}}_{(\backslash s,i),e_{s,i+1}}+\delta_{e_{s,i-1},e_{s,i}}\delta_{e_{s,i},e_{s,i+1}}+\gamma}{n^{e_{s,i}}_{(\backslash s,i),\cdot}+\delta_{e_{s,i-1},e_{s,i}}+M\gamma}$ | $\propto (n^{a}_{(\backslash s,i),t}+\alpha)\frac{n^{t}_{(\backslash s,i),m}+\beta}{n^{t}_{(\backslash s,i),\cdot}+M\beta}(n^{t_{s,i-1}}_{(\backslash s,i),t_{s,i}}+\gamma)$<br>$\cdot\frac{n^{t_{s,i}}_{(\backslash s,i),t_{s,i+1}}+\delta_{t_{s,i-1},t_{s,i}}\delta_{t_{s,i},t_{s,i+1}}+\gamma}{n^{t_{s,i}}_{(\backslash s,i),\cdot}+\delta_{t_{s,i-1},t_{s,i}}+T\gamma}$ |
| 音響トピック<br>(音響ワードは観測)<br>$p(z_{s,i} \mid \boldsymbol{e}, \boldsymbol{z}_{\backslash s,i}, \mathcal{A}, \alpha, \beta, \gamma)$ | $\propto (n^{a}_{(\backslash s,i),t}+\alpha)\frac{n^{t}_{(\backslash s,i),m}+\beta}{n^{t}_{(\backslash s,i),\cdot}+M\beta}$ | $\propto (n^{a}_{(\backslash s,i),t}+\alpha)\frac{n^{t}_{(\backslash s,i),m}+\beta}{n^{t}_{(\backslash s,i),\cdot}+M\beta}(n^{t_{s,i-1}}_{(\backslash s,i),t_{s,i}}+\gamma)$<br>$\cdot\frac{n^{t_{s,i}}_{(\backslash s,i),t_{s,i+1}}+\delta_{t_{s,i-1},t_{s,i}}\delta_{t_{s,i},t_{s,i+1}}+\gamma}{n^{t_{s,i}}_{(\backslash s,i),\cdot}+\delta_{t_{s,i-1},t_{s,i}}+T\gamma}$ |
| 音響シーン<br>$p(a_s \mid \boldsymbol{e}, \boldsymbol{z}, \boldsymbol{a}_{\backslash s}, \alpha, \beta, \gamma)$ | $\propto \frac{n^{a}_{(\backslash s),t}+\alpha}{n^{a}_{(\backslash s),\cdot}+T\alpha}$ | |

$$\frac{p(\boldsymbol{e}|\boldsymbol{z},\beta)}{p(\boldsymbol{e}_{\backslash s,i}|\boldsymbol{z}_{\backslash s,i},\beta)} = \frac{n^{t}_{(\backslash s,i),m}+\beta}{n^{t}_{(\backslash s,i),\cdot}+M\beta} \tag{4.22}$$

ただし，$n^{t_{s,i-1}}_{(\backslash s,i),t_{s,i}}$は$z_{s,i}$を除いた音響トピックのうち，$t_{s,i-1}$から$t_{s,i}$に遷移した数を表す。(4.21)と(4.22)を(4.16)に代入することで，最終的に音響トピックと音響ワードの事後確率は以下のようになる。

$$\begin{aligned} p(e_{s,i}, z_{s,i}|\boldsymbol{e}_{\backslash s,i}, \boldsymbol{z}_{\backslash s,i}, \mathcal{A}, \alpha, \beta, \gamma) \propto (n^{a}_{(\backslash s,i),t}+\alpha) \cdot \frac{n^{t}_{(\backslash s,i),m}+\beta}{n^{t}_{(\backslash s,i),\cdot}+M\beta} \\ \cdot \frac{(n^{t_{s,i-1}}_{(\backslash s,i),t_{s,i}}+\gamma)\left\{n^{t_{s,i}}_{(\backslash s,i),t_{s,i+1}}+\delta_{t_{s,i-1},t_{s,i}}\cdot\delta_{t_{s,i},t_{s,i+1}}+\gamma\right\}}{n^{t_{s,i}}_{(\backslash s,i),\cdot}+\delta_{t_{s,i-1},t_{s,i}}+T\gamma} \end{aligned} \tag{4.23}$$

なお，音響ワード$e_{s,i}$が欠損していない場合，それぞれの更新においては音響トピック$z_{s,i}$のみをサンプリングすれば良い。このとき，音響トピックは$z_{s,i}=t$に依存することを考慮すると，事後確率は(4.23)と同じとなる。

#### 音響シーンの事後確率

次に，音響シーンの事後確率$p(a_s|\boldsymbol{e}, \boldsymbol{z}, \boldsymbol{a}_{\backslash s}, \alpha, \beta, \gamma)$を考える。 $p(\boldsymbol{a}_s|\alpha)$がUniform分布に従うこと，および，$p(\boldsymbol{e})$が$\boldsymbol{a}_s$に依存していないことを踏まえると，音響シーンの事後確率は(4.11)と同じとなる。また，音響ワード遷移型教師あり音響トピックモデルおよび音響トピック遷移型教師あり音響トピックモデルにおける，崩壊型ギブスサンプリングのための音響トピック，音響ワード，および，音響シーンの事後確率を表4.2にまとめて記載する。

## 4.5 評価実験

### 4.5.1 実験条件

提案手法による音響シーンの分類性能と欠損した音響ワードの推定性能を評価するため評価実験を行った。本論文では，リビングで頻繁に発生する9種類の音響シーン（「会話をする」「料理をする」「食事をする」「PCを操作する」「新聞を読む」「掃除する」「移動する」「皿を洗う」「TVを見る」）を含む音を収録して実験に用いた。マイクロホンと音源の配置を図4.4に示す。音の収録にはマイクロホンとしてSony ECM-55Bを，マイクロホンアンプとしてGrace design m802を，A/D変換器としてMOTU 24I/Oを用いた。

各音ファイルを収録する際は収録の開始時刻をキューにより提示し，上記の9つの音響シーンのうち1つのみが含まれるようにした。収録後，各音ファイルを16秒毎の11,105の音クリップに分割し，9,802の音クリップをモデルパラメータの学習用に，1,303の音クリップを評価用に用いた。また，その他の実験条件を表4.3に示す。

音響シーン分類性能と欠損した音響ワードの推定性能を評価するため，図4.5の手順による評価を行った。評価手順では，まず事前処理として，入力された音クリップからフレーム毎の音響特徴量を算出した。本実験では，音響特徴量として12次元のメルケプストラム係数 (MFCCs: Mel-frequency cepstral coefficients)を用いた。 MFCCsは音声認識や音声合成などの研究において提案された音響特徴量であるが，音響シーン分類や音響イベント検出においても幅広く用いられているため，本研究においてもMFCCsを採用することとした。その後，GMMを用いたクラスタリングによりMFCCsをクラスタリング

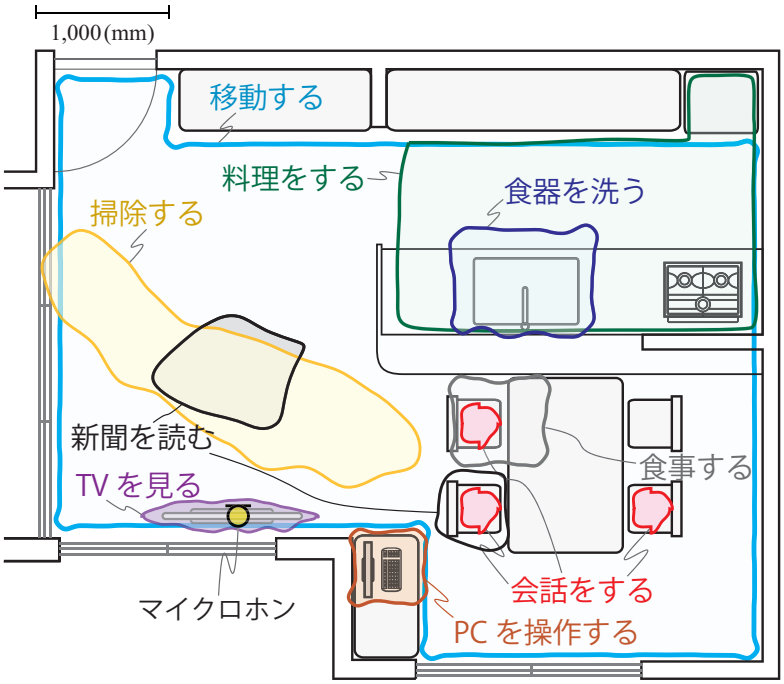



図 4.4: 実環境音を用いた音響シーン分類実験における音源とマイクロホン配置

表 4.3: 4章の実験条件

| | |
|---|---|
| サンプリング周波数 | 16 kHz |
| 量子化ビット数 | 16 bits |
| フレーム長 | 512 |
| 音響特徴量 | MFCCs (12次元) |
| 音響ワード系列の長さ | 1,000 |
| 音響ワードのクラス数 | 256 |
| 音響トピックのクラス数 | 20 |
| ハイパパラメータ $\alpha$ | 3.33 |
| ハイパパラメータ $\beta$ | 0.1 |
| ハイパパラメータ $\gamma$ | 0.5 |

し，音響ワードをモデル化/認識した。音響ワードのモデル化および認識では，GMMによりモデル化された個々のGaussian componentを1つの音響ワードとして定義した。その後，音響ワード系列$\boldsymbol{e}_s$と音響シーンラベル$a$を入力として，音響ワード遷移型教師あり音響トピックモデルおよび音響トピック遷移型教師あり音響トピックモデルを用いて音響シーンの分類と欠損した音響ワ

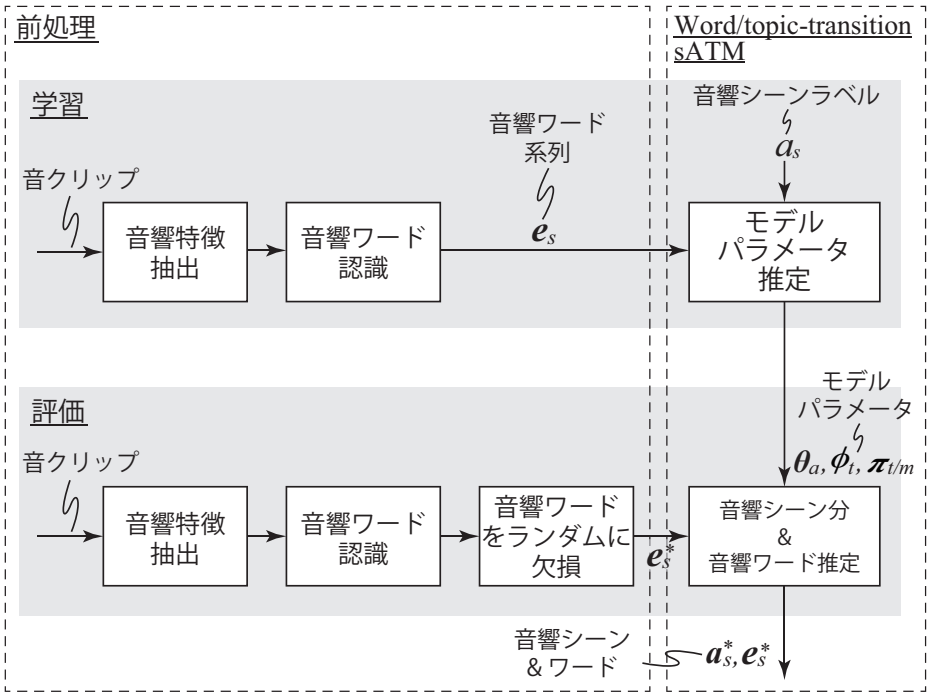



図 4.5: 音響シーン分類と欠損した音響ワードの推定の手順

ードの推定を行う。本実験では，音のクリップの検出やウインドノイズの検出[69]などの欠損検出システムを利用する代わりに，音響ワードを様々な割合でランダムに欠損させて作成した音響ワード系列$\boldsymbol{e}_s^*$を用いて実験を行った。音響シーンの分類と欠損した音響ワードの推定は，学習段階で推定されたパラメータを用いて$a_s$と$e_{s,i}$に対するMAP推定を用いて行った。

### 4.5.2 実験結果

#### 音響シーンの分類性能

提案手法と従来手法に対する，音響ワードの欠損率と音響シーン分類性能の関係を図4.6に示す。なお図4.6には，ランダムに選択された初期値を用いてパラメータ推定と音響シーン分類実験を10回行い得られたF-scoreの平均と分散を示している。また，提案手法との比較として従来の教師あり音響トピックモデルを用いた以下の3つの手法も合わせて示す。

- 観測された音響ワードのみを用いて教師あり音響トピックモデルを適用
- 欠損した音響ワードをランダムな音響ワードで補完して教師あり音響ト

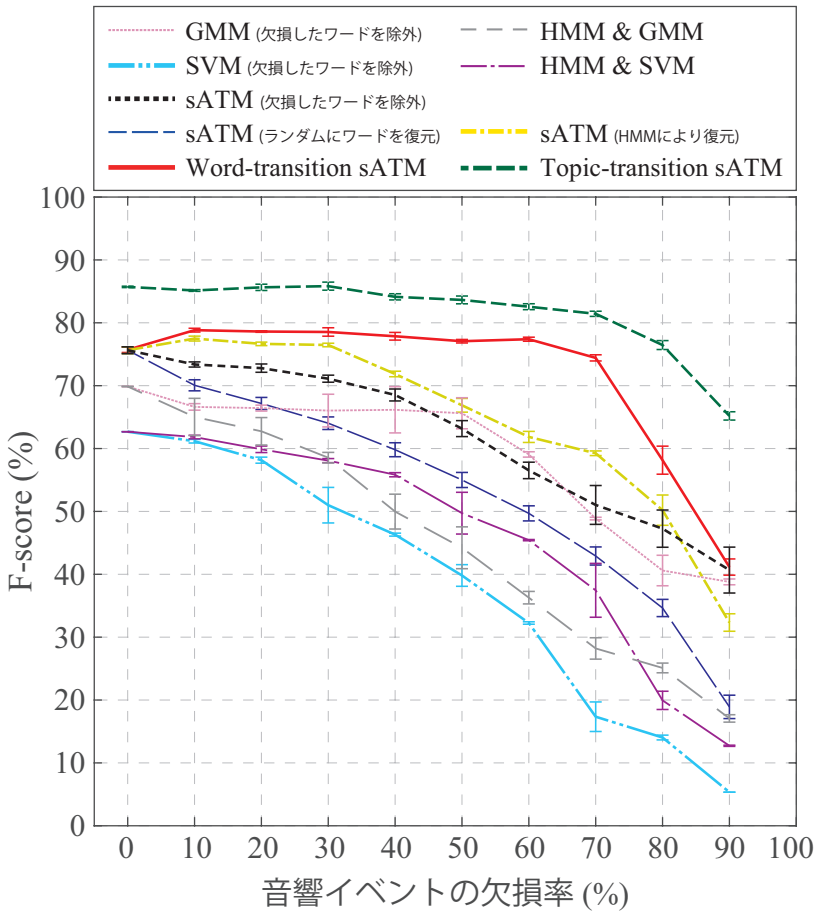



図 4.6: GMM，SVM，教師あり音響トピックモデル，音響ワード遷移型教師あり音響トピックモデル，音響トピック遷移型教師あり音響トピックモデルの各手法による音響シーン分類結果

ピックモデルを適用

- 欠損を有する音響ワード系列に対し事前にHMMを適用して音響ワードを補完し，教師あり音響トピックモデルを適用

実験結果より，音響ワードの欠損率が高くなるにつれて音響シーンに関する情報が失われ，分類性能が低下することが分かる。特に，従来の教師あり音響トピックモデルでは音響ワードの欠損率が50%を超えると音響シーンの分類性能が大幅に低下する。一方で，音響ワード遷移型教師あり音響トピックモデルと音響トピック遷移型教師あり音響トピックモデルでは欠損率が50%を超えた場合においても高い分類性能を保っていることが分かる。特に，音響ト

ピック遷移型教師あり音響トピックモデルでは，音響ワードの欠損率が80%の場合でも平均の音響シーン分類性能が76.5%を達成している。観測に雑音が混入する場合やパケットロスなど，音の観測が欠損する多くの場合において欠損率は50%を超えないと考えられるため，提案手法は欠損を有する観測に対して効果的に音響シーンを分類できると言える。また，プライバシーの観点から部分的な観測しか利用できない場合など，より限られた観測から音響シーンの分析が要求される場合も想定されるため，より高い欠損率の場合における音響シーン分類手法についてはさらに検討が必要である。

また，音響ワード遷移型教師あり音響トピックモデルと音響トピック遷移型教師あり音響トピックモデルを比較すると，音響トピック遷移型教師あり音響トピックモデルの方が7–8ポイント程度高い分類性能を達成している。これは，音響ワード遷移型教師あり音響トピックモデルでは音響ワードの遷移を直接考えているが，実際に観測される音響ワードは確率的にばらつくため，音響トピック遷移型教師あり音響トピックモデルのように潜在的なトピックの遷移を考慮することで確率的なばらつきをモデル化できるためと考えられる。これは，欠損率が0%の場合の音響シーン分類結果を比較することからも分かる。

### 欠損した音響ワードの推定性能

図4.7に欠損した音響ワードの平均推定精度を示す。なお，本実験は音響シーン分類実験と同様の条件で実施した。実験結果より，音響ワード遷移型教師あり音響トピックモデルと音響トピック遷移型教師あり音響トピックモデルによる音響ワードの推定精度は，ランダムに音響ワードを補完した場合と比較して良いことが分かる。また，図4.6と図4.7より音響ワードの欠損率が高く，かつ音響ワードの推定精度が高くない場合においても音響シーンの分類性能は大幅には低下していないことが分かる。このことより，提案手法ではたとえ欠損した音響ワードが正しく推定できていない場合においても，音響シーンと関連が深い音響ワードが推定されていると推測される。

推定された音響ワードがどのように分布しているかを確認するため，図4.8と図4.9に各音響クリップに含まれる音響ワードのヒストグラムを算出した。図4.8と図4.9では，欠損していない音響ワード系列の音響ワードヒストグラム，欠損を有する音響ワード系列の音響ワードヒストグラム，欠損した音

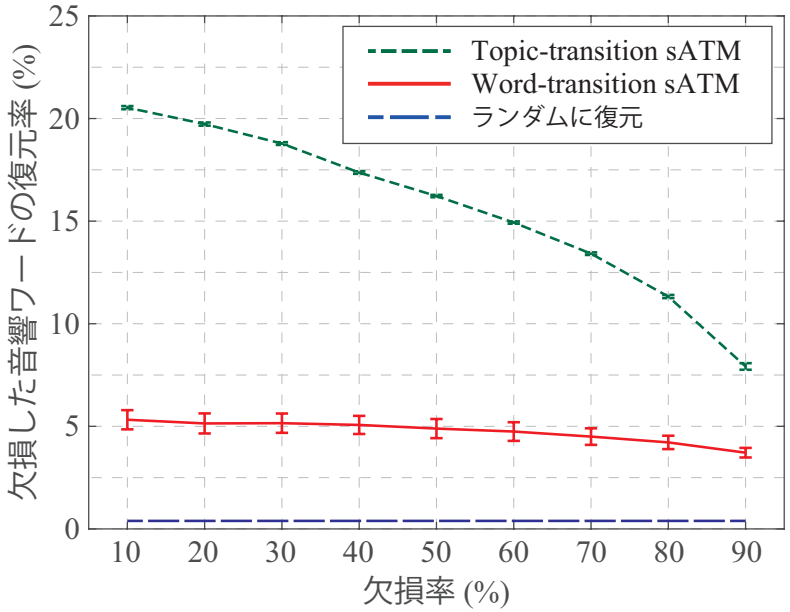



図 4.7: 音響ワード遷移型教師あり音響トピックモデル，音響トピック遷移型教師あり音響トピックモデルのそれぞれにより欠損した音響ワードを復元した場合の音響ワードの復元率

