博 士 論 文

多様な音を柔軟に生成可能とする統計的環境音合成の研究
(Research on Statistical Environmental Sound Synthesis for Generating Diverse Sounds)

2024 年 3 月

立命館大学大学院 情報理工学研究科
情報理工学専攻 博士課程後期課程

岡本 悠希

立命館大学審査博士論文

多様な音を柔軟に生成可能とする統計的環境音合成の研究

(Research on Statistical Environmental Sound Synthesis for Generating Diverse Sounds)

2024 年 3 月

March 2024

立命館大学大学院 情報理工学研究科
情報理工学専攻 博士課程後期課程

Doctoral Program in Advanced
Information Science and Engineering
Graduate School of Information Science and Engineering
Ritsumeikan University

岡本 悠希

OKAMOTO Yuki

研究指導教員 : 山下 洋一 教授

Supervisor : Professor YAMASHITA Yoichi

謝辞

本研究を遂行するにあたり，常に温かい御指導，御鞭撻をいただいた立命館大学情報理工学部 山下洋一 教授に深く感謝致します。立命館大学情報理工学部 福森隆寛講師には，数多くの御指導と御助言を頂きましたことを深く感謝いたします。

本研究を進めるにあたり，同志社大学 井本桂右准教授には，本研究に関するアイデアや実験方法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。所属が立命館大学から同志社大学へ変わられた後も，継続的にサポートと様々な助言を頂きました。東京大学 高道慎之介講師には，音声合成・変換の観点から本研究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。関西大学 山西良典准教授には，日頃より本研究の内容についてご討論頂き，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，日頃より本研究の内容についてご討論頂いた音声言語研究室の諸氏に御礼申し上げます。

博士課程 3 年間の半分の期間にあたる 1 年半の間，日立製作所研究開発グループにてインターンシップでお世話になりました。インターンシップにてお世話になった川口洋平博士，堀口翔太博士，山本正明博士，土肥宏太氏に深く感謝いたします。

最後に，日頃より多数の支援により私を支えてくれた家族に深く感謝します。

概要

本論文では，音声や楽音に限らないあらゆる音である環境音を対象とした音の合成技術を扱う。映画やアニメ，舞台などのコンテンツ制作において，「雨音」や「風の音」などといった環境音は状況の説明や登場人物の心象の象徴など，非常に重要な役割を担う。従来，音声や楽音に対しては，新たな音を生成するための合成・変換技術に関する検討が数多く行われてきた。一方，環境音を対象とした合成・変換技術に関する研究は極めて少ない。また，様々な環境音の特徴をどのようにすれば制御可能であるかについては明らかになっていない。本論文では，多様な環境音の特徴を柔軟に制御するため，様々な入力情報を用いた統計的手法に基づく環境音合成手法を提案する。多様な音を柔軟に制御可能とする環境音合成技術の実現は，映画や動画コンテンツ，ゲームなどの背景音，効果音作成，環境音分析のためのデータ拡張など様々な用途への応用が期待できる。

本研究ではまず，環境音に対してオノマトペを付与した環境音合成のためのデータセット構築を行う。統計的手法によって環境音を合成するにあたり，合成モデルの学習に使用するデータセットの整備は非常に重要な課題である。また，既に公開されているデータセットでは，環境音に対して音響イベントラベルが付与されたデータは多く存在する一方，その他のラベル情報が付与された環境音データセットは極めて少ない。そこで本論文では，環境音の時間的特徴を表現するために有効であるとされるオノマトペを環境音に付与したデータセットを構築する。オノマトペ付与の際には，オノマトペを記述した本人による自信度のスコアと，他者からのオノマトペに対する許容度のスコアを収集する。収集した自信度と許容度のスコアを分析することで，収集されたオノマトペが環境音に対して適切であることを示す。

続いて，環境音合成のための評価方法の整理を行い，合成された環境音をどのように評価するべきかについて検討する。本論文では，環境音合成のための主観評価手法の提案を行う。評価実験にて，環境音合成で使用される主観評価と客観評価手法による結果を比較して，合成された環境音をどのように評価するべきかについて示す。

さらに本論文では，様々な入力情報を用いた環境音合成手法を提案する。まず，「雨音」や「風の音」といった音響イベントラベルから環境音を合成するための手法を提案する。音響イベントラベルでは，生成したい環境音全体の印象を表現す

ることが可能である。そのため，音響イベントラベルを環境音合成の入力情報として使用することで，合成音の音源の種類の制御が期待できる。合成音の品質評価実験により，入力情報として音響イベントラベルを利用することが有効であることを示す。また，半数近くの音響イベントにおいてはデータセットに含まれる自然音と同程度の自然性を獲得できることを示す。

続いて，本論文にて構築したデータセットを用いて，オノマトペからの環境音合成手法を提案する。オノマトペは音の時間的変化を表現するために有効であるため，環境音合成の入力情報として利用することで，合成音の繰り返し回数などといった時間的変化の制御が期待できる。また，オノマトペと音響イベントラベルを同時に環境音合成の入力情報として用いることで，合成音の時間的変化と音源の種類の同時制御を行う。合成音に対する評価実験より，オノマトペを利用することが合成音の時間的変化の制御に有効であることを示す。また，音響イベントラベルとオノマトペを同時に入力することで，音響イベントラベルのみを入力する場合よりも多様な環境音が生成可能であることを示す。

最後に，環境音を模倣した音声を用いた環境音合成手法を提案する。環境音の音高やリズムを表現する方法として，人の声による環境音の模倣が挙げられる。声による環境音の模倣は，直感的に環境音の音高やリズムを表現することができる。そのため，環境音を模倣した音声（模倣音声）を環境音合成に利用することで，合成音の音高やリズムの制御が期待できる。合成音の評価実験より，環境音を模倣した音声が合成音の音高とリズムの制御に有効であることを示す。また，入力に使用する模倣音声の音高やリズムを変化させた場合，それに追従して合成音の音高とリズムが変化することも評価実験にて示す。

Abstract

This thesis deals with sound synthesis method for environmental sounds, which are any sounds that are not limited to speech or music. Environmental sounds such as "sound of wind" play an extremely important role in the production of contents such as movies and cartoon animations, as they explain the situation and symbolize the characters' mental images. In the past, many studies have been conducted on method of synthesis and conversion for generating new sounds for speech and music. On the other hand, there have been very few studies on method of synthesis and conversion for environmental sounds. It is also unclear how the characteristics of diverse environmental sounds can be controlled. This thesis proposes methods of environmental sound synthesis based on a statistical method using various input informations in order to flexibly control the characteristics of diverse environmental sounds. The realization of an method of environmental sound synthesis that enables flexible control of diverse sounds is expected to be applied to various applications, such as the creation of background sounds and sound effects for movies, video content, games, etc., and data augmentation for environmental sound analysis.

The thesis constructs a dataset for environmental sound synthesis from onomatopoeic words corresponding to environmental sounds. It is very important to construct a dataset to be used for training of environmental sound synthesis model using statistical approach. In addition, while there are many datasets with sound event labels for environmental sounds, there are very few environmental sound datasets with other label information. Therefore, this thesis constructs a dataset in which environmental sounds are assigned onomatopoeic words, which is considered to be effective in expressing the temporal-change characteristics of environmental sounds. When assigning onomatopoeic words, this thesis collects the score of confidence level by the person who described the onomatopoeic words and the score of acceptance level for the onomatopoeic words from others. By analyzing the collected confidence and acceptance scores, this thesis shows that the collected onomatopoeic words is appropriately assigned to the environmental sound.