響ワードをHMMにより補完した音響ワード系列の音響ワードヒストグラム，音響ワード遷移型教師あり音響トピックモデルにより音響ワードを補完した音響ワード系列の音響ワードヒストグラム，音響トピック遷移型教師あり音響トピックモデルにより音響ワードを補完した音響ワード系列の音響ワードヒストグラムをそれぞれ示している。結果より，音響ワードの欠損率が増加すると音響ワードヒストグラムの詳細なパターンが失われることが分かる。特に音響ワードの欠損率が80%の場合では，多くの音響ワードが識別困難になる。一方で，音響ワード遷移型教師あり音響トピックモデルと音響トピック遷移型教師あり音響トピックモデルにより音響ワードが補完された音響ワードヒストグラムでは，ヒストグラムのパターンが復元されている様子が見て取れる。特に音響トピック遷移型教師あり音響トピックモデルを適用した場合は音響ワードの欠損率が80%においても，正確に音響ワードヒストグラムのパターンを復元できている。これらの結果からも，提案手法では欠損した音響ワードが正しく推定できていない場合においても，音響シーンと関連が深い音響ワード系列が再構成されていることが分かる。

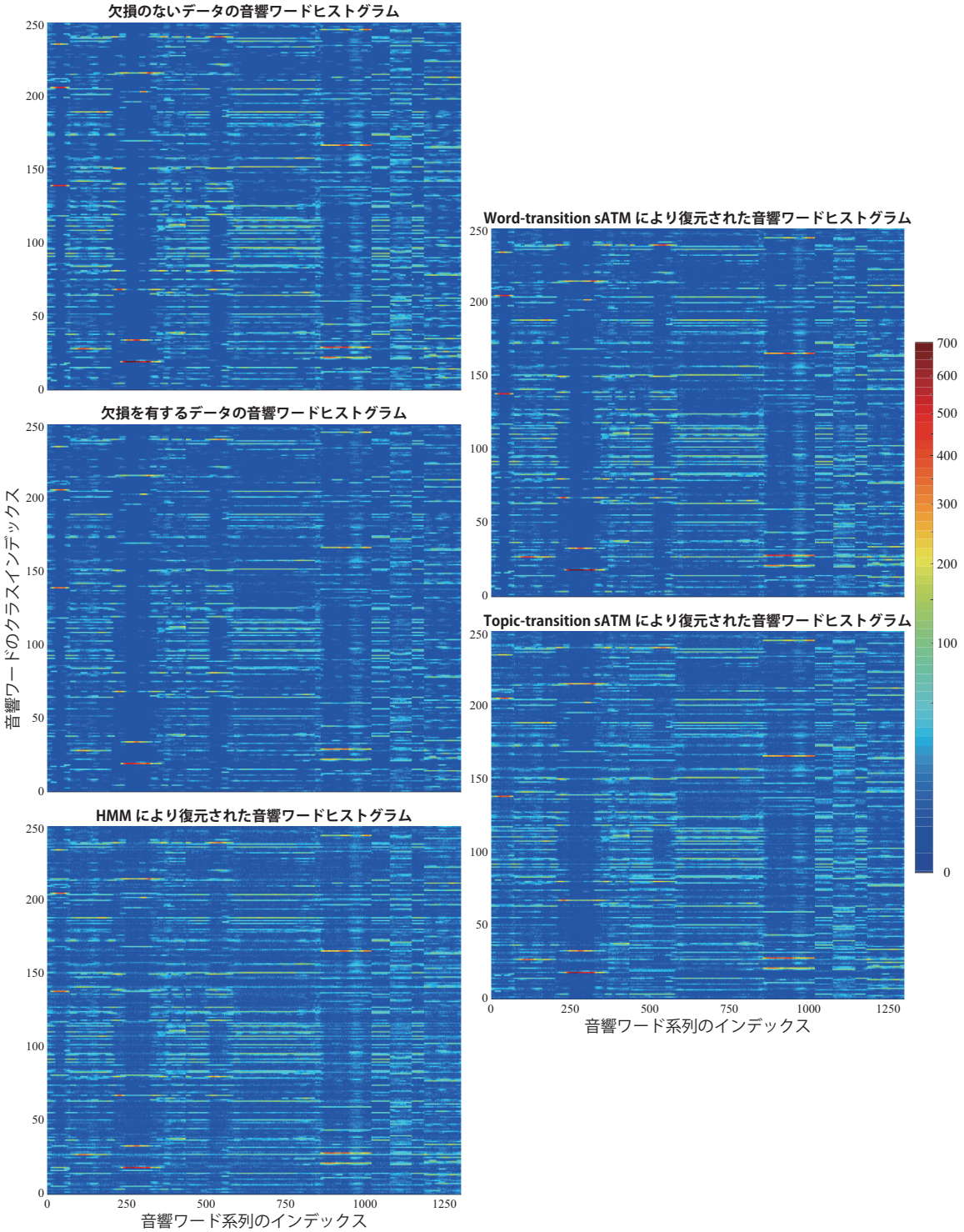



図 4.8: 音響ワード遷移型教師あり音響トピックモデルおよび音響トピック遷移型教師あり音響トピックモデルにより復元された音響ワード系列における音響ワードヒストグラム (欠損率40%)

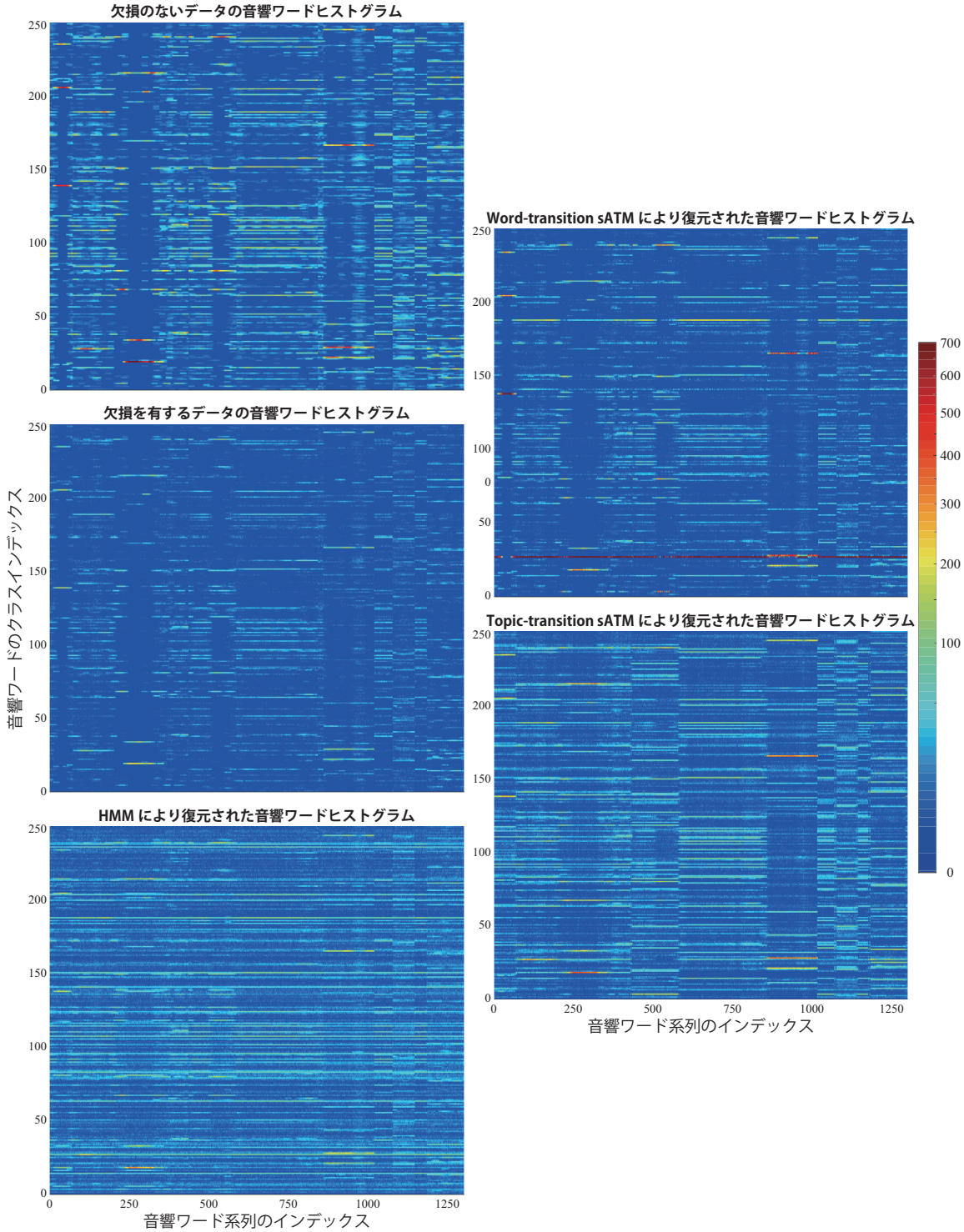



図 4.9: 音響ワード遷移型教師あり音響トピックモデルおよび音響トピック遷移型教師あり音響トピックモデルにより復元された音響ワード系列における音響ワードヒストグラム (欠損率80%)

## 4.6 4章のまとめ

欠損を有する音響ワード系列から音響シーンを分類しつつ，欠損した音響ワードを推定するため，本章では音響トピックモデルに基づく新たな音響ワード系列の生成モデルを提案した。具体的には，1) 音響ワードの時間遷移関係を単純マルコフモデルを用いてモデル化し，教師あり音響トピックモデルに組み込んだ，音響ワード遷移型教師あり音響トピックモデルと，2) 音響トピックの時間遷移関係を同様にマルコフモデルを用いてモデル化し，教師あり音響トピックモデルに組み込んだ，音響トピック遷移型教師あり音響トピックモデルを提案し，崩壊型ギブスサンプリングによるパラメータ推定方法を導出した。性能評価実験により，提案手法では音響ワードの欠損率が50%程度になった場合でも，欠損がない場合と同程度の音響シーン分類性能を実現することを示した。また，提案モデルでは，音響シーンと関連が深い音響ワードを正確に復元できることも明らかになった。

# 5 同期誤差を有する多チャネル観測に基づく音響シーン分類

## 5.1 はじめに

テレビのニュース番組から流れてくる音声とリビングで机を囲みながら会話をしている音声，両者は音の周波数情報という観点では非常に良く似ているが，音響シーンとしては大きな違いを持つ。一方，音の発生場所は2つの音響シーンで異なっているため，これらの音響シーンを分類するためには空間情報が有効となる。さらに，空間情報と周波数情報や時間-周波数情報を合わせれば，より正確な音響シーン分類が期待できる。

表5.1に示すように，音響シーン分類や音響イベント検出において空間情報を用いた手法はこれまでにも複数提案されている。例えば， Adavanneらは RNN-LSTMを用いてステレオチャネルの信号から音響シーンを分類する手法を提案しており，モノラルチャネル信号にRNN-LSTMを適用した場合よりも分類性能が向上することを示している[70]。 Giannoulisらは屋内の複数の部屋

表 5.1: 空間情報を用いた音響シーン分類の従来研究

| | | チャネル数 | |
|---|---|---|---|
| | | ステレオチャネル | 多チャネル（3チャネル以上） |
| マイク位置/時間同期 | 既知/同期あり | RNN-LSTMに基づく手法 [70] | MFCCs & GMM + SVMによる部屋分類手法 [71] |
| | 未知/同期なし | – | チャネル毎にシーン分類し結果を尤度に基づき統合 [22] |

で得られた多チャネル信号から，音が発生した部屋を分類する手法を提案している[71]。文献[71]では，多チャネル信号からMFCCsとGMMを用いた音声区間検出と，部屋の内外を分類する複数のSVMを組み合わせて音が発生した部屋を分類する手法を提案しており，手法の有効性を実環境収録音を用いて評価している。また，Phanらはランダムフォレストを多チャネル信号に適用し，音響イベントの分類と区間検出を行う手法を提案している[72]。 Kürbyらはマイクロホンの位置が未知の場合にも適用可能であり，かつ，マイクロホンが厳密に時間同期されていない場合にも利用可能な音響シーン分類手法を提案している[22]。文献[22]では，多チャンネルで得られた音響信号を事前に各チャネルごとに音響ワード系列で表現し，GMMを用いて各チャネルの音響シーン分類を行った後，結果をチャネル統合する手法を提案している。

文献[71, 70]では，多チャネル信号が正確に時間同期されていることを前提としている。しかしながら，スマートホンやIoT機器などにより構成されたマイクロホンアレイは正確に同期を取ることが容易でないという課題や，事前にマイクロホンの位置や音源の位置を知ることが難しいという課題があり，従来のマイクロホンアレイ処理により空間情報を直接取得することは出来ない。これらの課題に対処するため，Onoら[73]やHasegawaら[74]は時間同期誤差を持つマイクロホンのブラインドアライメント手法を提案している。また，Liu [75]やSchmalenstroeerら[76]，Miyabeら[77]はクロック同期誤差を持つマイクロホンのための同期補償手法を提案しており，同期補償を行った後に従来のマイクロホンアレイ技術を適用することで，クロックずれしたままの信号を用いてマイクロホンアレイ処理を適用した場合よりも，音源分離性能が向上することを示した。また，Raykarら [78]やHasegawaら[74]は音源位置やマイクロホン位置を収録音から推定し，マイクロホンアレイ処理を適用する手法

を提案している。しかしながら，これらの手法によりマイクロホン間の時間同期やクロック同期を行ったり，音源位置やマイクロホンの位置を推定するには大きな計算量が必要という課題が残る。また，背景雑音や音の反射，障害物の影響により十分な性能が得られない場合も多いため，音源やマイクロホンの位置情報を利用せず，また，時間同期誤差に対して頑健に空間情報を取得する手法の実現は依然重要である。一方，文献[22]では音源やマイクロホンの位置情報を必要としない手法を提案しているものの，多チャンネルで得られた音響信号から各チャネルごとに音響シーン分類し，その結果を尤度等に基づき統合しているため，空間情報を十分に利用できているとは言い難い。また，各チャネルで得られた振幅情報から音の発生の有無を判断し，バイナリ情報として音響シーン分類に利用する手法も考えられるが，依然として空間情報を十分に利用できているとは言い難い。

本章ではこれらの課題を解決可能な，空間情報に基づく音響シーン分類の実現を目指す。具体的には，音源やマイクロホンの位置情報を必要とせず，マイクロホン間の時間同期誤差に頑健な空間ケプストラムという特徴抽出法を提案する。空間ケプストラムは，マイクロホン間の時間同期誤差に敏感な位相情報を用いず，時間同期誤差により頑健な振幅情報のみ活用し，周波数特徴量であるケプストラムと同様の特徴量変換を行う手法である。本章では，空間ケプストラムにより空間情報を抽出することで，効果的で効率的な音響シーン分類が実現できることを示す。

次節以降の構成は以下のとおりである．まず，5.2節では空間情報を用いた音響トピックモデルについて述べ，同期誤差を含む多チャネル観測についても他の観測条件と同様の枠組みで扱うための方法について議論する。次に5.3節では，周波数特徴量として音声認識等で幅広く利用されているケプストラムについて述べる。 5.4節では，多チャネル観測から空間特徴量を抽出するための手法として，空間ケプストラムと呼ばれる特徴量について述べ，さらに，空間ケプストラムが特定の条件下でケプストラムと一致することを示す。 5.5節では，空間ケプストラムと同様の直交変換を周波数-空間対数振幅行列を連結したベクトルに対して適用することで定義される，一般化周波数-空間ケプストラムについて述べる。 5.6において，空間ケプストラムにより空間情報がどのように抽出されるかを確認するシミュレーション実験を行う。また，空間ケプストラムが定常的な拡散性雑音に対してどの程度頑健かを評価するシミュレーション実験結果についても示す。 5.7節では，空間ケプストラ

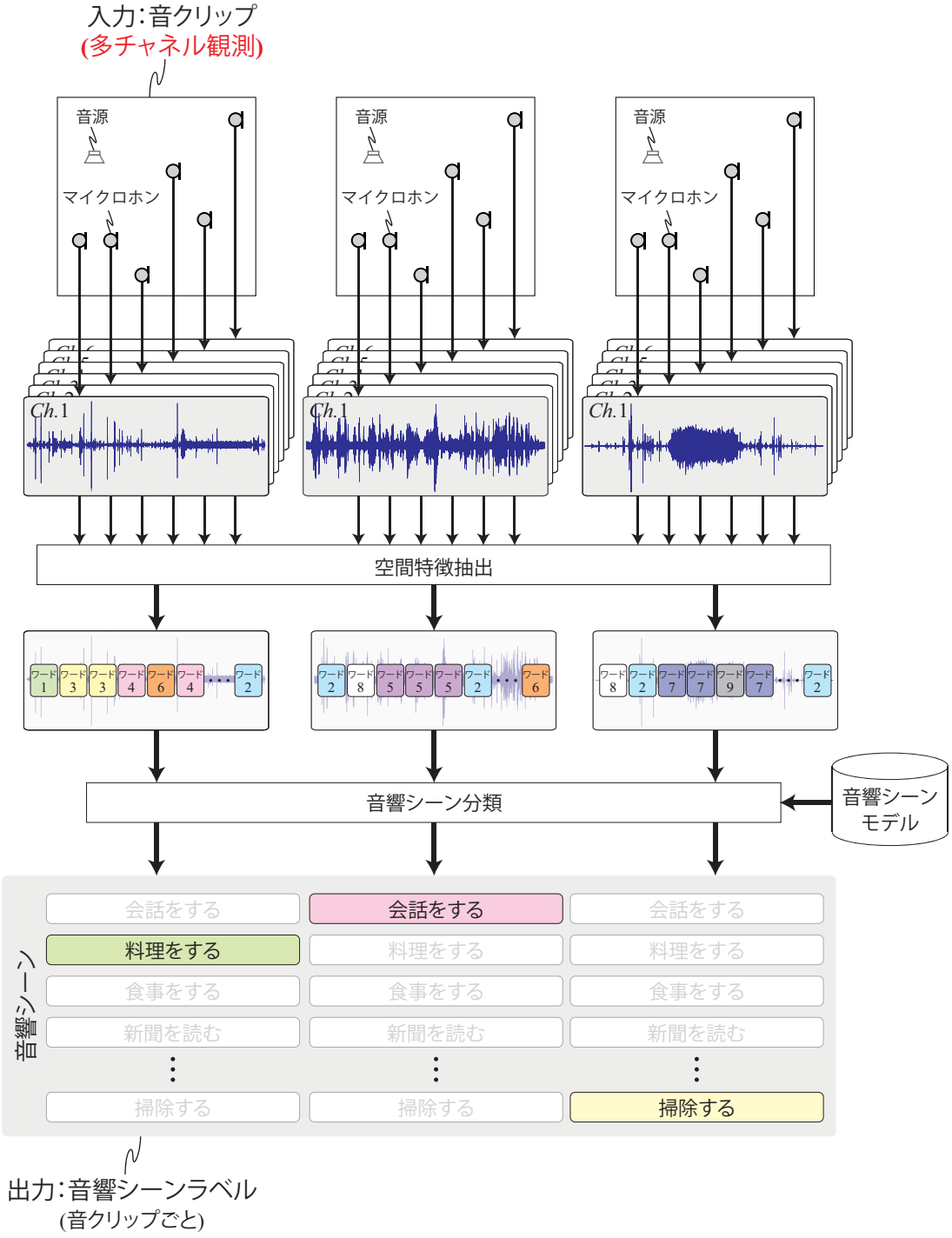



図 5.1: 空間特徴量に基づく音響トピックモデルの実現例1

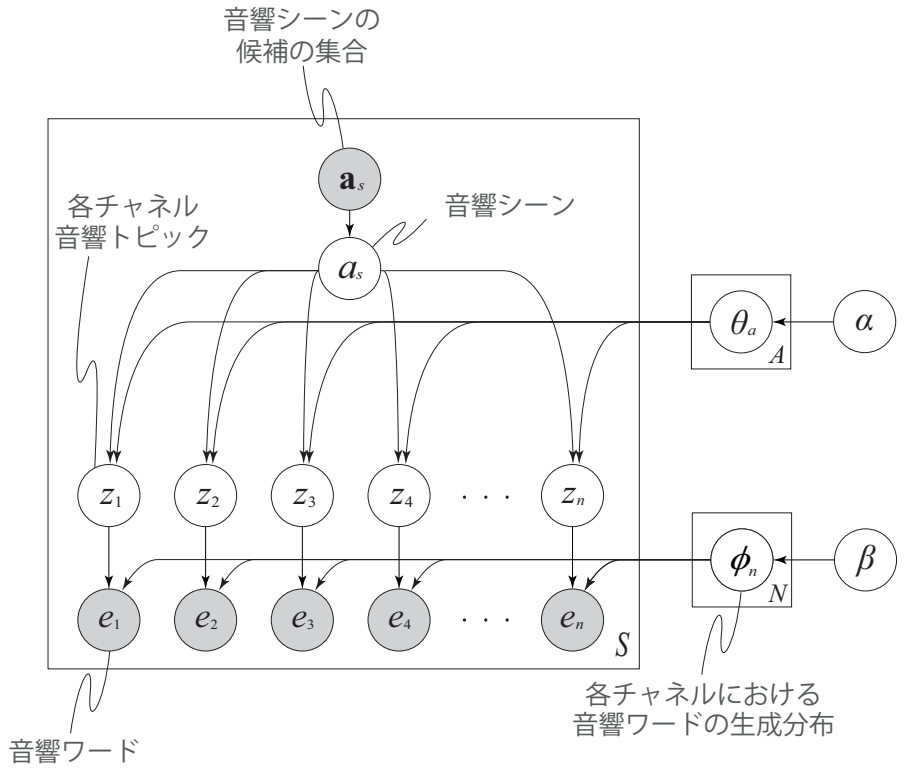



図 5.2: 多チャネルの音響ワードの生成を考慮した教師あり音響トピックモデルのグラフィカルモデル。チャネル数を$N$とする。

ムの音響シーン分類問題への有効性について，実環境収録音を用いた性能評価実験を行った結果について述べ，5.8節において，マイクロホンのチャネル間同期誤差に対する空間ケプストラムの頑健性について評価実験を行った結果について述べる。最後に5.9節で本章のまとめを述べる。

## 5.2 空間情報に基づく音響トピックモデル

音響シーン分類ではそれぞれの音響シーンに対して非常に幅広い音を含むデータを大量に取得する必要があるが，モデルの学習に十分な量のデータを収集することは容易ではないため，モデルの過剰適合がシーン分類性能低下の大きな要因となる。そのため，多チャネルのマイクロホンにより取得された音情報に対しても音響トピックモデルを適用できれば，空間情報を利用した音響シーン分類に有効な手法となり得る。また，音響トピックモデルの利点

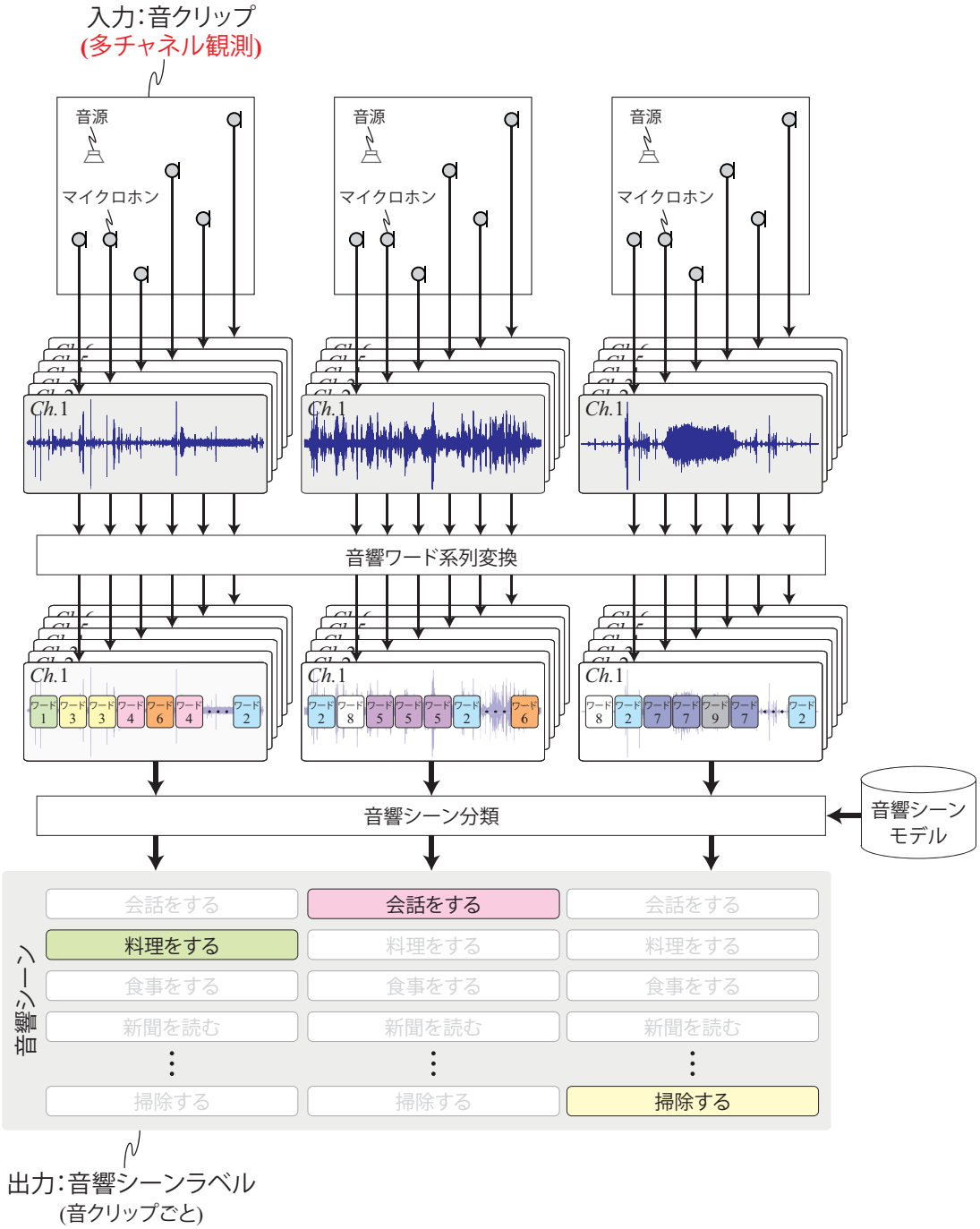



図 5.3: 空間特徴量に基づく音響トピックモデルの実現例2

の1つとして，音響ワードというラベル系列を入力とするため，観測された音クリップをラベル系列に変換できれば，その間の処理に依らず適用可能である点が挙げられる。つまり，空間特徴量を音響ワードラベル系列に変換すれば，音響トピックモデルの枠組みはそのままで空間情報を組み込むことができる。そこで本節では，空間情報を用いた音響トピックモデルの実現方法について検討を行う。なお，本論文では実現方法について議論を行うが，実際の評価実験については今後の課題とする。

多チャネルの観測に対して音響トピックを適用するための最も簡単な例として，Kürbyら[22]が提案した手法を生成モデルに拡張する方法が挙げられる。具体的には，多チャンネルで得られた音響信号を事前に各チャネルごとに音響ワード系列で表現し，各チャネルごとに音響トピックモデルによる音響シーン分類を行った後に，最も尤度の高い分類結果最終的な音響シーン分類結果とする方法である。しかしながら，この手法では最も尤度の高いチャネルの結果を選択しているに過ぎず，空間情報を十分に活用しているとは言い難い。

次に，空間特徴量を用いた音響トピックモデルによる音響シーン分類の例を図5.1に示す。図5.1に示す音響シーン分類では，任意の方法により抽出された空間特徴量を音響ワード系列に変換し，2章に示した流れにより音響シーン分類を行う。これは，従来の音響トピックモデルにおける音響シーン分類において，MFCCsなどの周波数特徴抽出を行う代わりに空間特徴抽出を行うものであり，従来の音響トピックモデルの枠組みを活用した上で空間情報を音響トピックモデルに組み込むことが可能となる。つまり，音響トピックモデルと，マイクロホンの位置情報を必要とせず時間同期誤差に頑健な空間特徴抽出法とを組み合わせることにより，両者の利点を容易に組み合わせることが可能となる。

また，空間情報を用いた音響トピックモデルの異なる実現方法として，図5.2，図5.3に示すように，多チャネルの音響ワード系列の生成過程をモデルし，そのパラメータを活用して音響シーンの分類を行う方法も考えられる。多チャネルの音響ワード系列の生成過程をモデル化する手法では，空間特徴の生成過程もモデルに組み込むことで一貫した音響シーンのモデル化と分類が可能になるが，チャネル数に応じてパラメータが線形に増加するため，計算コストや過学習が大きな問題となる。また，マイクロホン間の時間同期誤差は考慮されていないため，時間同期誤差がある観測に対して音響シーン分

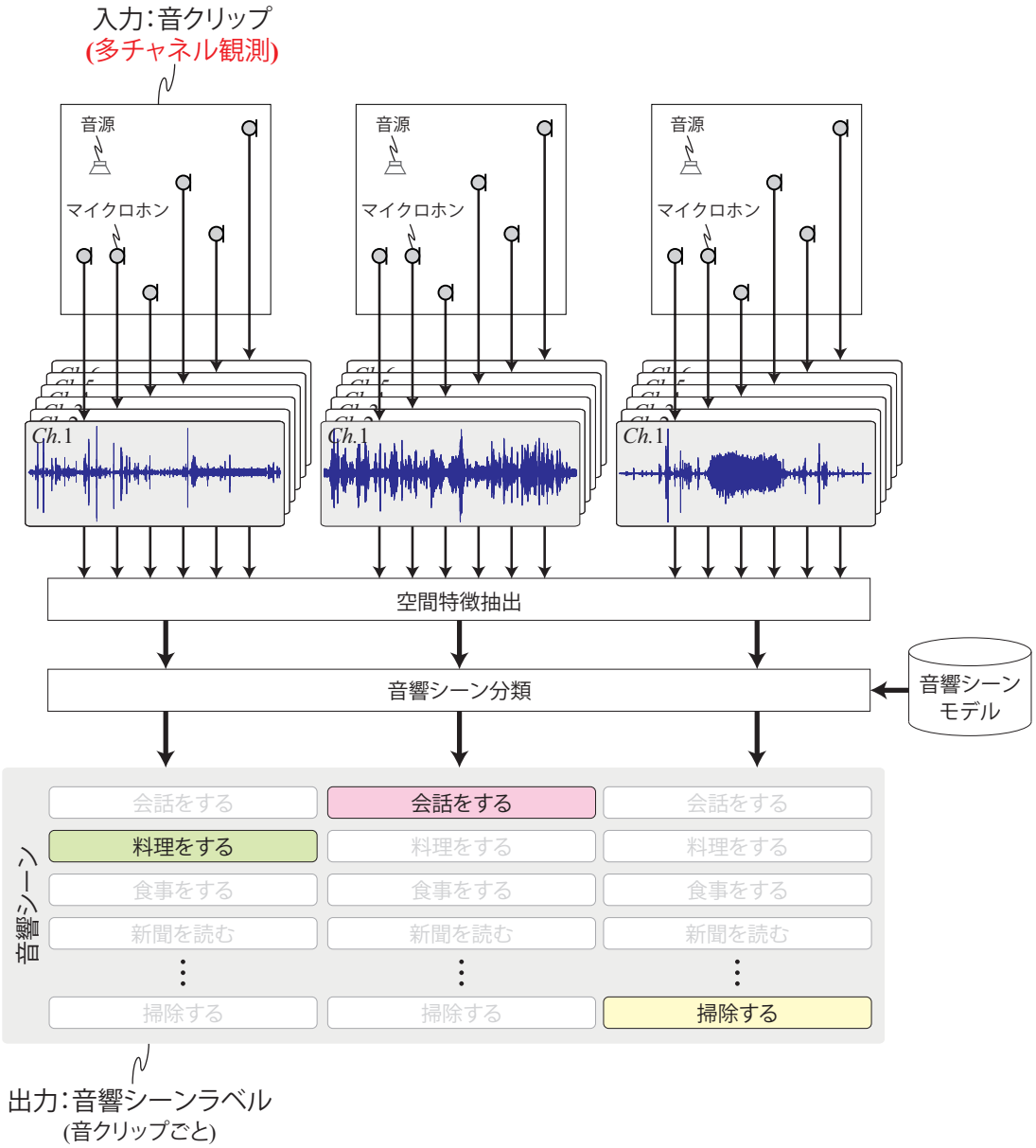



図 5.4: 5章で扱う音響シーン分類問題

類性能が低下する可能性もある。

本論文では，任意の空間特徴抽出方法と組み合わせることが可能であり，かつ，従来の音響トピックモデルの枠組みを活用可能であるという観点から，図5.1に示す音響シーン分類手法の予備検討を行う。ただし，空間情報を用いた音響トピックモデルの詳細な議論や性能評価は今後の検討とし，以降では，

音源やマイクロホンの位置情報やマイクロホン間の時間同期誤差の課題を解決可能な空間特徴抽出法について議論を行う。また，以下では議論を簡単にするため，図5.4に示すような音響シーン分類問題により空間特徴抽出法の評価を行う。

## 5.3 ケプストラム

5.4節で提案する空間ケプストラムとの比較のため，本節では周波数特徴量として音声認識や音声合成などで幅広く用いられているケプストラムについて述べる。

配置が固定された$N$個のマイクロホンを用いて音響信号を収録する場面を想定する。$\omega$を周波数のインデックス，$\tau$を時間フレームのインデックス，$n$をチャネルインデックスとし，$s_{\omega,\tau,n}$を多チャネル信号の観測の短時間フーリエ変換（STFT: Short-time Fourier transform）とする。ここでケプストラムでは，信号の振幅情報$a_{\omega,\tau,n} = |s_{\omega,\tau,n}|$に着目し，周波数分割された対数振幅ベクトル$\mathbf{p}_\tau$を利用する。

$$\mathbf{p}_\tau = \begin{pmatrix} \log \bar{a}_{1,\tau} \\ \log \bar{a}_{2,\tau} \\ \vdots \\ \log \bar{a}_{\omega,\tau} \\ \vdots \\ \log \bar{a}_{\Omega,\tau} \end{pmatrix} \tag{5.1}$$

ここで，$\Omega$は周波数ビンの数とし，

$$\bar{a}_{\omega,\tau} = \sqrt{\frac{1}{N}\sum_n a^2_{\omega,\tau,n}} \tag{5.2}$$

を全チャネルにわたる振幅スペクトルの二乗平均平方根 (RMS: Root mean square)とする。対数振幅ベクトル$\mathbf{p}_\tau$に逆離散フーリエ変換 (IDFT: Inverse

discrete Fourier transform)を適用し，ケプストラムは以下のように定義される。

$$\mathbf{c}_\tau = \mathbf{Z}_\Omega \mathbf{p}_\tau \tag{5.3}$$

ただし，$\mathbf{Z}_\Omega$は$\Omega \times \Omega$の逆離散フーリエ変換行列とする。また，ケプストラムと同様の周波数特徴量として，メル周波数表現されたスペクトルに対して離散コサイン変換 (DCT: Discrete cosine transform)を用いて算出されるメル周波数ケプストラム係数 (MFCCs: Mel-frequency cepstrum coefficients)も音声認識等で幅広く用いられている。

## 5.4 空間ケプストラム

### 5.4.1 空間ケプストラムの定義と期待される効用

周波数領域におけるケプストラムの考え方を空間領域での特徴量抽出に適用することで空間ケプストラム (SC: Spatial cepstrum)が定義可能である。分散マイクロホンアレイではチャネル間同期が非常に大きな問題であり，正確に同期が取れたマイクロホンアレイを用いることができない場合，サンプリング周波数の同期ずれに敏感な位相情報は信頼性に乏しい。そこで空間ケプストラムではケプストラム同様，サンプリング周波数の同期ずれにより頑健な信号の振幅情報$a_{\omega,\tau,n} = |s_{\omega,\tau,n}|$にのみ着目する。

ケプストラム同様，配置が固定された$N$個のマイクロホンを用いて音響信号を収録する場面を想定する。時間フレーム$\tau$において各マイクロホンで得られた対数振幅ベクトル$\mathbf{q}_\tau$は以下のように表すことができる。

$$\mathbf{q}_\tau = \begin{pmatrix} \log \tilde{a}_{\tau,1} \\ \log \tilde{a}_{\tau,2} \\ \vdots \\ \log \tilde{a}_{\tau,n} \\ \vdots \\ \log \tilde{a}_{\tau,N} \end{pmatrix} \tag{5.4}$$

ここで，

$$\tilde{a}_{\tau,n} = \sqrt{\frac{1}{\Omega}\sum_{\omega} a_{\omega,\tau,n}^2} \tag{5.5}$$

は各時間フレーム，各チャネルにおけるパワーである。

周波数領域のケプストラム算出において，各サブバンドの振幅$\bar{a}_{\omega,\tau}$は周波数軸上やメル周波数軸上で均等な間隔で定義されているため，逆離散フーリエ変換や離散コサイン変換による基底変換が可能である。一方，空間領域の場合はマイクロホンが均等な間隔で配置されるとは限らないため，逆離散フーリエ変換や離散コサイン変換での基底変換を適用することができない。そこで，空間ケプストラムでは基底変換として主成分分析 (PCA: Principal component analysis)を適用する。まず，$\mathbf{R}_q$を$\mathbf{q}_\tau$の共分散行列として以下のように定義する。

$$\mathbf{R}_q = \frac{1}{T}\sum_{\tau} \mathbf{q}_\tau \mathbf{q}_\tau^{\mathsf{T}} \tag{5.6}$$

ただし，$T$は時間フレームの総数，$^{\mathsf{T}}$は転置を表すものとする。共分散行列$\mathbf{R}_q$が対称行列であることを踏まえると，その固有値分解は以下で表現可能である。

$$\mathbf{R}_q = \mathbf{E}\mathbf{D}\mathbf{E}^{\mathsf{T}} \tag{5.7}$$

ただし，$\mathbf{E}$，$\mathbf{D}$はそれぞれ固有ベクトルを連結した行列と，固有値を対角成分に持つ対角行列を表す。また，$\mathbf{D}$は値の大きいものから降順に固有値が並んでいるものとする。行列$\mathbf{E}$を用いることで空間ケプストラム$\mathbf{d}_\tau$は以下のように計算される。

$$\mathbf{d}_\tau = \mathbf{E}^{\mathsf{T}}\mathbf{q}_\tau \tag{5.8}$$

空間ケプストラムではPCAを用いて$\mathbf{d}_\tau$の各要素を無相関化していることより，固有値の小さい要素に対応する次元を削減することで効率的に次元圧縮することも可能である。

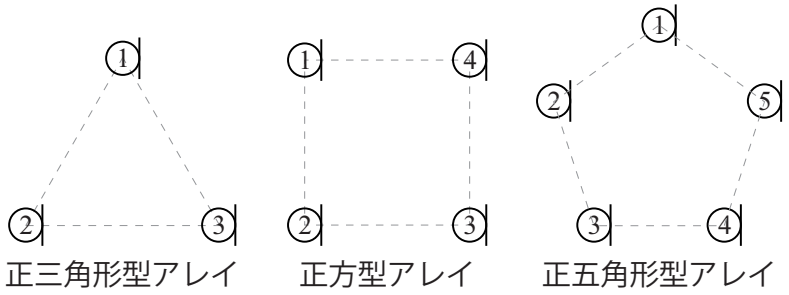



図 5.5: 円対称なマイクロホン配置の例

空間ケプストラムの算出にはマイクロホンの位置情報が不要であるため，分散配置されたマイクロホンに対するアレイ信号処理に適した手法と言える。また，周波数領域におけるケプストラムとの類似性を考慮すると，ケプストラム平均正規化 (CMN: Cepstral mean normalization) [79]やケプストラム分散正規化 (CVN: Cepstral variance normalization) [80]などの手法がそのまま適用でき，これは空間ケプストラムを利用する際に大きな利点となる。

### 5.4.2 等方性音場での円対称アレイにおける空間ケプストラム

主成分分析における白色化行列は観測信号の共分散行列の固有値分解によって決定されるため，一般的に$\mathbf{E}$は観測信号毎に可変の行列となる。しかしながら，マイクロホンの配置が円対称になりかつ音場が等方性を有するとき，つまり，1) 音響パワーが位置に依らず同一であり，かつ，2) 2つのマイクロホンの観測の相互相関がマイクロホン位置のなす角度に依存しない場合，白色化行列は一意に決定することができる。例えば，円対称なマイクロホン配置として，図5.5に示すような正多角形が考えられる。以下では正$N$角形を例にとって等方性音場に配置された円対称アレイにおける空間ケプストラムについて議論を行う。

図5.5に示すように，マイクロホンのインデックス$n$は環状に与えられているものとする。 $\alpha_1$をパワースペクトルとし，$\alpha_i$ $(i \neq 1)$を2つのマイクロホンの観測信号のクロススペクトルとすると，共分散行列$\mathbf{R}_q$は以下のように与えられる。

$$\mathbf{R}_q = \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_n & \cdots & \alpha_{N-1} & \alpha_N \\ \alpha_2 & \alpha_1 & \cdots & \alpha_{n-1} & \cdots & \alpha_{N-2} & \alpha_{N-1} \\ \vdots & \vdots & \ddots & \vdots & & \vdots & \vdots \\ \alpha_n & \alpha_{n-1} & \cdots & \alpha_1 & \cdots & \alpha_{N-n} & \alpha_{N-n+1} \\ \vdots & \vdots & & \vdots & \ddots & \vdots & \vdots \\ \alpha_{N-1} & \alpha_{N-2} & \cdots & \alpha_{N-n} & \cdots & \alpha_1 & \alpha_2 \\ \alpha_N & \alpha_{N-1} & \cdots & \alpha_{N-n+1} & \cdots & \alpha_2 & \alpha_1 \end{bmatrix} \tag{5.9}$$