The thesis also organizes the evaluation methods for environmental sound syn-

thesis and discusses how the generated environmental sounds should be evaluated. This thesis proposes subjective evaluation methods for environmental sound synthesis. Through evaluation experiments, this thesis compares the subjective and objective evaluation methods used in environmental sound synthesis, and indicates how the synthesized environmental sounds should be evaluated.

Furthermore, this thesis discusses methods for environmental sound synthesis using various input informations. First, this thesis proposes a method of environmental sound synthesis from sound event labels such as sound of "rain" and "wind." Sound event labels can express the sound events of the environmental sound to be generated. Therefore, by using sound event labels as input information for environmental sound synthesis, it can be expected to control the sound event of the synthesized sounds. From the experimental evaluation of the quality of the synthesized sound shows that the use of sound event labels as input information is effective. This thesis also shows that nearly half of the sound events have the same level of naturalness as the natural sounds in the dataset.

Second, this thesis proposes a method of environmental sound synthesis from onomatopoeic words using the dataset constructed in this thesis. Since onomatopoeic words are effective for expressing the temporal-change characteristics of sounds, it can be expected to control temporal-change characteristics such as the number of repetitions of synthesized sounds by using onomatopoeic words as input information for environmental sound synthesis. In addition, by using onomatopoeic words and sound event labels simultaneously as input information for environmental sound synthesis, the temporal-change characteristics of the synthesized sound and the sound event of the synthesized sound can be controlled simultaneously. From the evaluation experiments on the synthesized sounds, this thesis shows that the use of onomatopoeic words is effective in controlling the temporal-change features of the synthesized sounds. This thesis also shows that the simultaneous input of sound event labels and onomatopoeic words can generate a wider diversity of environmental sounds than when only sound event labels are input.

Finally, this thesis proposes a method of environmental sound synthesis using vocal imitation that imitating environmental sounds using. One of the methods to express the pitch and rhythm of environmental sounds is to imitate environmental sounds by human voices. Vocal imitations can intuitively express the pitch and rhythm of environmental sounds. Therefore, it is possible to control the pitch

and rhythm of the synthesized sound by using a vocal imitation that imitates the environmental sound for environmental sound synthesis. From the evaluation experiment of the synthesized sound, this thesis shows that the vocal imitation of the environmental sound is effective in controlling the pitch and rhythm of the synthesized sound. In addition, when the pitch and rhythm of the vocal imitations used for input is changed, the pitch and rhythm of the synthesized sound changes in accordance with the change in pitch and rhythm.

目 次

第 1 章　序論 **1**
1.1　研究背景・目的 1
1.2　本論文の構成 3

第 2 章　統計的環境音合成と本研究の着眼点 **4**
2.1　はじめに 4
2.2　本論文で用いる用語の定義 4
2.3　統計的環境音合成 5
2.3.1　問題の定式化 5
2.3.2　統計的環境音合成において用いられる深層学習の技術 5
2.4　環境音合成における関連研究と問題設定 9
2.4.1　音響シーンを対象とした環境音合成 9
2.4.2　環境音変換 10
2.4.3　マルチモーダル環境音合成 11
2.5　本研究の方針 11

第 3 章　環境音合成のための
データセット構築 **12**
3.1　はじめに 12
3.2　オノマトペデータセットの構築 13
3.2.1　RWCP-SSD の概要 13
3.2.2　構築したデータセットの概要 13
3.2.3　相互相関による音響イベントクラス内のクラスタリング . . 14
3.2.4　クラウドソーシングによるオノマトペおよび自信度, 許容度
の付与 15
3.3　付与されたオノマトペの分析 16
3.3.1　付与されたオノマトペの結果 16
3.3.2　付与された自信度と許容度の分析 17
3.4　3 章のまとめ 19

第 4 章　環境音合成のための評価手法　21
4.1　はじめに 21
4.2　環境音合成における従来の評価手法 21
4.2.1　従来の客観評価手法 21
4.2.2　従来の主観評価手法 22
4.3　環境音合成のための主観評価手法の提案 23
4.3.1　音響イベントラベルから合成された環境音の主観評価手法 . 23
4.3.2　オノマトペから合成された環境音の主観評価手法 24
4.4　評価実験 24
4.4.1　実験条件 25
4.4.2　音響イベントラベルから合成された環境音の主観評価結果 . 27
4.4.3　オノマトペから合成された環境音の主観評価結果 27
4.4.4　音響イベント分類器による合成音の客観評価結果 29
4.4.5　主観評価と客観評価結果の比較 30
4.5　4 章のまとめ 31

第 5 章　音響イベントラベルからの環境音合成　32
5.1　はじめに 32
5.2　統計的手法による音響イベントラベルからの環境音合成 32
5.2.1　提案手法の概要 32
5.2.2　音響イベントラベルを用いた WaveNet によるモデル構築 . . 33
5.3　評価実験 35
5.3.1　実験条件 35
5.3.2　合成された環境音の了解性に関する評価 37
5.3.3　実在する音としての自然性の評価 38
5.3.4　環境音としての自然性の評価 39
5.4　5 章のまとめ 40

第 6 章　オノマトペからの環境音合成　42
6.1　はじめに 42
6.2　統計的手法によるオノマトペからの環境音合成手法 43
6.2.1　提案手法の概要 43
6.2.2　オノマトペのみを入力とする音響モデルの構築 44
6.2.3　オノマトペと音響イベントラベルを入力とする音響モデルの構築 44
6.3　評価実験 47

6.3.1 実験条件 47
6.3.2 環境音の自然性に関する評価 48
6.3.3 環境音の多様性に関する評価 51
6.4 6 章のまとめ 55

第 7 章 環境音を模倣した音声を用いた環境音合成 58
7.1 はじめに 58
7.2 関連研究 59
7.3 統計的手法による模倣音声を利用した環境音合成 59
7.4 模倣音声データセットの構築 60
7.5 評価実験 62
7.5.1 実験条件 62
7.5.2 合成音の自然性の評価 63
7.5.3 入力音声の音高に対する合成音の妥当性の評価 65
7.5.4 入力音声のリズムに対する合成音の妥当性の評価 67
7.5.5 入力音声の音高変化による合成音の評価 67
7.5.6 入力音声のリズム変化による合成音の評価 69
7.5.7 スペクトログラムによる合成音の変化の確認 70
7.6 7 章のまとめ 71

第 8 章 結論 72

付 録 A Transformer を用いたオノマトペからの環境音合成 74
A.1 はじめに 74
A.2 Transfomer を用いたオノマトペからの環境音合成手法 74
A.3 評価実験 76
A.3.1 実験条件 76
A.3.2 オノマトペに対する環境音の評価 77
A.3.3 環境音の品質に関する評価 79
A.4 付録 A のまとめ 80

付 録 B オノマトペを用いた環境音抽出 82
B.1 はじめに 82
B.2 関連研究 83
B.3 オノマトペを用いた環境音抽出手法 84
B.3.1 オノマトペを用いた環境音抽出の概要 84
B.3.2 オノマトペを用いた環境音抽出手法の提案 84

B.4 評価実験 86
B.4.1 学習・評価用データの作成 86
B.4.2 実験条件 87
B.4.3 実験結果 88
B.5 付録 B のまとめ 89

付 録 C 単一音源に説明文を付与した環境音データセットの構築 90
C.1 はじめに 90
C.2 データセットの構築 91
C.2.1 構築したデータセットの概要 91
C.2.2 環境音の収録環境と条件 93
C.2.3 収録した音に対する説明文の収集 93
C.2.4 収集した音の説明文の評価 94
C.2.5 収集したデータの分割 95
C.3 評価実験 95
C.3.1 音の説明文に基づく環境音抽出のモデル構造 95
C.3.2 学習・評価用データの作成 96
C.3.3 実験条件 97
C.3.4 実験結果 99
C.4 付録 C のまとめ 99

参考文献 100

本論文に関連する研究業績 111

その他の研究業績 114

図 目 次

1.1 本論文の構成 2

2.1 本論文で用いる用語の概念図 4
2.2 音響シーンを対象とした環境音合成における問題設定 9
2.3 環境音変換の問題設定 10

3.1 オノマトペを用いた環境音合成の概要 13
3.2 RWCP-SSD に含まれる笛の音のスペクトログラム 15
3.3 音響イベントごとの自信度/許容度の平均 17
3.4 オノマトペに対する自信度/許容度の平均の散布図 18
3.5 音響イベントごとの自信度の平均 19
3.6 音響イベントごとの許容度の平均 20

4.1 音響イベント分類器を用いた環境音の客観評価の概要 22
4.2 音響イベントラベルに対する環境音の妥当性の絶対評価結果 (音と音響イベントラベルを提示) 27
4.3 音響イベントラベルに対する環境音の妥当性の相対評価結果 28
4.4 オノマトペに対する環境音の妥当性の絶対評価結果 29
4.5 オノマトペに対する環境音の妥当性の相対評価結果 29
4.6 音響イベント分類器による合成音の分類結果 30
4.7 主観評価と客観評価の比較結果 31

5.1 統計的手法による音響イベントラベルを用いた環境音合成の概要 . . 33
5.2 WaveNet を用いた環境音合成モデルの構築 34
5.3 自然音に対して被験者が回答した音響イベントラベルの正解率 . . . 37
5.4 合成音に対して被験者が回答した音響イベントラベルの正解率 . . . 38
5.5 自然音と合成音のスペクトログラム 39
5.6 各音響イベントの音に対して被験者が自然音を認識した割合 40
5.7 自然音と合成音に対する自然性のスコア平均 41

6.1 統計的手法によるオノマトペからの環境音合成の概要 43
6.2 オノマトペのみを入力とするモデル学習 45
6.3 オノマトペと音響イベントラベルを入力とするモデル学習 46
6.4 オノマトペに対する環境音の許容度に関する評価結果 49
6.5 オノマトペに対する環境音の表現性に関する評価結果 50
6.6 オノマトペのみを用いた提案手法によって合成した環境音のスペクトログラム 51
6.7 KanaWave 並びにオノマトペと音響イベントラベルを同時入力とする提案手法による合成音のスペクトログラム 52
6.8 環境音に対する自然性の評価結果 53
6.9 音響イベントクラス内における環境音の多様性の評価結果 54
6.10 同一オノマトペから合成された環境音の多様性の評価結果 56
6.11 ビイッ (/ ー b i: i q /) というオノマトペを各手法において入力した際の合成音のスペクトログラム 57

7.1 音響イベントラベルと模倣音声を入力とする環境音合成の概要 . . . 60
7.2 “Label+vocal” と “Label” の評価スコアの分布 65
7.3 入力音声の音高を変化させた場合の合成音の主観/客観評価結果 . . 69
7.4 入力音声のリズムを変化させた場合の合成音の主観/客観評価結果 . 69
7.5 入力音声の音高とリズムを変化させた場合の合成音のスペクトログラム 70

A.1 Transformer によるオノマトペからの環境音合成の概要 75
A.2 オノマトペに対する環境音の許容度の評価結果 77
A.3 オノマトペに対する環境音の表現性の評価結果 78
A.4 データセットに含まれる自然と各手法によるオノマトペからの合成音のスペクトログラム 79
A.5 環境音の全体的な印象に関する評価結果 80
A.6 環境音の自然性に関する評価結果 81

B.1 オノマトペを用いた環境音抽出の概要 83
B.2 オノマトペを用いた環境音抽出のモデル構造 84
B.3 オノマトペを用いた環境音抽出によって抽出された環境音のスペクトログラム 89

C.1 音の説明文に対して収集した妥当性スコアのヒストグラム 94
C.2 音の説明文に基づく環境音抽出のモデル概要 95

C.3 音の説明文に基づく環境音抽出によって抽出された環境音のスペクトログラム 98

表 目 次

2.1 各入力情報によって制御が期待できる環境音の特徴 11

3.1 付与されたオノマトペの例 16

4.1 各主観/客観評価に使用した環境音合成手法の一覧 25
4.2 各主観評価実験において被験者に提示した情報 25
4.3 評価実験で使用した音響イベントと学習・評価サンプル数 25
4.4 各合成手法のパラメータ設定及び使用した音響特徴量 26
4.5 音響イベント分類器のパラメータ設定および使用した音響特徴量 . . 26

5.1 WaveNet のパラメータ設定 36

6.1 seq2seq のパラメータ設定および利用した音響特徴量 47
6.2 各評価実験に使用した合成音のサンプル数と被験者数 47
6.3 各実験にて評価した環境音合成手法の一覧 48

7.1 実験に使用した ESC-50 データセットの音響イベントとサンプル数 . 61
7.2 環境音を模倣した音声からの環境音合成の実験条件 63
7.3 合成音の自然性の評価結果 64
7.4 入力音声の音高に対する合成音の妥当性の評価結果 66
7.5 入力音声のリズムに対する合成音の妥当性の評価結果 68

A.1 Transformer を用いたオノマトペからの環境音合成の実験条件 . . . 76

B.1 オノマトペを用いた環境音抽出で使用した各音響イベントクラスの superclass と subclass 86
B.2 オノマトペを用いた環境音抽出の実験条件 87
B.3 オノマトペを用いた環境音抽出の SDRi [dB] による評価結果 88

C.1 構築したデータセットに含まれる音の説明文の例 91
C.2 収録した音響イベントクラス 92

C.3 環境音の収録に使用した機材 93
C.4 構築したデータセットにおける学習，検証，評価セットの統計的情報 95
C.5 音の説明文に基づく環境音抽出の実験条件 97
C.6 音の説明文に基づく環境音抽出の SDRi [dB] による評価結果 98