ここで，

$$\alpha_n = \alpha_{N-n+2} \quad (n > \frac{N}{2} + 1) \tag{5.10}$$

音場の等方性の仮定から，観測信号のクロススペクトルはマイクロホン間の距離のみに従って決定されることを踏まえると，マイクロホンの配置が円対称となるとき，共分散行列$\mathbf{R}_q$は循環行列になる[81, 82]。さらに，循環行列は以下の逆離散フーリエ変換行列$\mathbf{Z}_N$によって対角化されることが知られている[83]。

$$\mathbf{Z}_N = \frac{1}{\sqrt{N}} \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 & 1 \\ 1 & \zeta^1 & \zeta^2 & \cdots & \zeta^{N-2} & \zeta^{N-1} \\ 1 & \zeta^2 & \zeta^4 & \cdots & \zeta^{2(N-2)} & \zeta^{2(N-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & \zeta^{N-2} & \zeta^{2(N-2)} & \cdots & \zeta^{(N-2)^2} & \zeta^{(N-1)(N-2)} \\ 1 & \zeta^{N-1} & \zeta^{2(N-1)} & \cdots & \zeta^{(N-2)(N-1)} & \zeta^{(N-1)^2} \end{bmatrix} \tag{5.11}$$

$$\zeta = e^{j2\pi/N} \tag{5.12}$$

以上より，マイクロホンの配置が円対称になりかつ音場が等方性を有する場合，主成分分析は逆離散フーリエ変換による直行化と一致し，(5.8)はケプストラムの定義と厳密に一致する事が分かる。特定の条件下でのみこの関係は成り立つが，本論文では$\mathbf{d}_\tau$を空間ケプストラムと称する。

## 5.5 一般化周波数-空間ケプストラム

ケプストラムと空間ケプストラムはともに対数振幅ベクトルの線形直交変換により算出される事を踏まえ，空間ケプストラムと同様の直交変換を，周波数-空間対数振幅行列を列ごとに連結したベクトルに適用することで，一般化周波数-空間ケプストラムを定義する。

まず，以下のような周波数-空間対数振幅行列を考える。

$$\mathrm{S}_\tau = [\ \mathbf{p}_{\tau,1}\ \ \cdots\ \ \mathbf{p}_{\tau,n}\ \ \cdots\ \ \mathbf{p}_{\tau,N}\ ] \tag{5.13}$$

ただし，

$$\mathbf{p}_{\tau,n} = \begin{pmatrix} \log a_{1,\tau,n} \\ \log a_{2,\tau,n} \\ \vdots \\ \log a_{\omega,\tau,n} \\ \vdots \\ \log a_{\Omega,\tau,n} \end{pmatrix} \tag{5.14}$$

ここで，時間フレーム$\tau$における周波数-空間特徴量をベクトル表現するため，(5.13)で表される周波数-空間対数振幅行列を直列に結合し，$(\Omega N) \times 1$次元の特徴量ベクトルとして以下のように表現する。

$$\mathbf{s}_\tau = \begin{pmatrix} \mathbf{p}_{\tau,1} \\ \mathbf{p}_{\tau,2} \\ \vdots \\ \mathbf{p}_{\tau,n} \\ \vdots \\ \mathbf{p}_{\tau,N} \end{pmatrix} \tag{5.15}$$

ベクトル表現された$\mathbf{s}_\tau$に対し，(5.7)と(5.8)で示される主成分分析を行うことで以下の特徴量が得られる。

$$\mathbf{g}_\tau = \mathbf{E}^{\mathsf{T}}\mathbf{s}_\tau \tag{5.16}$$

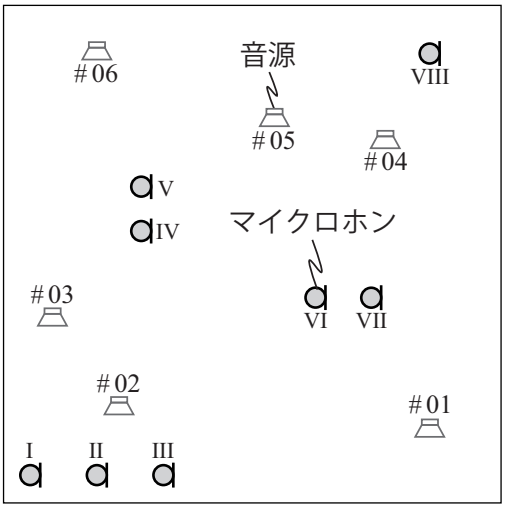



図 5.6: シミュレーション実験における音源とマイクロホンの配置

本稿では，$\mathbf{g}_\tau$を一般化周波数-空間ケプストラム (GFSC: Generalized frequency-spatial cepstrum)と呼ぶ。

## 5.6 空間ケプストラムによる空間パターン表現

空間ケプストラムが特徴量空間上で空間情報をどのように表すのかを評価すること，また，空間ケプストラムの拡散性雑音への頑健性を評価することを目的としてシミュレーション実験を行った。図5.6にシミュレーション実験におけるマイクロホンとスピーカの配置を示す。本実験では，自由空間内にマイクロホンとスピーカが配置されているものと仮定し，スピーカからは球面波が伝搬するものとする。また，スピーカからは固定長の定常的な白色ガウス雑音を順番に駆動し，それぞれのマイクロホンでの観測の理論値を算出し，その後，変換行列$\mathbf{E}^\mathsf{T}$を用いて空間ケプストラムを計算した。加えて本実験では，以下で計算されるCMNを適用した。

$$\mathbf{d}_\tau = \mathbf{E}^\mathsf{T}(\mathbf{q}_\tau - \bar{\mathbf{q}}_\tau) \tag{5.17}$$

ここで，$\bar{\mathbf{q}}_\tau$は$\mathbf{q}_\tau$の時間平均を表す。空間ケプストラムによって空間情報がどのように表現されるかを，図5.7に示す変換行列$\mathbf{E}^\mathsf{T}$によって確認した。変換行

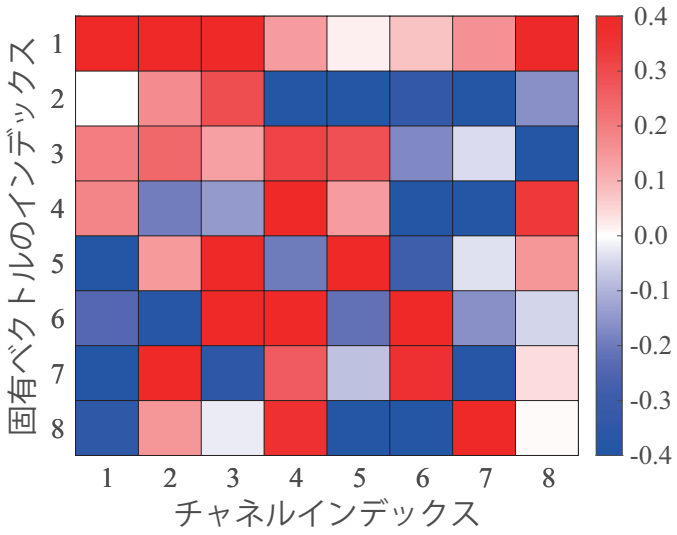
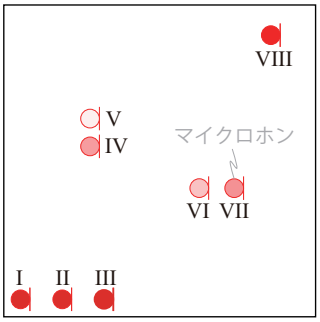



図 5.7: シミュレーション実験条件における変換行列$\mathbf{E}^{\mathsf{T}}$のカラーマップ表現

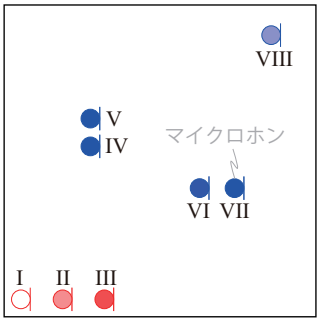
列$\mathbf{E}^{\mathsf{T}}$の$k$番目の行ベクトルは$\mathbf{R}_q$の$k$番目に固有値が大きい固有ベクトルを示している。ここで，$k$番目の空間ケプストラムは変換行列$\mathbf{E}^{\mathsf{T}}$の$k$番目の行ベクトルを用いて以下のように計算される。

$$d_{\tau,k} = \mathbf{e}_k^{\mathsf{T}}\mathbf{q}_\tau = \sum_{n=1}^{N} e_{k,n} q_{\tau,k} \tag{5.18}$$

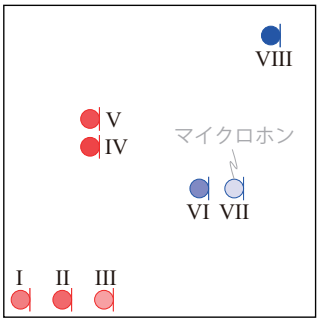
ここで，$d_{\tau,k}$, $\mathbf{e}_k^{\mathsf{T}}$, $e_{k,n}$, $q_{\tau,k}$はそれぞれ，$k$番目の空間ケプストラム，$\mathbf{E}^{\mathsf{T}}$の$k$番目の行ベクトル，$\mathbf{E}^{\mathsf{T}}$の$(k,n)$要素，$\mathbf{q}_\tau$の$k$番目の要素をそれぞれ表す。(5.18)より，$k$番目の空間ケプストラムは対数振幅の線形結合で表現でき，$\mathbf{e}_k$の$n$番目の要素$e_{k,n}$は線形結合の重みとなることが分かる。また，これらの重みを空間的に表現するため，図5.8にそれぞれの重みをマイクロホンの色として表現した図を示す。なお，ここで示すカラーマップは図5.7で用いたものと同じものとする。図5.8 (a)では，全ての重みが正の値となっていることより，1次の空間ケプストラム係数は空間全体の平均的な音圧レベルを表していると考えられる。また，図5.8 (b)–(d)を見ると，近くにあるマイクロホンでは重み係数の値が類似していることが分かる。このことより，本シミュレーション実験において2次から4次までの空間ケプストラム係数は大まかな空間情報を表現していることが分かる。これは，（周波数特徴量である）ケプストラム特徴量にお



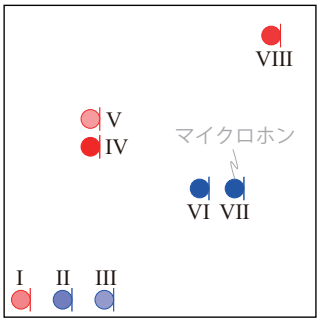

(a) 1次の空間ケプストラム



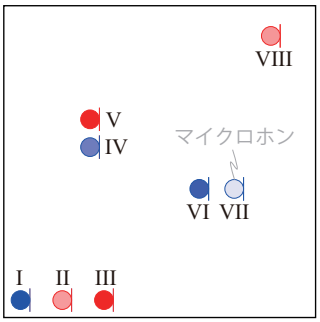

(b) 2次の空間ケプストラム



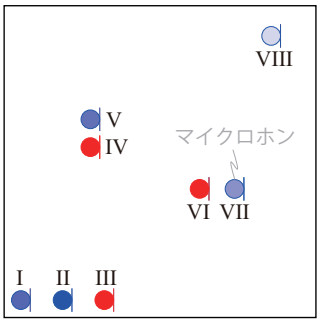

(c) 3次の空間ケプストラム



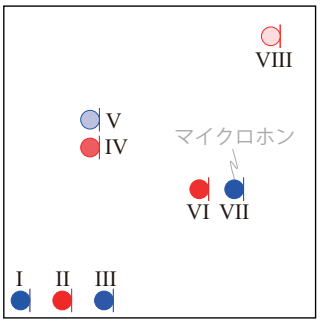

(d) 4次の空間ケプストラム



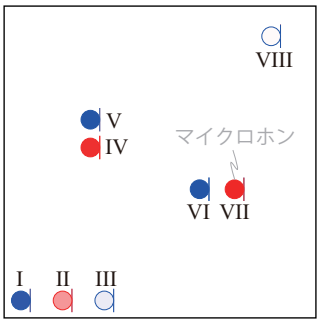

(e) 5次の空間ケプストラム




(f) 6次の空間ケプストラム




(g) 7次の空間ケプストラム




(h) 8次の空間ケプストラム

図 5.8: 各マイクロホンにおける対数振幅に対する変換の重み。マイクロホンは図5.7に示すカラーマップと同じ色により配色されている。

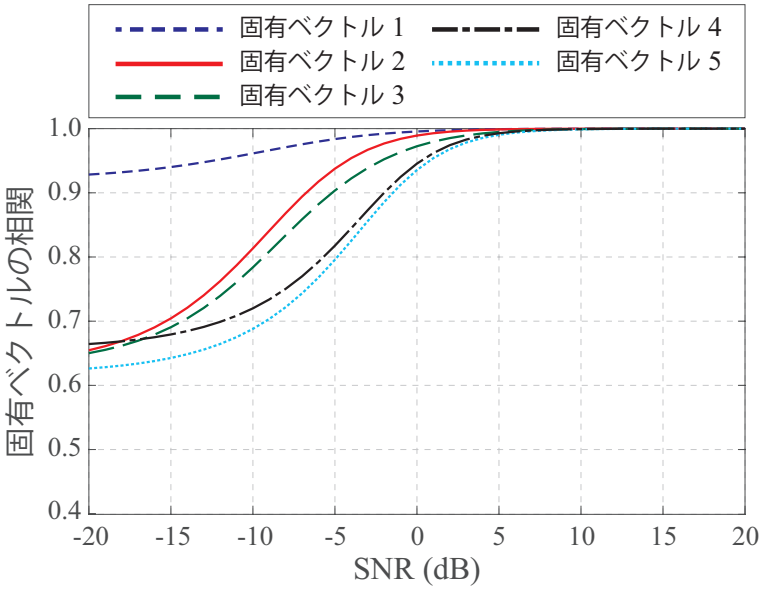



図 5.9: 雑音がない環境と雑音下環境での観測における固有ベクトルの相関係数

いて，低次の空間ケプストラム係数が大まかなスペクトル情報を表現することに対応していると言える。一方で図5.8 (f)–(h)より，6次から8次では近くに配置されているマイクロホンにおいて異なる符号の重みを持つことが分かる。つまり，これらの空間ケプストラム係数は，（周波数特徴量である）ケプストラム特徴量における高次の係数に対応する空間特徴量となっていると考えられる。

次に，空間ケプストラムが背景雑音にどの程度頑健かを評価するための評価実験を行った。ここで，本実験では背景雑音は拡散性を持つ定常的な雑音と仮定し，各マイクロホンでの観測は以下で表されるものとする。

$$\mathbf{q} = \begin{pmatrix} \log(\sqrt{\tilde{a}_1^2 + \nu}) \\ \log(\sqrt{\tilde{a}_2^2 + \nu}) \\ \vdots \\ \log(\sqrt{\tilde{a}_n^2 + \nu}) \\ \vdots \\ \log(\sqrt{\tilde{a}_N^2 + \nu}) \end{pmatrix} \tag{5.19}$$

ここで，$\nu$は背景雑音のパワーとする。背景雑音への頑健性を評価するため，雑音環境下と雑音のないクリーンな環境下での，共分散行列$\mathbf{R}_q$の固有ベクトルの相関係数を比較した。

$$r = \frac{|\mathbf{e}_{noise}^{\mathsf{T}} \mathbf{e}_{clean}|}{\left(\mathbf{e}_{noise}^{\mathsf{T}} \mathbf{e}_{noise}\right)^{1/2} \left(\mathbf{e}_{clean}^{\mathsf{T}} \mathbf{e}_{clean}\right)^{1/2}} \tag{5.20}$$

ここで，$\mathbf{e}_{noise}$と$\mathbf{e}_{clean}$はそれぞれ，雑音環境下と雑音のないクリーンな環境下での，共分散行列$\mathbf{R}_q$の固有ベクトルの天地を表す。図5.9に様々なS/N比で計算した相関係数を示す。ここで，本シミュレーション実験では6つの音源から1フレームずつの観測を想定していること，また，(5.17)で示されるCMNを適用していることより，相関行列$\mathbf{R}_q$のランクが5 (音源数−1)になっている。そのため，図5.9には，5次元分の固有ベクトルの相関係数を示している。実験結果より，S/N比が0 dBの場合においても相関係数は0.95程度となっており雑音環境下においても固有ベクトルは大きく変化しないことが分かる。このことから，定常的な拡散性雑音がある環境に対し，空間ケプストラムは頑健であることが分かる。

## 5.7 実環境収録音による空間ケプストラムの可視化と音響シーン分類実験

### 5.7.1 実環境収録音の収録

空間ケプストラムおよび一般化周波数-空間ケプストラムによる音響シーン分類性能を評価するため，屋内で収録された実環境収録音データセットを用いた空間ケプストラム可視化実験と音響シーン分類実験を行った。評価実験に利用する実環境音データセットを収録するため，13個の時間同期されたマイクロホンを図5.10のように配置した。本データセットでは，屋内のリビングで頻繁に発生する9つの行動（「会話をする」「料理をする」「食事をする」「PCを操作する」「新聞を読む」「掃除する」「皿を洗う」「TVを見る」「洗濯する」）を収録の対象とした。それぞれの音響シーンに含まれる代表的な音響イベントは表5.2に示す通りである。なお，本データセットでは音響シーンの重複は

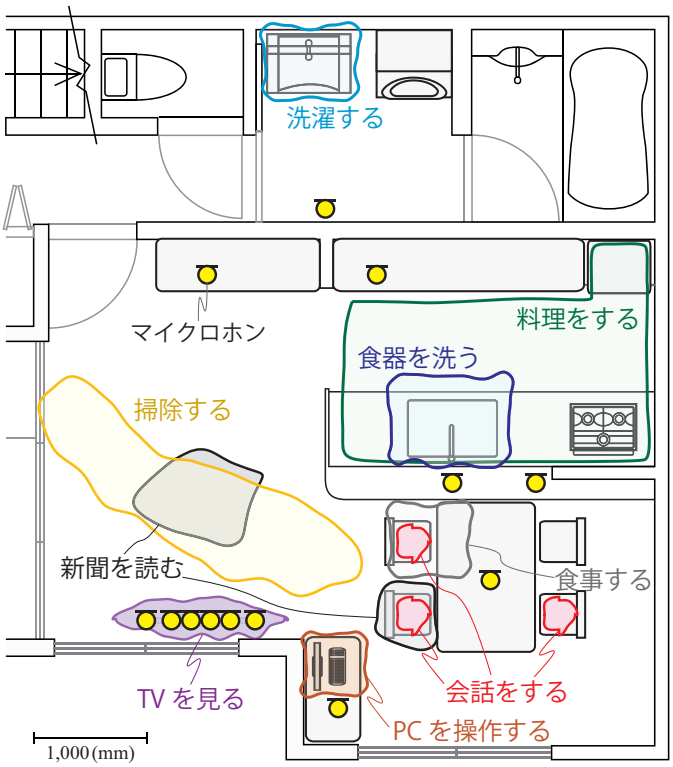



図 5.10: 実環境収録音を用いた音響シーン分類実験における音源とマイクロホンの配置

ないものとし，音響シーンラベルは音の収録を行った後に人手により付与した。本データセットは257.1分の収録時間であり，評価実験では収録音を8秒毎の音クリップに分割し，5,180の音クリップをモデル学習用データとして，また，2,532の音クリップを評価用データとして利用した。収録にはマイクロホンとしてSony ECM-55Bを，マイクロホンアンプとしてGrace Design m802を，A/D変換器としてMOTU 24I/Oを用いた。その他の実験条件は表5.3に示すとおりである。ただし本稿では，実環境収録音におけるS/N比は以下のように定義した。

$$\mathrm{SNR}_{\mathrm{dB}} = 20 \log_{10} \frac{A_{signal}}{A_{noise}} \tag{5.21}$$

ここで，$A_{signal}$および$A_{noise}$はそれぞれ，上記の音響シーンに関連する音の振幅の実効値，および，屋外騒音や収録機器の騒音などの上記音響シーンに関連

表 5.2: それぞれの音響シーンでの代表的な音響イベント

| 音響シーン | 代表的な音響イベント |
|---|---|
| 会話をする | 音声, 咳払い |
| 料理をする | 包丁, 水が流れる音, 食器がぶつかる音 |
| 食事する | 食器がぶつかる音, 音声, 咳払い |
| 洗濯する | 水が流れる音, アラーム音 |
| 食器を洗う | 水が流れる音, 食器がぶつかる音，食器をこする音 |
| 新聞を読む | 新聞をめくる音, 足音 |
| 掃除する | 掃除機の排気音, 足音 |
| TVを見る | 音声, 音楽, 効果音, 歓声 |
| PCを操作する | マウスをクリックする音, キーボードを押下する音, ファンノイズ, 効果音 |

表 5.3: 5章での実環境収録音を用いた音響シーン分類実験における実験条件

| | |
|---|---|
| マイクロホン数 | 13 |
| サンプリング周波数 | 48 kHz |
| 量子化ビット数 | 16 bits |
| 部屋の残響時間 | 0.31 sec. |
| 部屋の床面積 | 22.8 $m^2$ |
| 部屋の平均SNR | 25.2 dB |
| 音クリップの長さ | 8 sec. |
| 時間フレーム長 | 20 msec. |
| FFT長 | 2,048 |
| GMMのクラス数 | 3 |
| 周波数ビン数 (GFSC) | 8 |

しない音の振幅の実効値を表す。具体的には，音クリップから音響シーンに関連する音が含まれる部分を抽出して振幅の実効値を計算したものを$A_{signal}$とし，屋外の自動車騒音や収録機器の騒音が含まれた信号を抽出して振幅の実効値を計算したものを$A_{noise}$とした。

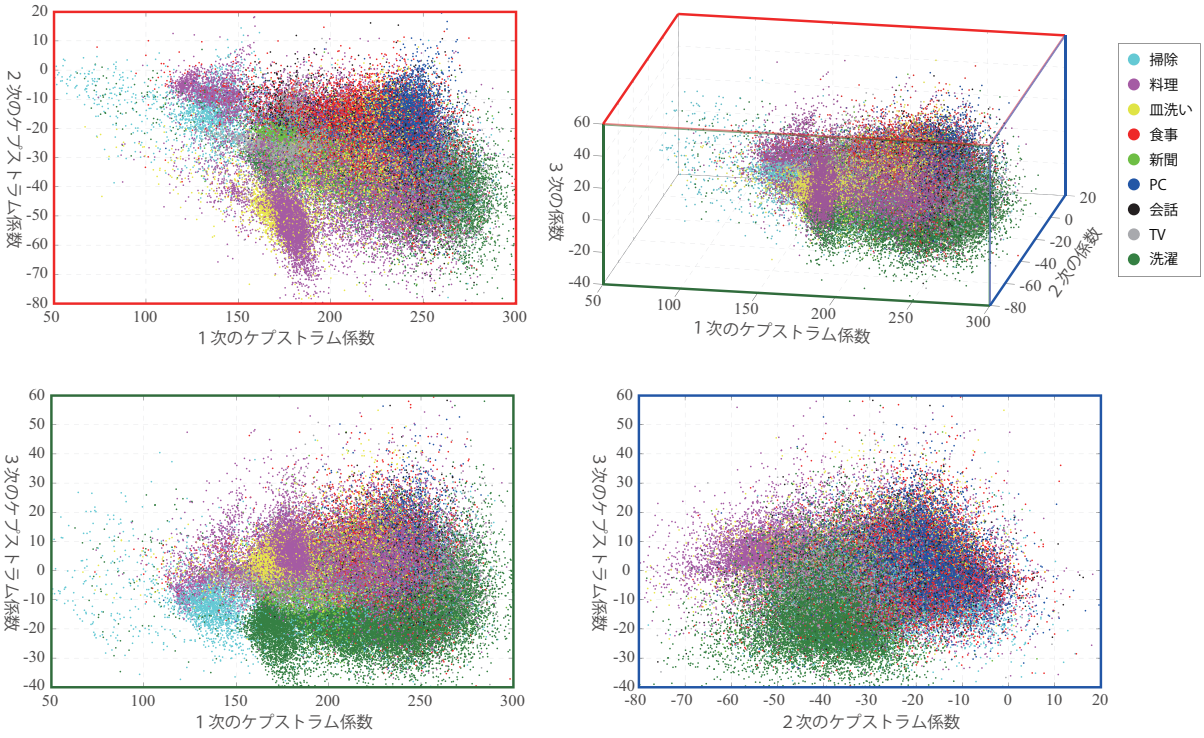



図 5.11: 空間ケプストラム領域で表現される空間情報（空間ケプストラムの上位3次元をプロット）

### 5.7.2 空間ケプストラムと一般化周波数-空間ケプストラムの算出

空間ケプストラムおよび一般化周波数-空間ケプストラムを算出するため，学習データを時間フレーム分割し，(5.6)を用いて共分散行列$\mathbf{R}_q$および$\mathbf{R}_s$を計算した後，主成分分析により固有ベクトルを算出した。また，一般化周波数-空間ケプストラムにおいては，周波数情報を用いるためサブバンド分析を行い周波数-空間特徴行列$\mathbf{S}_\tau$を計算した。なお本実験では，空間ケプストラムおよび一般化周波数-空間ケプストラムを算出した後，各学習用データと評価用データに対してそれぞれCMNを適用した。

### 5.7.3 空間ケプストラムにより表現される空間情報の可視化

空間ケプストラムにより各音響シーンの空間情報がどのように表現されるの

かを確認するため，評価データの空間ケプストラムの上位3次元を音響シーン毎に色分けしてプロットした。具体的には学習用データから変換行列$\mathbf{E}^{\mathsf{T}}$を求めた後，評価データに対して(5.8)を用いて空間ケプストラムを算出した。

図5.11に結果を示す。図より，実際の空間において他の音響シーンと大きく異なる「洗濯」するシーンは，空間ケプストラム領域上でも他の音響シーンと大きく異なる値を取っていることが分かる。また，実際の空間において「皿洗い」をするシーンは「料理」をするシーンと同じ空間で発生しているが，空間ケプストラム領域上でも同じように「料理」をするシーンと値の範囲が重なっていることが分かる。これらの結果より，空間ケプストラム特徴量により空間情報を取得可能であることが分かる。

### 5.7.4 音響シーンの分類

空間ケプストラムによる音響シーンの分類性能を確認するため，対角化共分散行列を用いたGMMにより音響シーンをモデル化/分類した。具体的には，音響シーン$x$毎の特徴量ベクトル$\mathbf{d}_\tau$や$\mathbf{g}_\tau$に対してGMMを学習し，音クリップ$c$に対する音響シーン$x_c$を時間フレーム毎の尤度の積により以下のように推定した。

$$x_c = \arg\max_x \prod_{\tau=1}^{T_c} p_\tau(\mathbf{f}_\tau|x) \tag{5.22}$$

ここで，$T_c, \mathbf{f}_\tau, p_\tau(\mathbf{f}_\tau|x)$はそれぞれ，音クリップ$c$における時間フレーム数，音響特徴ベクトル（$\mathbf{d}_\tau, \mathbf{g}_\tau$など），音響シーン$x$に対する尤度を表す。

### 5.7.5 音響シーン分類の比較手法

提案手法との比較のため，センサネットワークにおけるパターン認識で利用されている，1) Early fusion-based methodおよび，2) Late fusion-based methodを用いた音響シーン分類も評価を行った。 Early fusion-based methodでは，多チャネルのマイクロホンで観測された音クリップは事前に集約され，チャネル平均された後に時間フレーム毎の音響特徴量が算出される。本実験では音響特徴量として対数振幅ベクトル，ケプストラム係数，MFCCs，空間ケプストラ

ム係数とMFCCを連結した特徴量を用いた。ケプストラム係数およびMFCCsにおいては，1024次のケプストラム係数と20次元のMFCCsを算出した後，低次の12次元の係数を用いた。音響シーンのモデル化と分類は5.7.4に示す方法と同様の手順で実施した。また，音響シーン分類の代表的な手法との比較を行うため，Bag-of-acoustic wordとSupport vector machine (SVM) を用いた手法 [37]の評価も行った。具体的な手順としては，まず，Early fusion-based methodと同様に多チャネルの信号を事前に集約し，チャネル平均した後時間フレーム毎のMFCCsを算出した。その後，256混合のGMMにより音響ワードラベルに変換し，音クリップ毎の音響ワードヒストグラムを算出し，この音響ワードヒストグラムを特徴量としてSVMを用いて音響シーンのモデル化と分類を実施した。

次に，Late fusion-based methodによる音響シーン分類について述べる。 Late fusion-based methodではまず，チャネル毎に音響特徴量を算出した後，(5.22)を用いてチャネル毎の音響シーンの分類を行う。その後，尤度が最大となる音響シーンのラベルとその尤度を集約し，最大の尤度を持つチャネルの音響シーン分類結果を最終的な分類結果と判断する。本実験では，ケプストラム係数とMFCCsを音響特徴量として用いた。

### 5.7.6 音響シーン分類結果

図5.12, 5.13に音響シーン分類結果としてConfusion matrixを，図5.14にF-scoreを示す。これらの結果より，空間ケプストラムや一般化周波数-空間ケプストラムにより表現された空間情報は，ケプストラムやMFCCsにより表現される周波数情報と同様に音響シーン分類に効果的であることが分かる。特に，図5.14より，周波数特徴量ではよく似ているが異なる空間パターンを持つ「食器を洗う」シーンと「洗濯する」シーンは，空間ケプストラムを用いることで効果的に分類できることが分かる。一方で，空間ケプストラムを用いた場合，「料理をする」シーンと「食器を洗う」シーンのように類似する空間パターンを持つ音響シーンは混同されやすいことが分かる。また，空間ケプストラムでは，音源に動きを伴う音響シーンである「掃除する」シーンも頑健に分類できることが分かる。

音響特徴量の次元数，全ての音響シーンに対する平均F-score，各チャネル

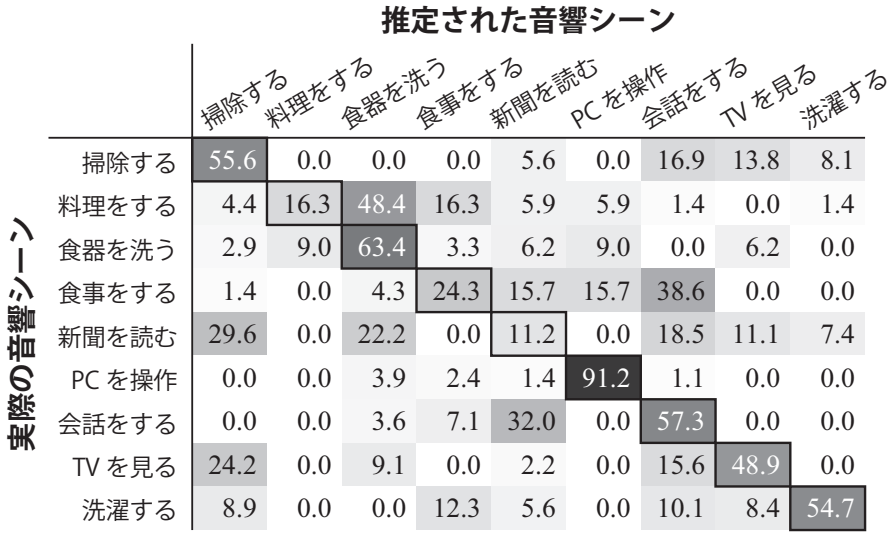



図 5.12: 13次元の空間ケプストラムを特徴量として用いた場合の音響シーン分類性能（再現率）

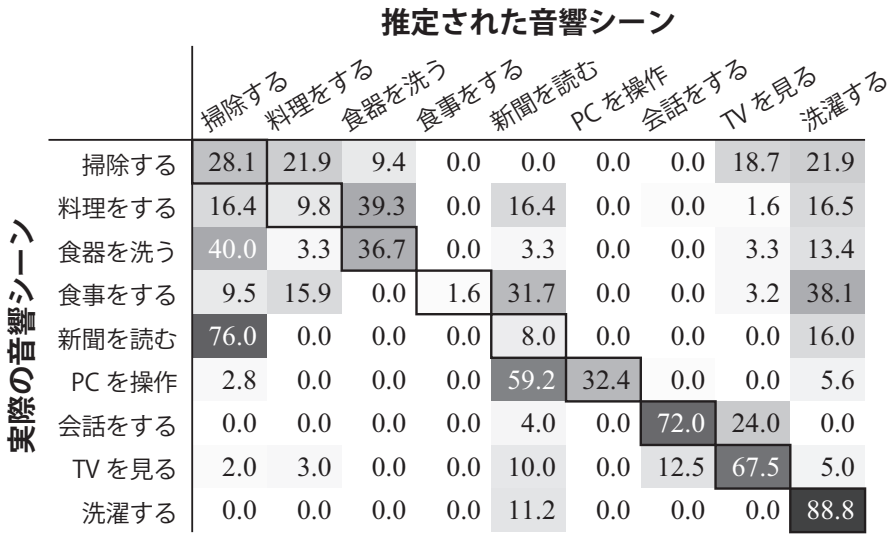



図 5.13: 12次元のMFCCsを特徴量として用いた場合の音響シーン分類性能（再現率）

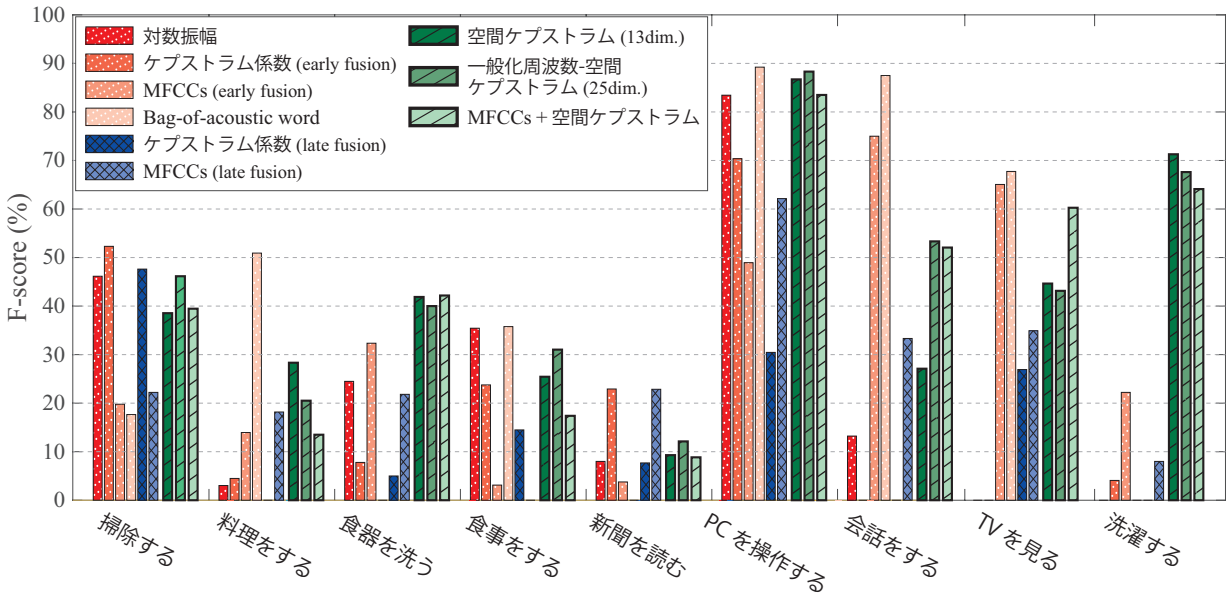



図 5.14: 様々な特徴量に対する音響シーン分類性能（F-score）

を集約する際に掛かるコミュニケーションコストを表5.7.5に示す。ここで$m$は時間フレーム長（時間フレームのサンプル数）を表す。また，コミュニケーションコストは音クリップ毎に計算するものとし，Early fusion-based methodでは音響信号そのものが通信されるものとし，空間ケプストラムや一般化周波数-空間ケプストラムでは時間フレーム毎のパワーを通信するものとする。また，Late fusion-based methodでは，尤度が最大となる音響シーンのラベルとその尤度が音クリップ毎に通信されるものとする。

実験結果より，周波数特徴量と空間特徴量を組み合わせることで，いずれかの特徴量のみを用いるよりも音響シーン分類の性能を向上できることが分かる。また，本実験ではMFCCsと空間ケプストラムを組み合わせた場合に最も高い分類性能 (55.4%)が得られた。一方で，提案手法はチャネル間の情報集約に高いコミュニケーションコストが必要となるため，空間ケプストラムを用いる際は通信環境を考慮する必要がある。

図5.15に空間ケプストラムおよび一般化周波数-空間ケプストラムにおいて，音響シーンのモデル化/分類に用いる特徴量次元数を変化させた場合の分類性能を示す。表5.7.5および図5.15より，元の特徴次元よりも小さい特徴量次元を用いた場合においても，性能が大きく劣化することなく音響シーン分類が実

表 5.4: 空間ケプストラムと従来の音響特徴量に対する平均分類性能（F-score），特徴量の次元，音響シーンのモデル化手法，通信コスト

| 音響特徴量 | 特徴量次元数 | シーンモデル化手法 | 平均 **F-score** | 通信コスト (/channel) | |
|---|---|---|---|---|---|
| | | | | 連続値 | 離散値 |
| 対数振幅 | 13 | GMM | 25.4% | $T_c$ | – |
| ケプストラム係数 (Early fusion) | 12 | GMM | 21.5% | $m \cdot T_c$ | – |
| MFCCs (Early fusion) | 12 | GMM | 42.4% | $m \cdot T_c$ | – |
| Bag-of-acoustic word (Early fusion) | 256 | SVM | 39.8% | $m \cdot T_c$ | – |
| ケプストラム係数 (Late fusion) | 12 | GMM | 21.5% | 1 | 1 |
| MFCCs (Late fusion) | 12 | GMM | 33.3% | 1 | 1 |
| 空間ケプストラム | 13 | GMM | 46.8% | $T_c$ | – |
| 空間ケプストラム | 6 | GMM | 43.8% | $T_c$ | – |
| 一般化周波数-空間ケプストラム | 25 | GMM | 50.4% | $\Omega \cdot T_c$ | – |
| 一般化周波数-空間ケプストラム | 13 | GMM | 54.0% | $\Omega \cdot T_c$ | – |
| MFCCs + 空間ケプストラム | 25 | GMM | **55.4%** | $(m + \Omega) \cdot T_c$ | – |

現できていることが分かる。また，空間ケプストラムを用いた場合と対数振幅特徴量を用いた場合においては，音響シーン分類性能に大きな差があることが分かる。これは，GMMにおいて対角化共分散行列を用いたことに起因すると考えられ，対数振幅特徴量を用いた場合においてはチャネル間相関が大きいためこのような結果になっていると考えられる。