第1章　序論

1.1　研究背景・目的

映画やアニメ，舞台などのコンテンツ制作において，「雨の音」や「風の音」などといった環境音は非常に重要な役割を担っている。環境音は状況の説明，登場人物の心象を象徴させるなどの効果があり，コンテンツ制作においては欠かせない存在である。現状，これらの環境音は大量のデータベースから目的とする音を探しての利用や，実際に環境音を収録して，編集・再生するのが一般的である。しかしながら，大量のデータベースから目的の音を探すには膨大な時間を要する。また，目的とする音がデータベースに存在しない場合もある。さらに，環境音の再編集には経験や編集者の技量が求められるため容易ではない。このような課題があるなか，近年では YouTube などの動画投稿サイトの普及に伴い，誰でも自作の映画やアニメといったコンテンツを制作・発信できるようになり，ますます環境音を利用する場面が増加している。

従来, 音声や楽音といった音に対しては, 新たな音を生成するための合成・変換技術に関する検討が行われてきた。特に音声合成技術においては, 2016 年の WaveNet [1] の登場以来，自然音と合成音の区別がつかないほどの品質に近づきつつあり，飛躍的に進歩した。しかしながら，音声や楽音に限らないあらゆる音である環境音を対象とした合成・変換技術に関する研究は極めて少ない。環境音合成の実現は映画や動画コンテンツ，ゲームなどの背景音，効果音作成 [2]，バーチャルリアリティコンテンツの作成補助 [3]，環境音分析のためのデータ拡張 [4] など様々な用途への応用が考えられる。そこで，既存の音のみならず，自由に新たな音を生成可能とする環境音合成技術を実現することができれば，誰でも目的とする環境音を容易に得ることができ，コンテンツ制作の大きな手助けとなる。

これまでの環境音合成の取り組み例として，音響シーン（電車の中，会議中など）と呼ばれる音が収録された状況を表すラベル情報を用いた手法の提案が行われてきた [5]。従来の研究は音響シーン全体の様子を音波形として再現することを目指した研究である。そのため，音響シーンを表した音波形に含まれる個々の音に関する再現精度は高いとは言えない。特に，メディアコンテンツなどに用いる環境音は作品の印象などを決定付けるための重要な情報であり，個々の音に対し

ても高い再現精度が求められる。そこで本研究では，環境音の中でも「雨の音」，「風の音」といった個々の音を柔軟に合成可能とする手法の実現を目的とし，「音響イベントラベルを入力とする環境音合成」，「オノマトペを入力とする環境音合成」，「環境音を模倣した音声を利用した環境音合成」の 3 つの環境音合成手法を提案する。また，上述したように，環境音合成の研究は検討例が少なく，使用可能なデータセットの整備並びに，合成された環境音をどのように評価するべきであるかについての議論が不十分である。そこで，本研究において，合成された環境音をどのように評価するべきであるかの議論と環境音合成のためのデータセット構築も行う。

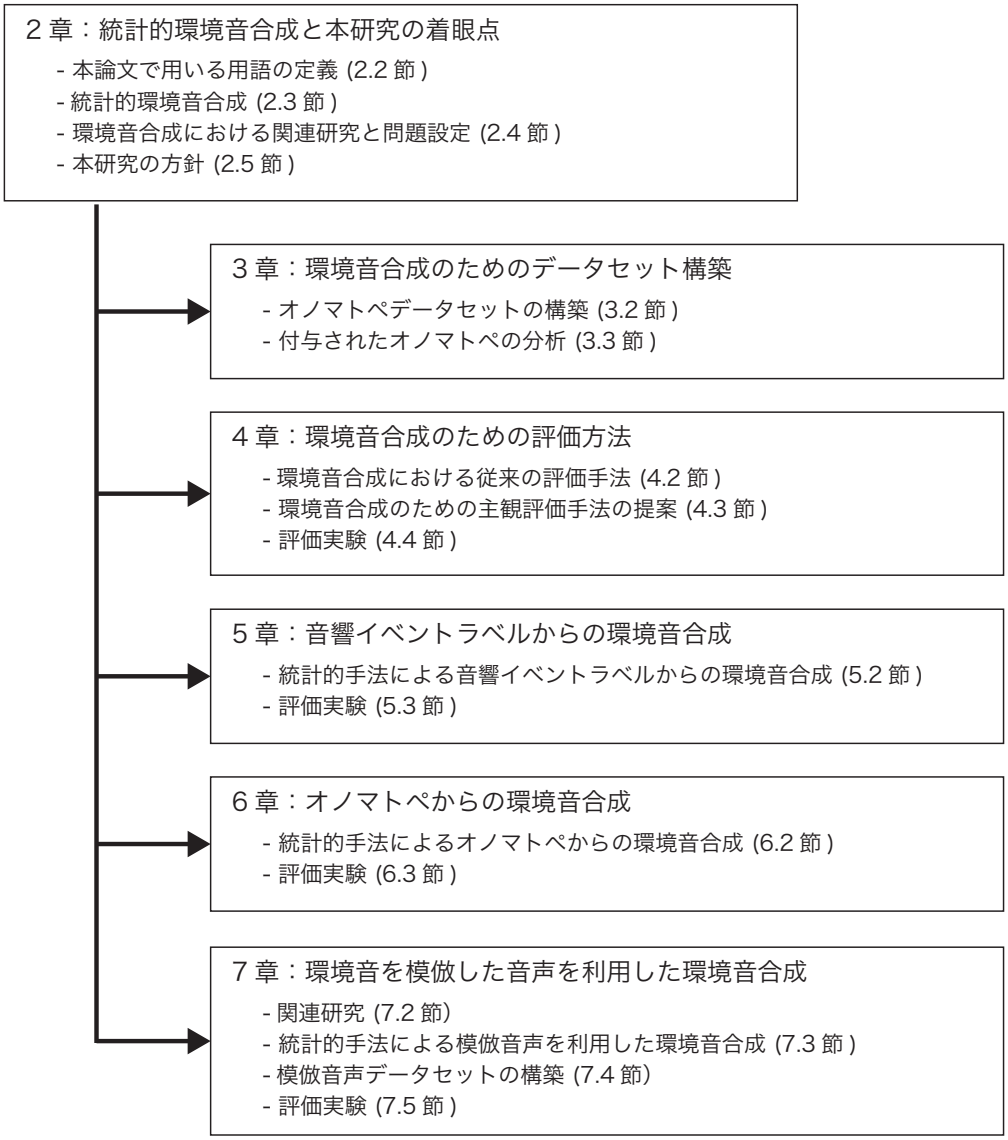


図 1.1: 本論文の構成

1.2　本論文の構成

本論文の構成を図 1.1 に示す。2 章では，まず環境音合成で用いられる用語，統計的環境音合成の定式化，問題設定の整理を行う。その後，本研究で取り扱う個々の音を対象とした環境音合成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環境音合成の着眼点および方針について述べる。3 章では，6 章で提案するオノマトペを入力とした環境音合成手法を実現するため，環境音にオノマトペを対応づけた環境音合成のためのデータセット構築について述べる。4 章では，環境音合成のための評価方法について整理を行い，合成された環境音をどのように評価するべきであるかについて検討する。5 章では，音響イベントラベルと呼ばれる音響イベントの種類（風の音，雨音などといった音の種類）を入力とする環境音合成手法について提案して，提案手法により個々の音の合成が実現できたことを示す。6 章では，3 章にて構築したデータセットを用いて，オノマトペから環境音を合成する手法を提案して，環境音の繰り返し回数などといった時間的な変化を制御可能であることを示す。また，オノマトペと同時に音響イベントラベルを入力することで，より柔軟に合成音を制御するための手法の提案も行う。7 章では，環境音を模倣した音声を利用した環境音合成手法を提案して，音声を入力とすることで，合成音のリズムや音高を制御可能であることを示す。最後に，8 章で本論文の結論を述べる。