これらの実験結果より，空間ケプストラムや一般化周波数-空間ケプストラムを用いることで，効果的で効率的な音響シーン分類を実現できることが分かる。

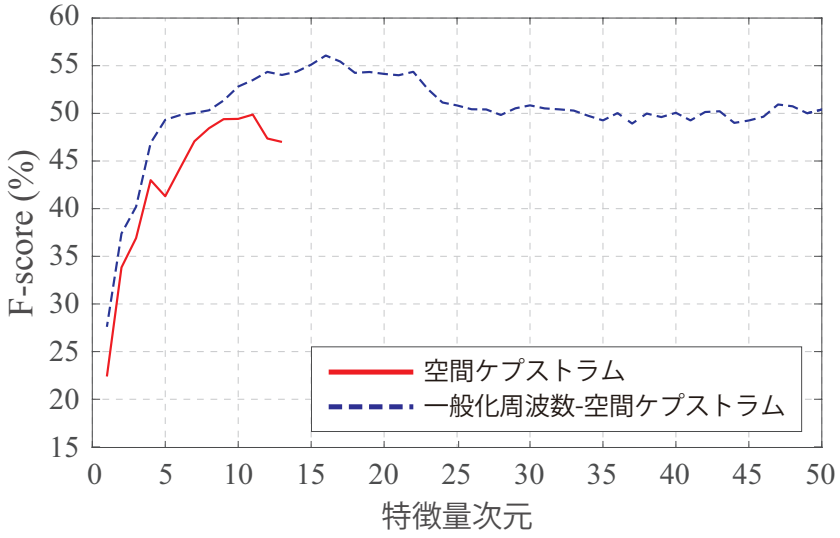



図 5.15: 様々な特徴量次元の空間ケプストラムと一般化周波数-空間ケプストラムにおける音響シーン分類性能

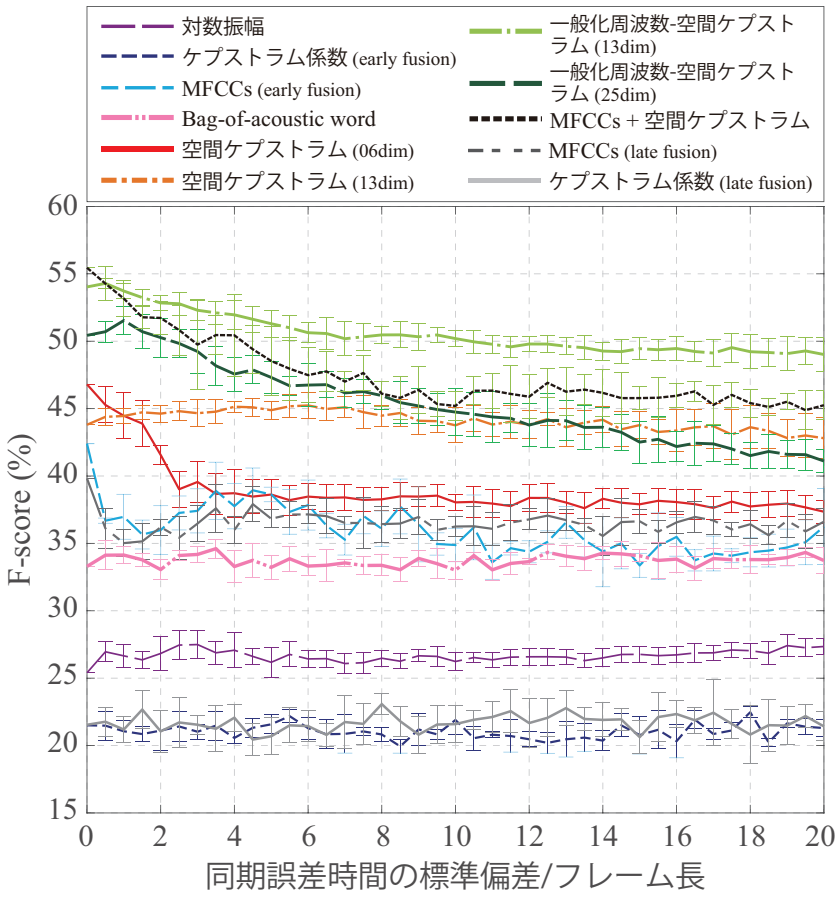



図 5.16: 様々なチャネル間の同期誤差時間を持つ観測に対する音響シーン分類性能

## 5.8 チャネル間の時間同期誤差に対する頑健性の評価実験

チャネル間に時間同期誤差がある場合における音響シーン分類性能を評価するため，様々な同期誤差時間を持つデータセットを用いた音響シーン分類実験を行った。本実験では，5.7節と同じデータセットを用いるが，評価用の音クリップに様々な時間同期誤差を加える。ここで，各チャネルに与える誤差は平均が$\mu = 0$で分散が$\sigma$の正規分布からランダムにサンプルした値を用いる。その他の実験条件は5.7節の実験と同じとする。

図5.16に音響シーン分類結果（F-score）を示す。それぞれの実験条件において異なる同期誤差を用いた分類実験を10回行い，その平均のF-scoreと分散を図に示している。実験結果より，空間ケプストラムおよび一般化周波数-空間ケプストラムでは，平均の同期誤差が小さい場合に音響シーン分類性能は大幅には低下しないことが分かる。一方でMFCCsとEarly fusion-based methodを用いた手法では，わずかな時間同期誤差に対しても音響シーン分類性能が急激に低下していることが分かる。これらの結果より，空間ケプストラムは正確に同期を取ることが困難な状況においても，チャネル間の相互相関を取り大まかに同期を取ったり，キューとなる音を手掛かりに大まかに同期を取ることさえできれば，大幅な性能低下を招くことなく音響シーンの分類が可能であると考えられる。

また，13次元の空間ケプストラム係数を用いた場合と比較して，6次元の空間ケプストラム係数を用いた場合の方が，同期誤差の増加による音響シーン分類性能の低下が小さいことが分かる。これは，同期誤差により生じた誤差成分が空間ケプストラムの高次の係数に含まれるためと考えられる。同様の理由により，一般化周波数-空間ケプストラム係数はMFCCsと空間ケプストラムを連結した特徴量よりも同期誤差に頑健であることが実験結果から分かる。

## 5.9 5章のまとめ

本章では，空間情報を用いた音響トピックモデルを実現するための検討を行い，任意の空間特徴抽出法と従来の音響トピックモデルを組み合わせた手法

や複数チャネルの音響ワードを同時に生成する音響トピックモデルについて議論を行った。また，時間同期誤差を持つ多チャネル観測から空間情報を取得し，音響シーン分類に活用するための手法の検討を行い，空間ケプストラムと呼ばれる新たな特徴抽出方法を提案した。空間ケプストラムは，多チャネルの観測の対数振幅をPCAにより基底変換することで得られる特徴量であり，特定の条件下ではケプストラムとその算出方法が一致することから，“空間ケプストラム”と称することを述べた。また本章では，周波数情報と空間情報を同時に抽出できるように空間ケプストラムを拡張した，一般化周波数-空間ケプストラムを提案した。実環境収録音を用いた音響シーン分類実験により，空間ケプストラムおよび一般化周波数-空間ケプストラムは周波数特徴量同様に音響シーン分類に効果的であることが分かった。また，実験結果よりチャネル間の同期誤差がある場合においても提案手法は頑健に音響シーンを分類できることも分かった。

# 6 結論

本論文では，様々な条件の下で観測された音情報から音響シーンを分類することを目的として，音響トピックモデルに基づいた新たな音響シーン分類手法を提案した。本論文の学術的貢献は，音響シーン分類問題において音情報の観測条件としてこれまでに取り扱われていなかった，逐次的に音情報が観測される場合や間欠的に欠損を有する音情報が観測される場合，また，同期誤差を持つ分散マイクロホンアレイにより多チャネルの音が観測される場合において，音響シーンを高精度に分類するための新たなアプローチを提案したことである。

第1章では，研究背景と，音響イベント検出や音響シーン分析における音情報の観測の多様性について述べた。

第2章では，音響イベント検出や音響シーン分析で用いられる用語と問題設定の整理を行い，本論文では音響シーン分類問題を扱うことを述べた。その後，音響シーン分類手法として近年盛んに取り組まれている機械学習に基づく手法を概説し，本論文で提案する音響シーン分類の方針について述べた。従来手法を用いた音響シーン分類では，音響シーンのモデルを学習するため

に膨大な量のデータを事前に用意しておく必要があるが，音声に限らない多様な性質を持つ音を対象とする音響シーン分類においては，モデル学習に十分といえる量のデータを事前に収集することは困難であるため，高い分類性能を実現することが出来なかった。そこで本論文では，音の観測情報が持つ潜在的構造のスパースさを利用して，限られたデータからでも過学習することなく音響シーンの学習と分類が可能な音響トピックモデルに着目することを述べた。

3章では，音情報が逐次的に得られる場合に，適応的に音響シーンをモデル化する手法について検討を行った。本論文では，音響シーンに関する情報をパラメータとしてモデルに保存しておき，逐次的にこのパラメータを更新可能であること，また，音のスパース性をモデルに導入でき限られたデータからでも過学習することなく音響シーンをモデル化可能であることから，音響トピックモデルに基づく逐次学習手法を提案した。提案手法ではさらに，少ない計算コストで高い精度のパラメータ推定を可能にするため，崩壊型変分ベイズ法の0次近似を用いた手法と崩壊型ギブスサンプリングを組合せたハイブリッド型パラメータ推定方法も導入した。実環境収録音を用いた評価実験の結果，提案手法により，逐次学習した音響シーンモデルから高精度に音響シーン分類が実現できることを示した。

4章では，観測の一部が完全に欠損している場合に，音響シーンを分類し，さらに欠損した観測を同時に推定する問題について検討を行った。本論文では，欠損した観測を潜在的な確率変数とみなすことで，音響ワード系列と欠損した観測を同一の生成モデルで扱うことができる点に着目し，音響トピックモデルを拡張した新たな音響シーン分類手法を提案した。提案モデルでは，音の時間連続性に基づき観測の時間遷移を教師あり音響トピックモデルに組み込むことで，欠損した観測を推定しながら音響シーンの分類を行うことを可能とする。実環境収録音を用いた評価実験により，提案手法では70%程度音の観測に欠損があった場合でも，従来手法と同程度の音響シーン分類性能を実現できることを確認した。

5章では，マイクロホンの位置が未知であり，マイクロホン間の時間同期が正確でない多チャネル観測から音響シーンを分類する手法について検討を行った。本章では，まず音響トピックモデルにより空間情報を扱う方法について議論を行い，任意の空間特徴抽出法と組み合わせ可能な音響トピックモデルの実現例を示した。その後，マイクロホンの位置が未知であり，かつ，マ

イクロホン間の時間同期が正確でない多チャネル観測から頑健に空間情報を取得する方法として，空間ケプストラムを提案した。空間ケプストラムでは，多チャネル観測から対数振幅ベクトルに主成分分析を行うことで空間特徴を取得する。実環境収録音を用いた評価実験により，空間ケプストラムを用いることで，時間同期誤差を含む多チャネル観測に対して頑健な音響シーン分類が可能になることを示した。

スマートホンやウェアラブルデバイス，スマート家電，街頭センサなど，気軽に利用可能なマイクロホンは今後も大きく増加することが想定され，音の収録環境は今後ますます多様なものになっていくと考えられる。本論文では，これまでに音響シーン分類問題の観測条件として扱われていなかった，逐次的な観測や間欠的に欠損を有する観測，同期誤差を持つ多チャネル観測に対する音響シーン分類を可能とした。今後は，高残響環境下など，本研究で取り扱わなかった収録条件における音響シーン分類や，欠損を有する音情報が逐次的に得られる場合など，複合的な収録条件下における音響シーン分類についての検討が課題となる。また，本研究では，同一時刻に音響シーンが重なることは考慮していないため，複数の音響シーンが同時に発生した場合の音響シーン分類についても検討が必要である。

# 参考文献

[1] A. Harma, M. F. McKinney, and J. Skowronek. Automatic surveillance of the acoustic activity in our living environment. *Proc. IEEE International Conference on Multimedia and Expo* (*ICME*), 2005.

[2] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo. CLEAR evaluation of acoustic event detection and classification systems. *Springer Berlin Hidelberg*, pages 311–322, 2007.

[3] Y. Peng, C. Lin, M. Sun, and K. Tsai. Healthcare audio event classification using hidden Markov models and hierarchical hidden Markov models. *Proc. IEEE International Conference on Multimedia and Expo* (*ICME*), pages 1218–1221, 2009.

[4] A. Temko and C. Nadeu. Acoustic event detection in meeting-room environments. *Pattern Recognition Letters*, 30(14):1281–1288, 2009.

[5] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen. Acoustic event detection in real life recordings. *Proc. 18th European Signal Processing Conference* (*EUSIPCO*), pages 1267–1271, 2010.

[6] Y. Ohishi, D. Mochihashi, T. Matsui, M. Nakano, H. Kameoka, T. Izumitani, and K. Kashino. Bayesian semi-supervised audio event transcription based on Markov Indian buffet process. *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pages 3163–3167, 2013.

[7] A. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi. Audio-based context recognition. *IEEE Trans. Audio Speech Lang. Process.*, pages 321–329, 2006.

[8] K. Imoto, S. Shimauchi, H. Uematsu, and H. Ohmuro. User activity estimation method based on probabilistic generative model of acoustic event sequence with user activity and its subordinate categories. *Proc. INTERSPEECH*, 2013.

[9] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley. Acoustic scene classification: Classifying environments from the sounds they produce. *IEEE Signal Process. Mag.*, pages 16–34, 2015.

[10] T. Zhang and C. J. Kuo. Audio content analysis for online audiovisual data segmentation and classification. *IEEE Trans. Audio Speech Lang. Process.*, 9(4):441–457, 2001.

[11] Q. Jin, P. F. Schulam, S. Rawat, S. Burger, D. Ding, and F. Metze. Event-based video retrieval using audio. *Proc. INTERSPEECH*, 2012.

[12] R. Radhakrishnan, A. Divakaran, and P. Smaragdis. Audio analysis for surveillance applications. *Proc. 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), pages 158–161, 2005.

[13] S. Ntalampiras, I. Potamitis, and N. Fakotakis. On acoustic surveillance of hazardous situations. *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pages 165–168, 2009.

[14] S. Lecomte, R. Lengellé, C. Richard, F. Capman, and B. Ravera. Abnormal events detection using unsupervised one-class svm-application to audio surveillance and evaluation. *Proc. 8th IEEE International Conference on Advanced Video and Signal-Based Surveillance* (*AVSS*), pages 124–129, 2011.

[15] Y. Lee, D. K. Han, and H. Ko. Acoustic signal based abnormal event detection in indoor environment using multiclass adaboost. *IEEE Trans. Consum. Electron.*, 59(3):615–622, 2013.

[16] P. Guyot, J. Pinquier, and R. André-Obrecht. Water sound recognition based on physical models. *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pages 793–797, 2013.

[17] M. A. M. Ahaikh, M. K. I. Molla, and K. Hirose. Automatic life-logging: A novel approach to sense real-world activities by environmental sound cues and common sense. *Proc. 11th International Conference on Computer and Information Technology* (*ICCIT*), pages 294–299, 2008.

[18] E. Marchi, F. Vesperini, F. Eyben, S. Squartini, and B. Schuller. A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional lstm neural networks. *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Process. (ICASSP), 2015*, pages 1996–2000, 2015.

[19] N. Waldo, R. Gerard, and H. Perfecto. Automatic event classification using front end single channel noise reduction, mfcc features and a support vector machine classifier. *Proc. the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2013.

[20] J. Schröder, B. Cauchi, M. R. Schädler, N. Moritz, K. Adiloglu, J. Anemüller, S. Doclo, B. Kollmeier, and S. Goetze. Acoustic event detection using signal enhancement and spectro-temporal feature extraction. *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), 2013.

[21] T. Nakatani and M. Masato. Blind dereverberation of single channel speech signal based on harmonic structure. *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pages I–92, 2003.

[22] J. Kürby, R. Grzeszick, A. Plinge, and G. A. Fink. Bag-of-features acoustic event detection for sensor networks. *Proc. the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, pages 55–59, September 2016.

[23] Y. Ohishi. Toward detection and discrimination of all sounds -present and future of audio event detection-. *Proc. 2014 Autumn meeting of Acoustical Society of Japan*, pages 1521–1524, 2014.

[24] M. D. Plumbley and T. Virtanen. Dcase challenge: Philosophy, tasks and results. *Proc. the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, September 2016.

[25] G. Richard. Acoustic scene and events recognition: How similar is it to speech recognition and music genre recognition? *Proc. the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, September 2016.

[26] H. G. Kim, M. Nicolas, and S. Thomas. *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval.* Wiley, October 2005.

[27] K. Muller, S. Mika, G. Ratsch, K. Tsukada, and B. Scholkopf. An introduction to kernel-based learning algorithms. *IEEE Trans. Neural Networks*, pages 181–201, 2001.

[28] V. Franc and H. Vaclav. Multi-class support vector machine. *Proc. 16th Int. Conf. on Pattern Recognition*, pages 236–239, 2002.

[29] J. T. Geiger, B. Schuller, and R. Gerhard. Large-scale audio feature extraction and SVM for acoustic scene classification. *Proc. Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), 2013.

[30] M. Chum, A. Habshush, A. Rahman, and C. Sang. IEEE AASP scene classification challenge using hidden Markov models and frame based classification. *Proc. Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), 2013.

[31] V. Bisot, R. Serizel, and S. Essid. Acoustic scene classification with matrix factorization for unsupervised feature learning. *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pages 6445–6449, 2016.

[32] Y. Xu, Q. Huang, W. Wang, and M. D. Plumbley. Hierarchical learning for DNN-based acoustic scene classification. *Proc. the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, pages 110–114, September 2016.

[33] S. H. Bae, I. Choi, and N. S. Kim. Acoustic scene classification using parallel combination of LSTM and CNN. *Proc. the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, pages 11–15, September 2016.

[34] M. Valenti, A. Diment, G. Parascandolo, S. Squartini, and T. Virtanen. DCASE 2016 acoustic scene classification using convolutional neural networks. *Proc. the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, pages 95–99, September 2016.

[35] M. Zöhrer and F. Pernkopf. Gated recurrent networks applied to acoustic scene classification. *Proc. the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, pages 115–119, September 2016.

[36] E. Marchi, D. Tonelli, X. Xu, F. Ringeval, J. Deng, S. Squartini, and B. Schuller. Pairwise decomposition with deep neural networks and multiscale kernel subspace learning for acoustic scene classification. *Proc. the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, pages 65–69, September 2016.

[37] G. Guo and S. Z. Li. Content-based audio classification and retrieval by support vector machines. *IEEE Trans. Neural Networks*, pages 209–215, 2003.

[38] T. Heittola, A. Mesaros, A. Eronen, and A. Klapuri. Audio content recognition using audio event histograms. *Proc. 18th European Signal Processing Conference* (*EUSIPCO*), pages 1272–1276, 2010.

[39] B. Elizalde, A. Kumar, A. Shah, R. Badlani, E. Vincent, B. Raj, and L. Lane. Experiments on the DCASE challenge 2016: Acoustic scene classification and sound event detection in real life recording. *Proc. the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, pages 20–24, September 2016.

[40] K. Lee and D. P. W. Ellis. Audio-based semantic concept classification for consumer video. *IEEE Trans. Audio Speech Lang. Process.*, pages 1406–1416, 2010.

[41] http://www.cs.tut.fi/sgn/arg/dcase2016/.

[42] https://archive.org/details/chime-home.

[43] http://dirha.fbk.eu/simcorpora.

[44] S. Kim, S. Narayanan, and S. Sundaram. Acoustic topic models for audio information retrieval. *Proc. 2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), pages 37–40, 2009.

[45] K. Imoto, Y. Ohishi, H. Uematsu, and H. Ohmuro. Acoustic scene analysis based on latent acoustic topic and event allocation. *Proc. IEEE International Workshop on Machine Learning for Signal Processing* (*MLSP*), 2013.

[46] K. Imoto and N. Ono. Acoustic scene analysis from acoustic event sequence with intermittent missing event. *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pages 156–160, 2015.

[47] K. Imoto and S. Shimauchi. Acoustic scene analysis based on hierarchical generative model of acoustic event sequence. *IEICE Trans. Inf. & Syst.*, E99-D(10):2539–2549, October 2016.

[48] T. Joachims. Learning to classify text using support vector machines: Methods, theory, and algorithms. *J. Comput. Linguist.*, 29:655–664, 2003.

[49] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *J. Machine Learn. Res.*, 3:993–1022, 2003.

[50] H. Attias. A variational Bayesian framework for graphical models. *Adv. Neural Inf. Proc. Syst. 12*, pages 209–215, 2000.

[51] R. M. Neal. Probabilistic inference using Markov chain Monte Carlo methods. *Department of Computer Science, University of Toronto, Tech. Rep. CRG-TR-93-1*, 1993.

[52] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 1:5228–5235, 2004.

[53] T. P. Minka and J. Lafferty. Expectation propagation for the generative aspect model. *Proc. 17th Conference on Uncertainty in Artificial Intelligence* (*UAI*), 2002.

[54] M. D. Hoffman and D. M. Blei. Structured stochastic variational inference. *arXiv preprint arXiv:1404.4114*, 2014.

[55] J. Paisley, D. M. Blei, and M. D. Jordan. Variational bayesian inference with stochastic search. *arXiv preprint arXiv:1206.6430*, 2012.

[56] M. Welling, Y. W. Teh, and H. Kappen. Hybrid variational/gibbs collapsed inference in topic models. *arXiv preprint arXiv:1206.3297*, 2012.

[57] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh. On smoothing and inference for topic models. *Proc. 25th Conference on Uncertainty in Artificial Intelligence* (*UAI*), 2009.

[58] Y. W. Teh, D. Newman, and M. Welling. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. *Proc. Adv. Neural Inf. Proc. Syst. 18*, pages 1353–1360, 2006.

[59] K. R. Canini, L. Shi, and T. L. Griffiths. Online inference of topics with latent Dirichlet allocation. *Proc. International Conference on Artificial Intelligence and Statistics* (*AISTATS*), 5:856–864, 2009.

[60] M. Sato. Online model selection based on the variational Bayes. *Neural Computation*, 13(7):1649–1681, 2001.

[61] M. D. Hoffman, D. M. Blei, and F. Bach. Online learning for latent Dirichlet allocation. *Proc. Adv. Neural Inf. Proc. Syst. 22*, pages 856–864, 2010.

[62] R. Martin. Spectral subtraction based on minimum statistics. *Proc. 1st European Signal Processing Conference* (*EUSIPCO*), pages 1182–1185, 1994.

[63] K. Lopatka, J. Kotus, and A. Czyzewski. Detection, classification and localization of acoustic events in the presence of background noise for acoustic surveillance of hazardous situations. *Multimedia Tools and Applications*, pages 1–33, 2015.

[64] S. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Audio Speech Lang. Process.*, 27(2):113–120, 1979.

[65] R. McAulay and M. Malpass. Speech enhancement using a soft-decision noise suppression filter. *IEEE Trans. Audio Speech Lang. Process.*, 28(2):137–145, 1980.

[66] A. Adami, L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, and S. Sivadas. QUALCOMM-ICSI-OGI features for ASR. *Proc. International Conference on Spoken Language Processing (ICSLP)*, pages 4–7, 2002.

[67] Y. Ephraim and M. David. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans. Audio Speech Lang. Process.*, 32(6):1109–1121, 1984.

[68] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc, of IEEE*, 77:257–286, 1989.

[69] C. M. Nelke, N. Nawroth, M. Jeub, C. Beaugeant, and P. Vary. Single microphone wind noise reduction using techniques of artificial bandwidth extension. *Proc. 19th European Signal Processing Conference* (*EUSIPCO*), pages 2328–2332, 2012.

[70] S. Adavanne, G. Parascandolo, P. Pertila, T. Heittola, and T. Virtanen. Sound event detection in multichannel audio using spatial and harmonic features. *Proc. the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, pages 6–10, September 2016.

[71] P. Giannoulis, A. Brutti, M. Matassoni, A. Abad, A. Katsamanis, M. Matos, G. Potamianos, and P. Maragos. Multi-room speech activity detection using a distributed microphone network in domestic environments. *Proc. 23rd European Signal Processing Conference* (*EUSIPCO*), pages 1271–1275, 2015.

[72] H. Phan, M. Maass, L. Hertel, R. Mazur, and A. Mertins. A multi-channel fusion framework for audio event detection. *Proc. 2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), pages 1–5, 2015.

[73] N. Ono, H. Kohno, and S. Sagayama. Blind alignment of asynchronously recorded signals for distributed microphone array. *Proc. Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), pages 161–164, 2009.

[74] K. Hasegawa, N. Ono, S. Miyabe, and S. Sagayama. Blind estimation of locations and time offsets for distributed recording devices. *Proc. Latent Variable Analysis and Signal Separation: 9th International Conference, LVA/ICA 2010*, pages 57–64, 2010.

[75] Z. Liu. Sound source separation with distributed microphone arrays in the presence of clock synchronization errors. *Proc. International Workshop on Acoustic Echo and Noise Control* (*IWAENC*), pages 1–4, 2008.

[76] J. Schmalenstroeer and R. Haeb-Umbach. Sampling rate synchronization in acoustic sensor networks with a pre-trained clock skew error model. *Proc. 21st European Signal Processing Conference* (*EUSIPCO*), pages 1–5, 2013.

[77] S. Miyabe, N. Ono, and S. Makino. Blind compensation of interchannel sampling frequency mismatch for ad hoc microphone array based on maximum likelihood estimation. *Elsevier Signal Processing*, 107:185–196, 2 2015.

[78] V. C. Raykar, I. V. Kozintsev, and R. Lienhart. Position calibration of microphones and loudspeakers in distributed computing platforms. *IEEE Trans. Audio Speech Lang. Process.*, 13(1):70–83, 2005.

[79] F. H. Liu, R. M. Stern, X. Huang, and A. Acero. Efficient cepstral normalization for robust speech recognition. *Proc. Workshop on Human Language Technology*, pages 69–74, 1993.

[80] O. Viikki and K. Laurila. Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication*, pages 133–147, 1998.

[81] H. Shimizu, N. Ono, K. Matsumoto, and S. Sagayama. Isotropic noise suppression in the power spectrum domain by symmetric microphone arrays. *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), pages 54–57, 2007.

[82] N. Ito, H. Shimizu, N. Ono, and S. Sagayama. Diffuse noise suppression using crystal-shaped microphone arrays. *IEEE Trans. Audio Speech Lang. Process.*, pages 2101–2110, 2011.

[83] G. Golub and C. Van Loan. *Matrix Computations.* Johns Hopkins University Press, 1996.

# A

# 音響トピックモデルにおけるモデルパラメータの推定

## A.1 音響トピックモデルにおけるVBによるパラメータ推定

音響トピックモデルのパラメータ推定として変分ベイズ法 (VB: Variational Bayes) [49, 50]や崩壊型ギブスサンプリング (CGS: Collapsed Gibbs sampling) [52]に基づく手法が知られている。 VB法は比較的少ない演算量でパラメータの推定が可能である一方で，各パラメータに対して平均場近似と呼ばれる独立性を仮定するため，パラメータの推定結果が局所解に陥りやすくCGS法と比較して推定精度が低下する場合があることが知られている。 CGS法では，事後確率に対する潜在変数のサンプリングを無限回繰り返すと理論的に大域的最適解を推定することが可能であるが，精度の良い推定を実現するためには多数回のサンプリングを行う必要があり，計算コストが非常に大きくなると

いう問題がある。本節では，VB法に基づく音響トピックモデルのパラメータ推定手法について述べる。また，本節の以下の議論では可読性を考慮し各変数の添字を省略した表記を用いる。

VB法では，推定したい潜在変数や生成分布をパラメータとして持つ変分事後分布$q(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi})$を定義し，Jensenの不等式と平均場近似を用いて変分事後分布を繰り返し真の事後分布$p(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{e})$に近づけることによりパラメータの推定を行う。まず，Jensenの不等式より，全ての未知量に対する周辺対数尤度の下限値$\mathcal{F}[q]$を計算すると以下のようになる。

$$\begin{aligned}\mathcal{L}(\boldsymbol{e}) &\triangleq \log p(\boldsymbol{e}|\alpha,\beta)\\ &= \log\iint\sum_{z} p(\boldsymbol{e},\boldsymbol{z},\boldsymbol{\phi},\boldsymbol{\theta}|\alpha,\beta)d\boldsymbol{\phi}d\boldsymbol{\theta}\\ &\geq \iint\sum_{z} q(\boldsymbol{z},\boldsymbol{\phi},\boldsymbol{\theta})\log\frac{p(\boldsymbol{e},\boldsymbol{z},\boldsymbol{\phi},\boldsymbol{\theta}|\alpha,\beta)}{q(\boldsymbol{z},\boldsymbol{\phi},\boldsymbol{\theta})}d\boldsymbol{\phi}d\boldsymbol{\theta}\\ &\triangleq \mathcal{F}[q]\end{aligned} \tag{A.1}$$

また，$\mathcal{L}(\boldsymbol{e})$と$\mathcal{F}[q]$はKLダイバージェンスを用いて以下のように表現する事も可能である。

$$\begin{aligned}\mathcal{L}(\boldsymbol{e})-\mathcal{F}[q] &= \iint\sum_{z} q(\boldsymbol{z},\boldsymbol{\phi},\boldsymbol{\theta})\log\frac{q(\boldsymbol{z},\boldsymbol{\phi},\boldsymbol{\theta})}{p(\boldsymbol{z},\boldsymbol{\phi},\boldsymbol{\theta}|\boldsymbol{e})}d\boldsymbol{\phi}d\boldsymbol{\theta}\\ &= \mathrm{KL}\Big(q(\boldsymbol{z},\boldsymbol{\phi},\boldsymbol{\theta}),p(\boldsymbol{z},\boldsymbol{\phi},\boldsymbol{\theta}|\boldsymbol{e})\Big)\end{aligned} \tag{A.2}$$

Jensenの不等式より下限値$\mathcal{F}[q]$を最大化すれば，変分事後分布$q(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi})$は$p(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{e})$の最良近似となることが分かる。また，これは(A.2)よりKLダイバージェンス基準において変分事後分布を真の分布に近似していると解釈する事が可能である。ここで，VB法による音響トピックモデルのパラメータ推定では，以下の平均場近似を仮定する。

$$q(\boldsymbol{z},\boldsymbol{\phi},\boldsymbol{\theta}) = q(\boldsymbol{\phi})q(\boldsymbol{\theta})q(\boldsymbol{z}) \tag{A.3}$$

このとき，(A.1)より$\mathcal{F}[q]$は以下のように変形することができる。

$$\begin{aligned}\mathcal{F}[q] &= \iint \sum_{z} q(\boldsymbol{\phi})q(\boldsymbol{\theta})q(z)\log\frac{p(\boldsymbol{e},z,\boldsymbol{\phi},\boldsymbol{\theta}|\alpha,\beta)}{q(\boldsymbol{\phi})q(\boldsymbol{\theta})q(z)}d\boldsymbol{\phi}d\boldsymbol{\theta} \\ &= \iint \sum_{z} q(\boldsymbol{\phi})q(\boldsymbol{\theta})q(z)\log\left\{p(\boldsymbol{e}|z,\boldsymbol{\theta})p(z|\boldsymbol{\theta})\right\}d\boldsymbol{\phi}d\boldsymbol{\theta} \\ &\quad + \int q(\boldsymbol{\phi})\log\frac{p(\boldsymbol{\phi}|\beta)}{q(\boldsymbol{\phi})}d\boldsymbol{\phi} + \int q(\boldsymbol{\theta})\log\frac{p(\boldsymbol{\theta}|\alpha)}{q(\boldsymbol{\theta})}d\boldsymbol{\theta} - \sum_{z} q(z)\log q(z) \quad \text{(A.4)}\end{aligned}$$

$q(\boldsymbol{\phi}), q(\boldsymbol{\theta}), q(z)$のそれぞれの変数に対して$\mathcal{F}[q]$は凸関数であるので，$\partial\mathcal{F}[q]/\partial q(\boldsymbol{\phi}) = 0, \partial\mathcal{F}[q]/\partial q(\boldsymbol{\theta}) = 0, \partial\mathcal{F}[q]/\partial q(z) = 0$を繰り返し解くことで音響トピックモデルのパラメータは推定される。なお，音響トピックモデルと等価なモデルであるトピックモデルのパラメータ推定の詳細が[49]に示されている。

## A.2 教師あり音響トピックモデルにおけるVBによるパラメータ推定

教師あり音響トピックモデルに対するVB法では，音響トピックモデルの場合と同様，推定したい潜在変数や生成分布をパラメータとして持つ変分事後分布$q(\boldsymbol{a},\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi})$を定義し，Jensenの不等式と平均場近似を用いて変分事後分布を繰り返し真の事後分布$p(\boldsymbol{a},\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{e})$に近づけることによりパラメータの推定を行う。まず，Jensenの不等式より，全ての未知量に対する周辺対数尤度の下限値$\mathcal{F}[q]$を計算すると以下のようになる。

$$\begin{aligned}\mathcal{L}(\boldsymbol{e}) &\triangleq \log p(\boldsymbol{e}|\alpha,\beta,\mathrm{a}) \\ &= \log\iint \sum_{\boldsymbol{a}}\sum_{z} p(\boldsymbol{e},\boldsymbol{a},z,\boldsymbol{\phi},\boldsymbol{\theta}|\alpha,\beta,\mathrm{a})d\boldsymbol{\phi}d\boldsymbol{\theta} \\ &\geq \iint \sum_{\boldsymbol{a}}\sum_{z} q(\boldsymbol{a},z,\boldsymbol{\phi},\boldsymbol{\theta})\log\frac{p(\boldsymbol{e},\boldsymbol{a},z,\boldsymbol{\phi},\boldsymbol{\theta}|\alpha,\beta,\mathrm{a})}{q(\boldsymbol{a},z,\boldsymbol{\phi},\boldsymbol{\theta})}d\boldsymbol{\phi}d\boldsymbol{\theta} \\ &\triangleq \mathcal{F}[q] \quad \text{(A.5)}\end{aligned}$$

Jensenの不等式より，下限値$\mathcal{F}[q]$を最大化すれば，変分事後分布$q(\boldsymbol{a},\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi})$は$p(\boldsymbol{a},\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{e})$の最良近似となることが分かる。ここで，VB法によるパラメータ

推定では，以下の平均場近似を仮定する。

$$q(\boldsymbol{a}, \boldsymbol{z}, \boldsymbol{\phi}, \boldsymbol{\theta}) = q(\boldsymbol{a})q(\boldsymbol{z})q(\boldsymbol{\phi})q(\boldsymbol{\theta}) \tag{A.6}$$

このとき，(A.5)より$\mathcal{F}[q]$は以下のように変形することができる。

$$\begin{aligned}
\mathcal{F}[q] &= \iint \sum_{\boldsymbol{a}} \sum_{\boldsymbol{z}} q(\boldsymbol{a})q(\boldsymbol{z})q(\boldsymbol{\phi})q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{e}, \boldsymbol{a}, \boldsymbol{z}, \boldsymbol{\phi}, \boldsymbol{\theta}|\alpha, \beta, \mathrm{a})}{q(\boldsymbol{a})q(\boldsymbol{z})q(\boldsymbol{\phi})q(\boldsymbol{\theta})} d\boldsymbol{\phi} d\boldsymbol{\theta} \\
&= \iint \sum_{\boldsymbol{a}} \sum_{\boldsymbol{z}} q(\boldsymbol{a})q(\boldsymbol{\phi})q(\boldsymbol{\theta})q(\boldsymbol{z}) \log \left\{ p(\boldsymbol{e}|\boldsymbol{z}, \boldsymbol{\theta}) p(\boldsymbol{z}|\boldsymbol{\theta}) \right\} d\boldsymbol{\phi} d\boldsymbol{\theta} \\
&\quad + \int q(\boldsymbol{\phi}) \log \frac{p(\boldsymbol{\phi}|\beta)}{q(\boldsymbol{\phi})} d\boldsymbol{\phi} + \int q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta}|\alpha)}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} - \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log q(\boldsymbol{z}) \\
&\quad + q(\boldsymbol{a}) \log \frac{p(\boldsymbol{a}|\mathrm{a})}{q(\boldsymbol{a})}
\end{aligned} \tag{A.7}$$

$q(\boldsymbol{a}), q(\boldsymbol{z}), q(\boldsymbol{\phi}), q(\boldsymbol{\theta})$のそれぞれの変数に対して$\mathcal{F}[q]$は凸関数であるので，$\partial\mathcal{F}[q]/\partial q(\boldsymbol{a}) = 0$, $\partial\mathcal{F}[q]/\partial q(\boldsymbol{z}) = 0$, $\partial\mathcal{F}[q]/\partial q(\boldsymbol{\phi}) = 0$, $\partial\mathcal{F}[q]/\partial q(\boldsymbol{\theta}) = 0$を繰り返し解くことで，教師あり音響トピックモデルのパラメータは推定される。

# B

# 3章におけるパラメータ更新式の導出

## B.1 音響トピックモデルにおけるCVB0によるパラメータ更新式の導出

本章では，$\partial\mathcal{F}[\gamma_{sit}]/\partial\gamma_{sit} = 0$を各$\gamma_{sit}$について解く方法を述べる。 $q(z_{s,i} = t) = \gamma_{sit}$を踏まえ，(3.5)の右辺第一項$\sum_z q(\boldsymbol{z}) \log p(\boldsymbol{e}, \boldsymbol{z}|\alpha, \beta)$を$\gamma_{sit}$で偏微分すると以下を得る。

$$
\begin{aligned}
&\frac{\partial}{\partial\gamma_{sit}} \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log p(\boldsymbol{e}, \boldsymbol{z}|\alpha, \beta) \\
&\quad = \sum_{\boldsymbol{z}_{\backslash s,i}, z_{si}=t} \left\{ \prod_{s',i' \neq s,i} \prod_{t'=1}^{T} \gamma_{s'i't'}^{\delta_{s'i't'}} \right\} \log p(\boldsymbol{e}, \boldsymbol{z}|\alpha, \beta) \\
&\quad = \sum_{\boldsymbol{z}_{\backslash s,i}, z_{si}=t} \left\{ \prod_{s',i' \neq s,i} \prod_{t'=1}^{T} \gamma_{s'i't'}^{\delta_{s'i't'}} \right\} \cdot \log \left\{ p(\boldsymbol{e}_{\backslash s,i}, \boldsymbol{z}_{\backslash s,i}|\alpha, \beta) \cdot \frac{p(\boldsymbol{e}, \boldsymbol{z}|\alpha, \beta)}{p(\boldsymbol{e}_{\backslash s,i}, \boldsymbol{z}_{\backslash s,i}|\alpha, \beta)} \right\}
\end{aligned}
$$

$$= \sum_{\boldsymbol{z}_{\backslash s,i}, z_{si}=t} \left\{ \prod_{s',i'\neq s,i} \prod_{t'=1}^{T} \gamma_{s'i't'}^{\delta_{s'i't'}} \right\} \cdot \log \left\{ p(\boldsymbol{e}_{\backslash s,i}, \boldsymbol{z}_{\backslash s,i}|\alpha,\beta) \cdot \frac{\int d\boldsymbol{\theta} p(\boldsymbol{z},\boldsymbol{\theta}|\alpha)}{\int d\boldsymbol{\theta} p(\boldsymbol{z}_{\backslash s,i},\boldsymbol{\theta}|\alpha)} \frac{\int d\boldsymbol{\phi} p(\boldsymbol{e},\boldsymbol{\phi}|\boldsymbol{z},\beta)}{\int d\boldsymbol{\phi} p(\boldsymbol{e}_{\backslash s,i},\boldsymbol{\phi}|\boldsymbol{z}_{\backslash s,i},\beta)} \right\} \tag{B.1}$$

ここで，ディリクレ分布を積分することにより以下のディリクレ積分が得られる。

$$\int d\mu \mathcal{D}(\mu|\zeta) = \frac{\Gamma(\sum_m \zeta_m)}{\prod_j \zeta_j} \int d\mu \prod_j \mu_j^{\zeta_j - 1} = 1$$

ディリクレ積分に加え，ガンマ関数の性質$\Gamma(x+1)/\Gamma(x) = x$を利用すると(B.1)は，さらに以下のように整理できる。

$$\begin{aligned}
&\frac{\partial}{\partial \gamma_{sit}} \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log p(\boldsymbol{e},\boldsymbol{z}|\alpha,\beta) \\
&= \sum_{\boldsymbol{z}_{\backslash s,i}, z_{si}=t} \left\{ \prod_{s',i'\neq s,i} \prod_{t'=1}^{T} \gamma_{s'i't'}^{\delta_{s'i't'}} \right\} \cdot \log \left\{ p(\boldsymbol{e}_{\backslash s,i}, \boldsymbol{z}_{\backslash s,i}|\alpha,\beta) \cdot \frac{\dfrac{\Gamma(n_t^s + \alpha)}{\Gamma(n_{(\backslash s,i),t}^s + \alpha)}}{\dfrac{\Gamma(n_\cdot^s + T\alpha)}{\Gamma(n_{(\backslash s,i),\cdot}^s + T\alpha)}} \cdot \frac{\dfrac{\Gamma(n_m^t + \beta)}{\Gamma(n_{(\backslash s,i),m}^t + \beta)}}{\dfrac{\Gamma(n_\cdot^t + M\beta)}{\Gamma(n_{(\backslash s,i),\cdot}^t + M\beta)}} \right\} \\
&= \sum_{\boldsymbol{z}_{\backslash s,i}, z_{si}=t} \left\{ \prod_{s',i'\neq s,i} \prod_{t'=1}^{T} \gamma_{s'i't'}^{\delta_{s'i't'}} \right\} \cdot \log \left\{ p(\boldsymbol{e}_{\backslash s,i}, \boldsymbol{z}_{\backslash s,i}|\alpha,\beta) \cdot \frac{n_{(\backslash s,i),t}^s + \alpha}{n_{(\backslash s,i),\cdot}^s + T\alpha} \cdot \frac{n_{(\backslash s,i),m}^t + \beta}{n_{(\backslash s,i),\cdot}^t + M\beta} \right\}
\end{aligned} \tag{B.2}$$