第2章　統計的環境音合成と本研究の着眼点

2.1　はじめに

本章ではまず，2.2 節にて環境音合成に関連する用語の定義を行う。その後，2.3 節にて統計的環境音合成の定式化並びに，用いる深層学習の技術について述べる。また，2.4 節では環境音合成における関連研究の問題設定の整理を行う。最後に，従来研究の課題と本論文で提案する環境音合成手法の着眼点について述べる。

2.2　本論文で用いる用語の定義

本論文で用いる用語の概念図を図 2.1 に示す。図に示すように，音の収録された場所や状況のことを音響シーンと呼び，それに含まれる個々の音の種類のことを音響イベントと呼ぶ。本研究においては，個々の音である音響イベントを表す音波形の合成に取り組む。

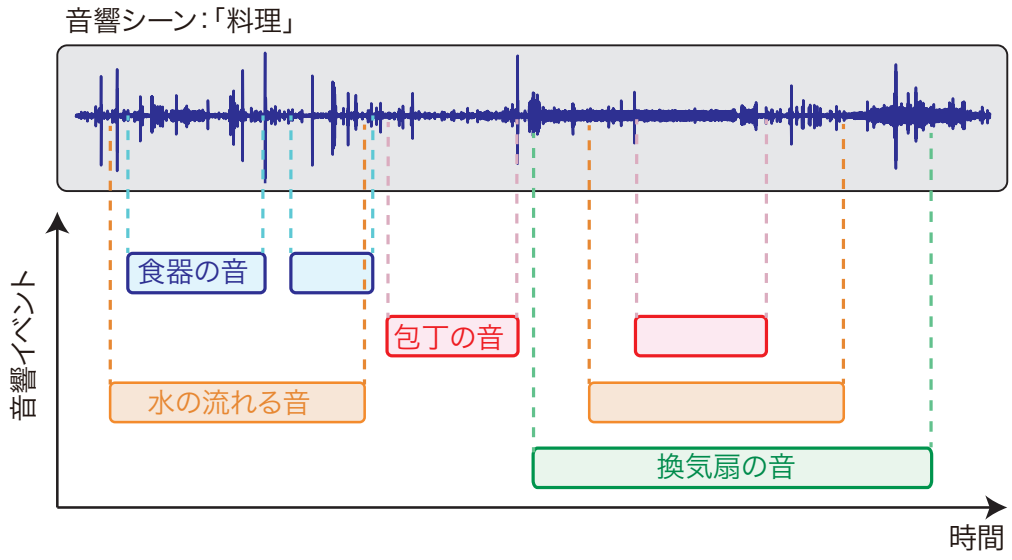


図 2.1: 本論文で用いる用語の概念図

2.3 統計的環境音合成

2.3.1 問題の定式化

統計的環境音合成は，式 (2.1) に示すように，入力特徴量の集合 $\boldsymbol{X}$，環境音波形の集合 $\boldsymbol{Y}$，未知の入力特徴量 $\boldsymbol{x}$ が与えられた際の環境音の波形 $\boldsymbol{y}$ に対する確率分布からのサンプリング問題とみなすことができる。

$$\boldsymbol{y} \sim P(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{X}, \boldsymbol{Y}) \tag{2.1}$$

このように，予測分布 $P(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{X}, \boldsymbol{Y})$ を求めることで，その分布からサンプリングすることによって環境音の波形を得ることができる。しかしながら，予測分布を直接計算することは困難であるため，モデルパラメータ $\boldsymbol{\lambda}$ を持つ統計モデルを導入することで，予測分布の計算を可能とする。モデルパラメータを導入することで，予測分布 $P(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{X}, \boldsymbol{Y})$ を式 (2.2) のように近似可能である。

$$P(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{X}, \boldsymbol{Y}) \approx P(\boldsymbol{y} \mid \boldsymbol{x}, \hat{\boldsymbol{\lambda}}) \tag{2.2}$$

$\hat{\boldsymbol{\lambda}}$ は $\boldsymbol{X}$, $\boldsymbol{Y}$ から推定されたモデルパラメータを表す。これらの近似によって，統計的環境音合成は以下の式で表現可能である。

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}} P(\boldsymbol{\lambda} \mid \boldsymbol{X}, \boldsymbol{Y}) \tag{2.3}$$

$$\boldsymbol{y} \sim P(\boldsymbol{y} \mid \boldsymbol{x}, \hat{\boldsymbol{\lambda}}) \tag{2.4}$$

なお環境音合成では，式中の入力特徴量の集合 $\boldsymbol{X}$ に様々な特徴量が使用される。環境音合成において想定される入力特徴量並びに関連研究に関しては 2.4 節にて述べる。

2.3.2 統計的環境音合成において用いられる深層学習の技術

本論文では，深層学習を用いて式 (2.3) にて説明したモデルパラメータの推定並びに環境音波形の合成を行う。そこで本節において，本論文で使用する深層学習の技術について述べる。

フィードフォワードニューラルネットワーク

フィードフォワードニューラルネットワーク (FFN: Feed-Forward Neural Network) [6] は，深層学習において最も基本的なネットワーク構造であり，入力ベクトルに対して線形変換と非線形変換を繰り返す一方向のみのニューラルネットワークとなっている。

$$z_j^{(i)} = \sum_k W_{j,k}^i h_k^{(i-1)} + b_j \tag{2.5}$$

$$h_j^i = \sigma(z_j^{(i)}) \tag{2.6}$$

FFN は式 (2.5)，(2.6) のように定式化され，$W_{j,k}^i$ は i 層におけるニューロン j と $i-1$ 層におけるニューロン k 間の重みパラメータ，$h_k^{(i-1)}$ は $i-1$ 層のニューロン k の出力，b_j は学習可能なバイアスを示す，σ は活性化関数と呼ばれる非線形変換を表しており，一般的にはシグモイド関数や双曲線正接関数，正規化線形関数 (ReLU) が用いられることが多い。

畳み込みニューラルネットワーク

畳み込みニューラルネットワーク (CNN: Convolutional Neural Network) [7] は，入力情報に対してフィルタによる重みを掛け合わせる畳み込み層, 最大値や平均値といった条件に従ってある範囲の数値を取り出すプーリング層から構成されるニューラルネットワークである。入力特徴量 $\boldsymbol{X} = \{x_{mn} \mid 0 \leq m < M, 0 \leq n < N\}$，畳み込みフィルタ（カーネル）のサイズが $P \times Q$ における畳み込み層における処理を以下に示す。

$$a_{mnc}^{(l)} = \sigma\left(\sum_{p=0}^{P-1}\sum_{q=0}^{Q-1} w_{pqc} x_{(m+p)(n+q)}^{(l-1)} + b_{pqc}\right) \tag{2.7}$$