次に，(3.5)の右辺第二項$\sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log q(\boldsymbol{z})$を$\gamma_{sit}$で偏微分すると以下を得る。但し，最後の式変形では，$\gamma_{sit}$に依らない項を定数としてまとめて表記している。

$$\begin{aligned}
&\frac{\partial}{\partial \gamma_{sit}} \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log q(\boldsymbol{z}) \\
&= \frac{\partial}{\partial \gamma_{sit}} \sum_{\boldsymbol{z}_{\backslash s,i}, z_{si}=t} \prod_{s',i'\neq s,i} \prod_{t'=1}^{T} \gamma_{s'i't'}^{\delta_{s'i't'}} \gamma_{sit} \left\{ \sum_{s',i'\neq s,i} \sum_{t'=1}^{T} \log \gamma_{s'i't'}^{\delta_{s'i't'}} + \log \gamma_{sit} \right\} \\
&= \sum_{\boldsymbol{z}_{\backslash s,i}, z_{si}=t} \prod_{s',i'\neq s,i} \prod_{t'=1}^{T} \gamma_{s'i't'}^{\delta_{s'i't'}} \sum_{s',i'\neq s,i} \sum_{t'=1}^{T} \log \gamma_{s'i't'}^{\delta_{s'i't'}} \\
&\quad + \sum_{\boldsymbol{z}_{\backslash s,i}, z_{si}=t} \prod_{s',i'\neq s,i} \prod_{t'=1}^{T} \gamma_{s'i't'}^{\delta_{s'i't'}} \left\{ \log \gamma_{sit} + \gamma_{sit} \cdot \frac{1}{\gamma_{sit}} \right\} \\
&= \log \gamma_{sit} + const.
\end{aligned} \tag{B.3}$$

さらに，これらを$\partial\mathcal{F}[\gamma_{sit}]/\partial\gamma_{sit}=0$に代入すると，$\gamma_{sit}$について以下の関係が得られる。

$$\frac{\partial}{\partial\gamma_{sit}}\sum_{z}q(z)\log p(\boldsymbol{e},z|\alpha,\beta)-\frac{\partial}{\partial\gamma_{sit}}\sum_{z}q(z)\log q(z)=0$$

$$\sum_{z_{\backslash s,i},z_{si}=t}\prod_{s',i'\neq s,i}\prod_{t'=1}^{T}\gamma_{s'i't'}^{\delta_{s'i't'}}\log\left\{p(\boldsymbol{e}_{\backslash s,i},\boldsymbol{z}_{\backslash s,i}|\alpha,\beta)\cdot\frac{n^{s}_{(\backslash s,i),t}+\alpha}{n^{s}_{(\backslash s,i),\cdot}+T\alpha}\cdot\frac{n^{t}_{(\backslash s,i),m}+\beta}{n^{t}_{(\backslash s,i),\cdot}+M\beta}\right\}$$

$$-\log\gamma_{sit}+const.=0 \quad \text{(B.4)}$$

ここで，$p(\boldsymbol{e}_{\backslash s,i},\boldsymbol{z}_{\backslash s,i}|\alpha,\beta)$および$(n^{s}_{(\backslash s,i),\cdot}+T\alpha)$が$\gamma_{sit}$に依存しないことを踏まえ，$\gamma_{sit}$について整理すると以下が得られる。

$$\log\gamma_{sit} = \sum_{z_{\backslash s,i},z_{si}=t}\prod_{s',i'\neq s,i}\prod_{t'=1}^{T}\gamma_{s'i't'}^{\delta_{s'i't'}}\cdot\log\left\{\frac{n^{s}_{(\backslash s,i),t}+\alpha}{n^{s}_{(\backslash s,i),\cdot}+T\alpha}\frac{n^{t}_{(\backslash s,i),m}+\beta}{n^{t}_{(\backslash s,i),\cdot}+M\beta}\right\}+const.$$

$$\gamma_{sit} \propto \exp\left[\sum_{z_{\backslash s,i},z_{si}=t}\prod_{s',i'\neq s,i}\prod_{t'=1}^{T}\gamma_{s'i't'}^{\delta_{s'i't'}}\cdot\log\left\{(n^{s}_{(\backslash s,i),t}+\alpha)\cdot\frac{n^{t}_{(\backslash s,i),m}+\beta}{n^{t}_{(\backslash s,i),\cdot}+M\beta}\right\}\right] \quad \text{(B.5)}$$

(B.5)を$q(\boldsymbol{z}_{\backslash s,i})$の期待値を用いて表すことで(3.6)が得られる。

## B.2 音響トピックモデルにおけるCGSによるパラメータの周辺積分の導出

CGS法の更新式の導出に必要となる，各パラメータの周辺積分$p(\boldsymbol{e}|\boldsymbol{z}^{GS},\boldsymbol{z}^{VB},\beta)$, $p(\boldsymbol{z}^{GS},\boldsymbol{z}^{VB}|\alpha)$を算出する。

### B.2.1 $p(e|z^{GS},z^{VB},\beta)$の導出

音響イベント系列の集合全体の中で，$n_t^{GS}$を音響トピック$t$が出現した回数を表すものとすると，$p(\boldsymbol{e},\boldsymbol{\phi}|\boldsymbol{z}^{GS},\boldsymbol{z}^{VB},\beta)$は以下の様に表現可能である。

$$p(\boldsymbol{e},\boldsymbol{\phi}|\boldsymbol{z}^{GS},\boldsymbol{z}^{VB},\beta)$$
$$=\prod_{t=1}^{T}p(\phi^{t}|\beta)\left\{\prod_{i=1}^{n_t^{GS}}\prod_{m=1}^{M}(\phi_m^t)^{\delta_{sim}}\right\}\left\{\prod_{i=1}^{n_t^{VB}}\prod_{m=1}^{M}(\phi_m^t)^{\delta_{sim}\gamma_{sit}}\right\}$$

$$= \left(\frac{\Gamma(M\beta)}{\Gamma(\beta)^M}\right)^T \left\{\prod_{t=1}^{T}\prod_{m=1}^{M}(\phi_m^t)^{\beta-1}\left\{\prod_{i=1}^{n_t^{GS}}(\phi_m^t)^{\delta_{sim}}\right\}\left\{\prod_{i=1}^{n_t^{VB}}(\phi_m^t)^{\delta_{sim}\gamma_{sit}}\right\}\right\}$$

$$= \left(\frac{\Gamma(M\beta)}{\Gamma(\beta)^M}\right)^T \prod_{t=1}^{T}\prod_{m=1}^{M}(\phi_m^t)^{n_m^{tGS}+(\sum_{m=1}^{M}\sum_{i=1}^{n_t^{VB}}\delta_{sim}\gamma_{sit})+\beta-1} \tag{B.6}$$

また，$\sum_{m=1}^{M}\zeta_m = n_\cdot^{tGS} + (\sum_{m=1}^{M}\sum_{i=1}^{n_t^{VB}}\gamma_{sit}) + M\beta$となるようなディリクレ積分は，以下のように表現可能である。

$$\int d\phi \prod_{m}^{M}(\phi_m^t)^{n_m^{tGS}+(\sum_{m=1}^{M}\sum_{i=1}^{n_t^{VB}}\delta_{sim}\gamma_{sit})+\beta-1} = \frac{\prod_{m=1}^{M}\Gamma(n_m^{tGS}+\sum_{m=1}^{M}\sum_{i=1}^{n_t^{VB}}\delta_{sim}\gamma_{sit}+\beta)}{\Gamma(n_\cdot^{tGS}+\sum_{m=1}^{M}\sum_{i=1}^{n_t^{VB}}\gamma_{sit}+M\beta)} \tag{B.7}$$

$p(\boldsymbol{e}|\boldsymbol{z}^{GS},\boldsymbol{z}^{VB},\beta)$に(B.6), (B.7)を代入すると以下の周辺積分を得る。

$$\begin{aligned} p(\boldsymbol{e}|\boldsymbol{z}^{GS},\boldsymbol{z}^{VB},\beta) &= \int d\boldsymbol{\phi}\, p(\boldsymbol{e},\boldsymbol{\phi}|\boldsymbol{z}^{GS},\boldsymbol{z}^{VB},\beta) \\ &= \left(\frac{\Gamma(M\beta)}{\Gamma(\beta)^M}\right)^T \prod_{t=1}^{T}\frac{\prod_{m=1}^{M}\Gamma(n_m^{tGS}+\sum_{m=1}^{M}\sum_{i=1}^{n_t^{VB}}\delta_{sim}\gamma_{sit}+\beta)}{\Gamma(n_\cdot^{tGS}+\sum_{m=1}^{M}\sum_{i=1}^{n_t^{VB}}\gamma_{sit}+M\beta)} \end{aligned} \tag{B.8}$$

(3.13)の右辺第一項に(B.8)を代入すると以下を得る。

$$\frac{p(\boldsymbol{e}|\boldsymbol{z}^{GS},\boldsymbol{z}^{VB})}{p(\boldsymbol{e}_{\backslash s,i}|\boldsymbol{z}_{\backslash s,i}^{GS},\boldsymbol{z}^{VB})} = \frac{\dfrac{\Gamma(n_m^{tGS}+\sum_{m=1}^{M}\sum_{i=1}^{n_t^{VB}}\delta_{sim}\gamma_{sit}+\beta)}{\Gamma(n_{(\backslash s,i),m}^{tGS}+\sum_{m=1}^{M}\sum_{i=1}^{n_t^{VB}}\delta_{sim}\gamma_{sit}+\beta)}}{\dfrac{\Gamma(n_\cdot^{tGS}+\sum_{m=1}^{M}\sum_{i=1}^{n_t^{VB}}\gamma_{sit}+M\beta)}{\Gamma(n_{(\backslash s,i),\cdot}^{tGS}+\sum_{m=1}^{M}\sum_{i=1}^{n_t^{VB}}\gamma_{sit}+M\beta)}} \tag{B.9}$$

(B.9)にガンマ関数の性質$\Gamma(x+1)/\Gamma(x) = x$を用いることで(3.14)が得られる。

### B.2.2 $p(z^{GS}, z^{VB}|\alpha)$の導出

音響イベント系列の集合全体の中で，$n_s^{GS}$を音響イベント系列$s$のうちCGS法に割り当てられた数を表すものとすると，$p(\boldsymbol{z}^{GS},\boldsymbol{z}^{VB},\boldsymbol{\theta}|\alpha)$は以下のように算出でき

る。

$$
\begin{aligned}
p(\boldsymbol{z}^{GS}, \boldsymbol{z}^{VB}, \boldsymbol{\theta}|\alpha) &= \prod_{s=1}^{S} p(\theta^s|\alpha)\left\{\prod_{i=1}^{n_s^{GS}}\prod_{t=1}^{T}(\theta_t^s)^{\delta_{sit}}\right\}\left\{\prod_{i=1}^{n_s^{VB}}\prod_{t=1}^{T}(\theta_t^s)^{\delta_{sit}\gamma_{sit}}\right\} \\
&= \left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\right)^S\left\{\prod_{s=1}^{S}\prod_{t=1}^{T}(\theta_t^s)^{\alpha-1}\left\{\prod_{i=1}^{n_s^{GS}}(\theta_t^s)^{\delta_{sit}}\right\}\left\{\prod_{i=1}^{n_s^{VB}}(\theta_t^s)^{\delta_{sit}\gamma_{sit}}\right\}\right\} \\
&= \left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\right)^S\prod_{s=1}^{S}\prod_{t=1}^{T}(\theta_t^s)^{n_t^{sGS}+(\sum_{t=1}^{T}\sum_{i=1}^{n_s^{VB}}\delta_{sit}\gamma_{sit})+\alpha-1} \qquad \text{(B.10)}
\end{aligned}
$$

また，$\sum_{t=1}^{T}\zeta_t = n_{\cdot}^{sGS} + (\sum_{t=1}^{T}\sum_{i=1}^{N_s^{VB}}\gamma_{sit}) + T\alpha$となるディリクレ積分は以下のように表現可能である。

$$
\int d\theta \prod_{t}^{T}(\theta_t^{sGS})^{n_t^{sGS}+(\sum_{t=1}^{T}\sum_{i=1}^{N_s^{VB}}\delta_{sit}\gamma_{sit})+\alpha-1} = \frac{\prod_{t=1}^{T}\Gamma(n_t^{sGS}+\sum_{t=1}^{T}\sum_{i=1}^{N_s^{VB}}\delta_{sit}\gamma_{sit}+\alpha)}{\Gamma(n_{\cdot}^{sGS}+\sum_{t=1}^{T}\sum_{i=1}^{N_s^{VB}}\gamma_{sit}+T\alpha)} \qquad \text{(B.11)}
$$

$p(\boldsymbol{z}^{GS}, \boldsymbol{z}^{VB}|\alpha)$に(B.10), (B.11)を代入すると以下を得る。

$$
\begin{aligned}
p(\boldsymbol{z}^{GS}, \boldsymbol{z}^{VB}|\alpha) &= \int d\boldsymbol{\theta} p(\boldsymbol{z}^{GS}, \boldsymbol{z}^{VB}, \boldsymbol{\theta}|\alpha) \\
&= \left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\right)^S\prod_{s=1}^{S}\frac{\prod_{t=1}^{T}\Gamma(n_t^{sGS}+\sum_{t=1}^{T}\sum_{i=1}^{N_s^{VB}}\delta_{sit}\gamma_{sit}+\alpha)}{\Gamma(n_{\cdot}^{sGS}+\sum_{t=1}^{T}\sum_{i=1}^{N_s^{VB}}\gamma_{sit}+T\alpha)} \qquad \text{(B.12)}
\end{aligned}
$$

B.2.1と同様，ガンマ関数の性質を利用することで(3.15)が得られる。

# C

# 4章におけるパラメータ更新式の導出

## C.1 音響ワード遷移型教師あり音響トピックモデルにおけるパラメータ更新式の導出

本章では，音響ワード遷移型教師あり音響トピックモデルおよび音響トピック遷移型教師あり音響トピックモデルにおいて，崩壊型ギブスサンプリングによりパラメータを更新する際に用いる音響トピックと音響ワードの同時事後確率$p(e_{s,i}, z_{s,i}|\boldsymbol{e}_{\backslash s,i}, \boldsymbol{z}_{\backslash s,i}, \boldsymbol{a}, \alpha, \beta, \gamma)$の導出の詳細について述べる。まず，(4.4)–(4.7)を(4.3)に代入すると，以下の式が得られる。

$$\frac{p(\boldsymbol{z}|\boldsymbol{a}, \alpha)}{p(\boldsymbol{z}_{\backslash s,i}|\boldsymbol{a}, \alpha)} = \left(\frac{\Gamma(n^a_t + \alpha)}{\Gamma(n^a_{(\backslash s,i),t} + \alpha)}\right) \Big/ \left(\frac{\Gamma(n^a_{\cdot} + T\alpha)}{\Gamma(n^a_{(\backslash s,i),\cdot} + T\alpha)}\right) \tag{C.1}$$

$$\frac{p(\boldsymbol{e}|\boldsymbol{z},\beta,\gamma)}{p(\boldsymbol{e}_{\backslash s,i}|\boldsymbol{z}_{\backslash s,i},\beta,\gamma)} = \frac{\frac{\Gamma(n^t_m+\beta)}{\Gamma(n^t_{(\backslash s,i),m}+\beta)}\frac{\Gamma(n^{m^-}_{m^+}+\gamma)}{\Gamma(n^{m^-}_{(\backslash s,i),m^+}+\gamma)}}{\frac{\Gamma(n^t_{\cdot}+M\beta)}{\Gamma(n^t_{(\backslash s,i),\cdot}+M\beta)}\frac{\Gamma(n^{m^-}_{\cdot}+M\gamma)}{\Gamma(n^{m^-}_{(\backslash s,i),\cdot}+M\gamma)}} \tag{C.2}$$

さらにガンマ関数の性質$\Gamma(x+1)/\Gamma(x) = x$を考慮すると，(C.1)および(C.2)は以下のように変形できる。

$$\frac{p(\boldsymbol{z}|\boldsymbol{a},\alpha)}{p(\boldsymbol{z}_{\backslash s,i}|\boldsymbol{a},\alpha)} = \frac{n^a_{(\backslash s,i),t}+\alpha}{n^a_{(\backslash s,i),\cdot}+T\alpha} \tag{C.3}$$

$$\frac{p(\boldsymbol{e}|\boldsymbol{z},\beta,\gamma)}{p(\boldsymbol{e}_{\backslash s,i}|\boldsymbol{z}_{\backslash s,i},\beta,\gamma)} = \frac{n^t_{(\backslash s,i),m}+\beta}{n^t_{(\backslash s,i),\cdot}+M\beta}\cdot\frac{n^{e_{s,i-1}}_{(\backslash s,i),e_{s,i}}+\gamma}{n^{e_{s,i-1}}_{(\backslash s,i),\cdot}+M\gamma}\cdot\frac{n^{e_{s,i}}_{(\backslash s,i),e_{s,i+1}}+\delta_{e_{s,i-1},e_{s,i}}\cdot\delta_{e_{s,i},e_{s,i+1}}+\gamma}{n^{e_{s,i}}_{(\backslash s,i),\cdot}+\delta_{e_{s,i-1},e_{s,i}}+M\gamma} \tag{C.4}$$

(C.3)と(C.4)を(4.3)に代入することで，音響トピックと音響ワードの事後確率(崩壊型ギブスサンプリングにおける更新式)は(4.8)で表される。

$$\begin{aligned} p(e_{s,i},z_{s,i}|\boldsymbol{e}_{\backslash s,i},\boldsymbol{z}_{\backslash s,i},\boldsymbol{a},\alpha,\beta,\gamma) \propto{}& (n^a_{(\backslash s,i),t}+\alpha)\cdot\frac{n^t_{(\backslash s,i),m}+\beta}{n^t_{(\backslash s,i),\cdot}+M\beta} \\ &\cdot\frac{(n^{e_{s,i-1}}_{(\backslash s,i),e_{s,i}}+\gamma)\left\{n^{e_{s,i}}_{(\backslash s,i),e_{s,i+1}}+\delta_{e_{s,i-1},e_{s,i}}\cdot\delta_{e_{s,i},e_{s,i+1}}+\gamma\right\}}{n^{e_{s,i}}_{(\backslash s,i),\cdot}+\delta_{e_{s,i-1},e_{s,i}}+M\gamma} \end{aligned} \tag{4.8}$$

また，音響ワード$e_{s,i}$が欠損していない場合，それぞれの更新においては音響トピック$z_{s,i}$のみをサンプリングすれば良い。このとき，(4.8)の右辺の最後の項は定数となるため，音響トピック$z_{s,i}$の事後確率は(4.9)で与えられる。

$$p(z_{s,i}|\boldsymbol{e}_{\backslash s,i},\boldsymbol{z}_{\backslash s,i},\boldsymbol{a},\alpha,\beta,\gamma) \propto (n^a_{(\backslash s,i),t}+\alpha)\cdot\frac{n^t_{(\backslash s,i),m}+\beta}{n^t_{(\backslash s,i),\cdot}+M\beta} \tag{4.9}$$

# 研究業績

## 学術論文

[J.1] **Keisuke Imoto**, Nobutaka Ono, “Spatial Cepstrum as a Spatial Feature Using Distributed Microphone Array for Acoustic Scene Analysis,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 25, No. 6.

[J.2] 井本桂右, 大石康智, 植松尚, 大室仲, 小野順貴, “逐次的な観測のための音響シーン分析手法の提案,” 日本音響学会誌, Vol. 72, No. 6, pp. 293-305, 2016.

## 査読付き国際会議

[C.1] **Keisuke Imoto**, Nobutaka Ono, “Online Acoustic Scene Analysis Based on Nonparametric Bayesian Model,” *Proceedings of The 2016 European Signal Processing Conference (EUSIPCO 2016)*, pp. 988–992, 2016.

[C.2] **Keisuke Imoto**, Nobutaka Ono, “Spatial-Feature-Based Acoustic Scene Analysis Using Distributed Microphone Array,” *Proceedings of The 2015 European Signal Processing Conference (EUSIPCO 2015)*, pp. 739–743, 2015.

[C.3] **Keisuke Imoto**, Nobutaka Ono, “Acoustic Scene Analysis from Acoustic Event Sequence with Intermittent Missing Event," *Proceedings of The IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 156–160, 2015.

## 国内学会発表

[D.1] 井本桂右, 小野順貴, “分散マイクロホンアレイを用いた音響イベントの空間パターン特徴表現,” 日本音響学会2015年秋季研究発表会, pp. 139–142, 2015.

[D.2] 井本桂右, 小野順貴, 植松尚, 大室仲,“イベント遷移を考慮した音響トピックモデルによる欠損を含む観測からの音響シーン推定,” 日本音響学会2014年秋季研究発表会, pp. 1531–1534, 2014.

## 特許

[P.1] 井本桂右, 植松尚, 大室仲, 小野順貴, “生成モデル作成装置，推定装置，それらの方法およびプログラム,” 特開2016–042123.