ここで，l，w_{pqc}，b_{pqc} はそれぞれ層のインデックス，c 番目におけるカーネルの重み，学習可能なバイアスを表す。畳み込み層から出力された $a_{mnc}^{(l)}$ を最大プーリング層もしくは平均プーリング層に通すことによってカーネルごとに数値を取り出す。なお，プーリング層に関しては，扱う入力データの特性に応じて，最大プーリングもしくは平均プーリングのどちらかを使用する。プーリングを行うことによって，多次元の特徴量を圧縮しながら特徴量の抽出を行うことが可能である。

再帰的ニューラルネットワーク

再帰的ニューラルネットワーク (RNN: Recurrent Neural Network) [8] は，FNN や CNN のような一方向のみのニューラルネットワークとは異なり，内部に循環構

造を持つニューラルネットワークである。RNN は系列データを扱うことに長けており，FNN や CNN とは異なり，現在の時刻より前の時刻の内部状態（隠れ状態）も用いてモデル学習を行うことが可能である。そのため，時系列情報を伴う環境音に対しても頻繁に使用される深層学習モデルである。RNN の処理を式 (2.8) に示す。なお，時間インデックスを t，入力特徴量を $\boldsymbol{V} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_T\}$，隠れ状態を $\boldsymbol{H} = \{\boldsymbol{h}_1, \ldots, \boldsymbol{h}_T\}$ とする。

$$\boldsymbol{h}_t = \sigma\left(\boldsymbol{W}^{(v)}\boldsymbol{v}_t + \boldsymbol{W}^{(h)}\boldsymbol{h}_{t-1} + \boldsymbol{b}\right) \tag{2.8}$$

ここで，$\boldsymbol{W}^{(v)}$，$\boldsymbol{W}^{(h)}$，$\boldsymbol{b}$ はそれぞれ入力特徴量に対する重み，隠れ状態に対する重み，学習可能なバイアスとする。なお式中の活性化関数は，環境音合成などの回帰問題においては恒等関数を利用する。RNN は過去の情報を利用できる一方，比較的長い系列長のデータに対しては，FNN や CNN と比べて重みと勾配が乗算される回数が多いため，勾配消失が起こりやすくなる。そのため，長期的な時系列の構造をうまくモデル化できないという課題がある。

長期的な時系列に対応した深層学習モデルとして，長・短期記憶 (LSTM: Long Short-Term Memory) [9] が提案された。LSTM は，関連する情報を選択的に保持して，関連しない情報を忘却するゲート構造を RNN に取り入れることで，勾配消失の問題を解決した。LSTM における処理を式 (2.9)–(2.15) に示す。なお，$\boldsymbol{f}_t$ を忘却ゲート，$\boldsymbol{i}_t$ を入力ゲート，$\boldsymbol{o}_t$ を出力ゲートとする。過去の隠れ状態 $\boldsymbol{h}_{t-1}$，過去の記憶セル $\boldsymbol{C}_{t-1}$，現在の時刻の入力特徴量 $\boldsymbol{x}_t$ を用いて，次の時刻の隠れ状態 $\boldsymbol{h}_t$ と記憶セル $\boldsymbol{C}_t$ を計算する。

$$\hat{\boldsymbol{h}}_{t-1} = \mathsf{Concat}\left(\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right) \tag{2.9}$$

$$\boldsymbol{f}_t = \sigma\left(\boldsymbol{W}^{(f)}\hat{\boldsymbol{h}}_{t-1} + \boldsymbol{b}^{(f)}\right) \tag{2.10}$$

$$\boldsymbol{i}_t = \sigma\left(\boldsymbol{W}^{(i)}\hat{\boldsymbol{h}}_{t-1} + \boldsymbol{b}^{(i)}\right) \tag{2.11}$$

$$\boldsymbol{g}_t = \mathsf{tanh}\left(\boldsymbol{W}^{(g)}\hat{\boldsymbol{h}}_{t-1} + \boldsymbol{b}^{(g)}\right) \tag{2.12}$$

$$\boldsymbol{C}_t = \boldsymbol{C}_{t-1} \odot \boldsymbol{f}_t + \boldsymbol{i}_t \odot \boldsymbol{g}_t \tag{2.13}$$

$$\boldsymbol{o}_t = \sigma\left(\boldsymbol{W}^{(o)}\hat{\boldsymbol{h}}_{t-1} + \boldsymbol{b}^{(o)}\right) \tag{2.14}$$

$$\boldsymbol{h}_t = \mathsf{tanh}\left(\boldsymbol{C}_t\right) \odot \boldsymbol{o}_t \tag{2.15}$$

ここで，$\mathsf{Concat}(\cdot)$，$\odot$，tanh はそれぞれベクトル同士の連結，アダマール積，双曲線正接関数を表す。$\boldsymbol{W}^{(*)}$，$\boldsymbol{b}^{(*)}$ はそれぞれ学習可能な重みパラメータ，バイアスを表す。忘却ゲートによって，過去のセル状態からどのぐらいの割合で情報を使

用するかを決定する。入力ゲートでは，現時刻における情報をどの程度長期記憶セルに保存するかを決定する。そして最後に出力ゲートによって，現時刻の情報のうち何を出力するべきかを決定する。

過去, 未来の双方向の隠れ状態を利用する双方向 LSTM (BiLSTM: Bi-directional LSTM) [9] も提案されている。BiLSTM では，現在の隠れ状態の前後の隠れ状態を利用することで，過去と未来の時間的依存性を捕捉することが可能となった。

注意機構

注意機構 (Attention) は, 入力された情報のどこに注目するべきかを動的に算出するモデル構造である。特に，この Attention を導入して高い性能を示している深層学習モデルとして Tansformer [10] がある。そのため，ここでは Transformer にて使用されている Attention について説明する。Transformer では，scaled-dotproduct attention [10] という枠組みが提案されており，クエリ $\boldsymbol{Q}$ とキー $\boldsymbol{K}$ とバリュー $\boldsymbol{V}$ を用いて以下のような式で表される。

$$\mathsf{Attention}\left(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}\right) = \mathsf{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{2.16}$$

ここで，$\boldsymbol{Q}$，$\boldsymbol{K}$，$\boldsymbol{V}$ はそれぞれ，クエリ，キー，バリューの行列を表す。また，$\mathsf{softmax}(\cdot)$, k はそれぞれソフトマックス関数，キーの次元数を表す。クエリとキーの要素毎の内積を算出してソフトマックス関数によって正規化する部分は，クエリとキーの類似度計算とみなすことができる。なお，$\sqrt{d_k}$ にて除算を行うのは，勾配消失を防ぐためである。そして，算出された類似度に対してバリューを乗算することでクエリに関する情報をキーとの関連度から検索して，出力結果に反映させる。特に，クエリー，キー，バリューが同じ行列の場合は自己注意機構（self-attention) と呼ばれる。さらに，注意機構を発展させた multi-head attention [10] という枠組みも提案されている。multi-head attention の処理を以下に示す。

$$\mathsf{MultiHeadAttention}\left(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}\right) = \mathsf{Concat}\left(\boldsymbol{H}_1, \ldots, \boldsymbol{H}_h\right)\boldsymbol{W}^O \tag{2.17}$$

$$\boldsymbol{H}_i = \mathsf{Attention}\left(\boldsymbol{Q}\boldsymbol{W}_i^Q, \boldsymbol{K}\boldsymbol{W}_i^K, \boldsymbol{V}\boldsymbol{W}_i^V\right) \tag{2.18}$$

ここで, $\boldsymbol{W}_i^O$, $\boldsymbol{W}_i^Q$, $\boldsymbol{W}_i^K$, $\boldsymbol{W}_i^V$ はそれぞれ学習可能な重み行列を表す。$\mathsf{Attention}(\cdot)$, h はそれぞれ scaled-dotproduct attention，ヘッド数を表す。multi-head attention は，式 (2.17)，(2.18) に示すように，先ほどの注意機構を複数用意して別々に計算を行うことで，それぞれで異なる特徴を抽出することが期待できる。環境音合成の研究においては，1 秒の比較的系列長の短い信号から，5 秒以上の長い信号を扱

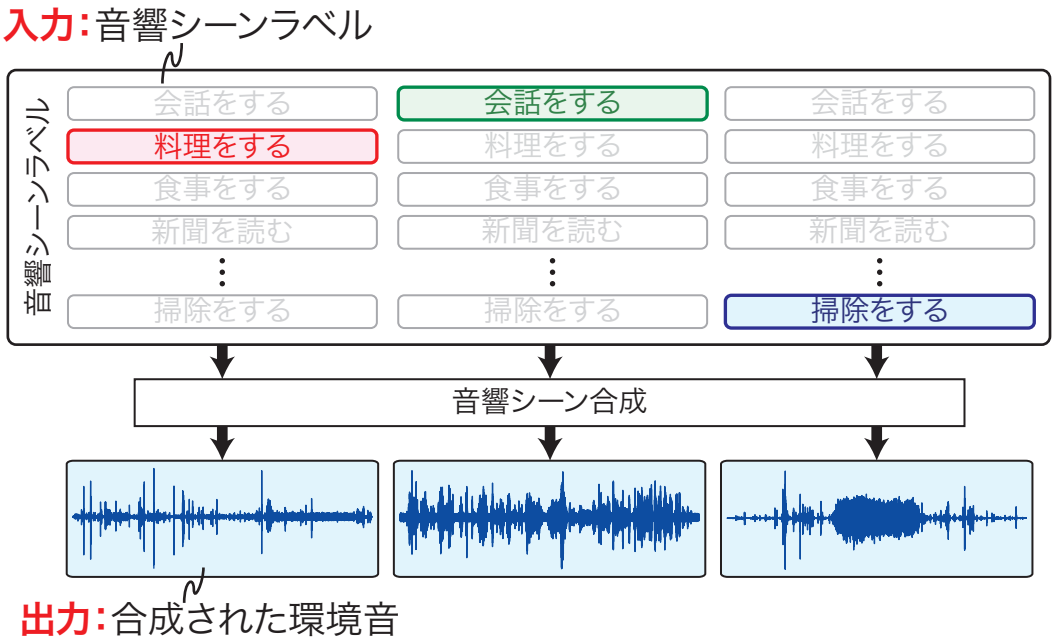


図 2.2: 音響シーンを対象とした環境音合成における問題設定

うこともある。そのため，時系列同士で特に注目を向けるべき点を重みづけることのできるこれらの注意機構の利用は有効であると考えられる。

2.4 環境音合成における関連研究と問題設定

環境音を統計的に合成する取り組みは検討例が少ない。そこで，本論文にて環境音合成において想定される問題設定とそれらに関連する技術について述べる。

2.4.1 音響シーンを対象とした環境音合成

複数の音響イベントを含む音響シーンを表す音波形を合成する取り組み例として，音響シーンラベルを入力とした環境音合成が挙げられる。図 2.2 に音響シーンラベルを入力とした環境音合成の問題設定を示す。例えば，Kong ら [5] は音響シーンラベルを one-hot 表現し，SampleRNN [11] への入力とすることで複数の音響シーンに対応する音波形を出力可能にしている。また，Gontier ら [12] は，CNN を用いて音響シーンを表す音波形を出力とする環境音合成手法を提案している。これらの技術は，深層学習など近年の統計的手法に基づく環境音分析の研究におけるデータ拡張手法として利用することも考えられる。一方，従来の音響シーンを対象とした環境音合成は音響シーン全体を表す音波形の合成を主眼としているた

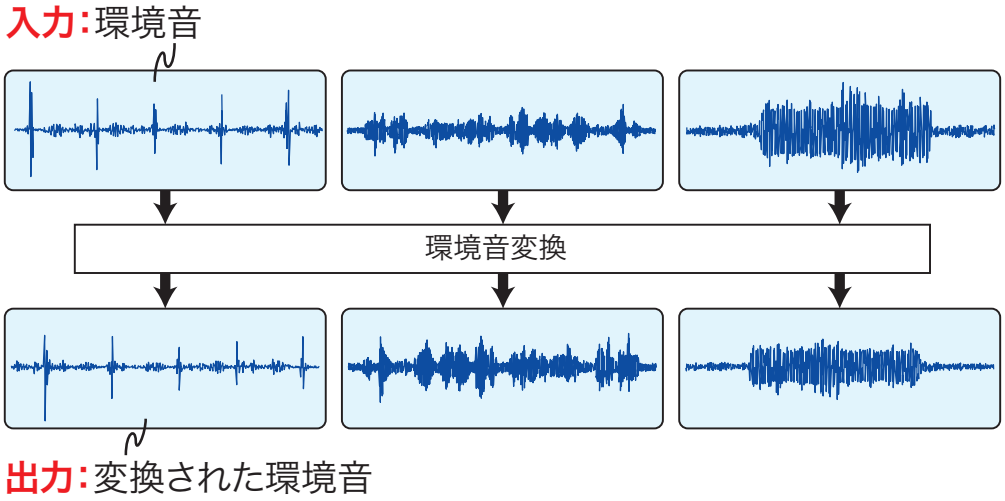


図 2.3: 環境音変換の問題設定

め，合成音の中に含まれる個々の音響イベントの音の再現精度が高くないという課題も残る。

2.4.2 環境音変換

環境音を合成する技術として，図 2.3 に示すような環境音を入力として，新たな環境音を出力する環境音変換の技術が挙げられる。例えば，動画像の制作を行う際，いくつかの背景音や効果音はすでに所持しているが，それらが作成したい動画像に適さない場合がある。その際，環境音の変換技術を利用して，所持している環境音を目的とする環境音に変換して利用することが考えられる。具体的には，車種 A のクラクションの音は所持しているが，車種 B のクラクションを用いた動画を作成したい場合などに環境音変換の技術を利用することで，目的の音を得ることが可能であると考える。

環境音変換の従来研究として，Grinstein ら [13] や Mital[14] らは画像処理で用いられている style transfer[15] という技術を応用した audio style transfer という手法を提案している。これらの手法では，音信号やスペクトログラムをコンテンツ情報（音サンプル内の主となる音）とスタイル情報（音の質感など背景音のような情報）に分けて学習し，スタイル情報のみを変換対象の音に転写させることで環境音変換を実現している。また，環境音を変換する技術として，sound texture synthesis と呼ばれる音の texture と呼ばれる背景音のような情報をターゲットとなる環境音に転写することで，環境音の texture 情報を変換する研究も提案されている [16, 17]。

表 2.1: 各入力情報によって制御が期待できる環境音の特徴

入力情報	音の高さ	リズム	時間的な変化	音源の種類
音響イベントラベル（提案法）				✓
画像				✓
オノマトペ（提案法）			✓	
音の説明文			✓	✓
音声（提案法）	✓	✓	✓	

2.4.3 マルチモーダル環境音合成

画像など，音響シーンラベルや音以外のメディア情報を入力として環境音を合成・変換する研究も行われている。Zhou ら [18] は動画に含まれる画像と音の対応関係を SampleRNN で学習することで，画像に対する環境音を合成する手法を提案している。

また，環境音の特徴を説明するための文（以下，音の説明文）を利用した環境音合成手法も提案されている [19, 20, 21]。音の説明文では，「包丁の音の後に水が流れる音が聞こえる」のように，音響シーンラベルよりもより詳細に音響シーンを説明することが可能である。

2.5 本研究の方針

2.4 節で述べた環境音合成の関連研究の多くは音響シーンを対象としているため，音響シーンに含まれる個々の音である音響イベントを柔軟に制御することが困難である。アニメや動画コンテンツにおいて環境音は作品の印象などを決定づけるため重要な情報であり，個々の音に関しても高い再現精度かつ柔軟な制御が必要であると考える。そこで本論文では音響イベントに着目し，音響イベントを表す音波形を高い精度かつ，柔軟に制御可能な合成手法の提案を目標とする。

本研究では，環境音の特徴を表現する方法として，「音響イベントラベル」，「オノマトペ」，「音声」の 3 つに着目し，これらを入力情報とした統計的手法による環境音合成の提案を行う。各入力情報によって制御が期待できる環境音の特徴を表 2.1 に示す。さらに，「オノマトペ」を利用した環境音合成実現のため，環境音に対しオノマトペが対応づいた大規模なデータセットの構築をする。また，合成された環境音をどのように評価するべきかについても本論文にて議論する。

第3章　環境音合成のための データセット構築

3.1　はじめに

環境音の特徴を表現する方法の1つに，音の特徴を自然言語によって表現するオノマトペが挙げられる [22, 23, 24]。オノマトペは音の特徴を表現する手段として有効であるとされており，オノマトペを検索クエリとして効果音を検索する技術 [25] など幅広い用途で使用されている。音のオノマトペ表現の特徴として，「ピュ」や「ピイイイ」のように文字列の長さや繰り返しによって音の時間的変化を表現可能であることが挙げられる。そのため，図 3.1 に示すように，環境音合成にオノマトペを利用することで，音の繰り返し回数やなどといった合成音の時間的変化の制御が期待できる。オノマトペを用いた環境音合成を実現するためには，オノマトペと環境音間の関係性を獲得する必要があり，そのためには環境音に対しオノマトペが付与されたデータセットが必要となる。しかしながら，現在そのようなデータセットは公開されていない。

本章において，RWCP 実環境音声・音響データベース (RWCP-SSD)[26] に含まれる 105 種類の音響イベントの環境音に対しオノマトペを付与して，データセットの構築を行う。なお，作成したデータセットを RWCP-SSD-onomatopoeia と呼ぶ。オノマトペ付与には，より多くの環境音に対しオノマトペを付与するため，クラウドソーシングサービスを用いる。クラウドソーシングサービスは大量のデータを収集するのに非常に効果的であり，近年様々な分野において利用されている [27, 28, 29, 30]。なお，1 つの環境音に対して複数のオノマトペが想起される場合もあるため，付与したオノマトペに対する自信度並びに，他者がそのオノマトペを見た場合の許容度を付与して，付与されたオノマトペの妥当性について分析を行う。

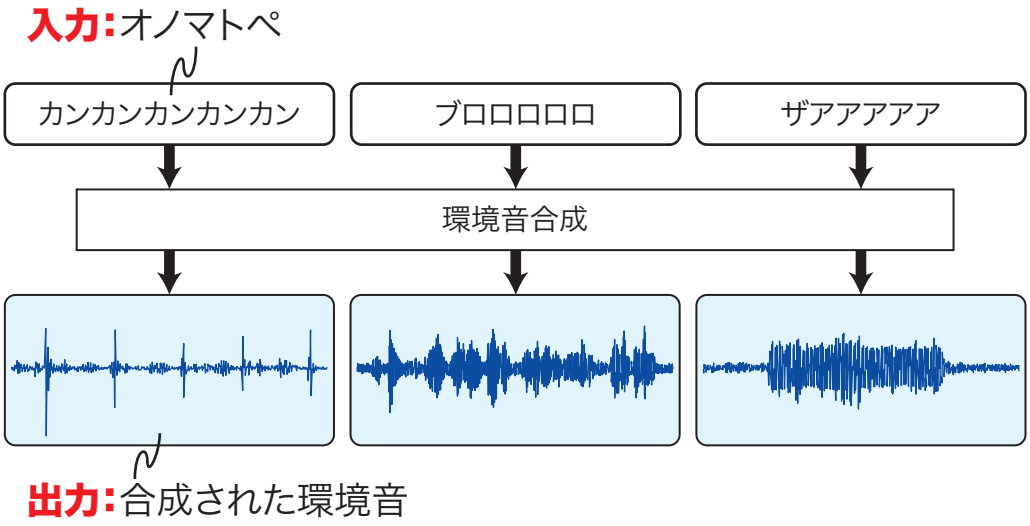


図 3.1: オノマトペを用いた環境音合成の概要

3.2 オノマトペデータセットの構築

3.2.1 RWCP-SSD の概要

本章では，複数の音響イベントを含む RWCP-SSD へのオノマトペ付与を行う。RWCP-SSD には残響が存在しない無響室で様々な音響イベントを収録した 0.5〜2.0 秒程度の長さをもつ環境音が 9,722 サンプル含まれている。音響イベントの種類としては，木板を木棒で叩く音などの衝突系音源（物体の単発的な衝突に起因），拍手の音などの動作系音源（特徴的な音色を持つ），電気カミソリの音などの特徴的音源（音色が音源自身を表す）の 3 タイプの音源が含まれている。これらの音は 48kHz，16bit でサンプリングされている。

3.2.2 構築したデータセットの概要

作成したデータセットには以下の内容を含む。

- 各環境音に対するオノマトペ
  RWCP-SSD 内の 9,722 音（105 種類の音響イベント）に対して 1 音 5 人以上，1 人につき 3 個のオノマトペを付与し，合計 155,568 個のオノマトペを収集した。それぞれのオノマトペは，日本語話者によってカタカナ表記で収集を行った。本データセットには，カタカナ表記のオノマトペを Julius Speech segmentation キット [31] の変換ルールに従い音素表記に変換したオノマトペも含む。

- **オノマトペに対する自信度**
  オノマトペを付与したワーカー自身によるオノマトペに対する自信度の付与を行った。オノマトペに対する自信度を付与することによって，付与されたオノマトペの適切性を評価することが可能となる。
- **他者によって付与されたオノマトペに対する許容度**
  オノマトペを付与したワーカー以外から，音に対して付与されたオノマトペに対する許容度の収集を行った。
- **作業者 ID**
  オノマトペ，自信度，許容度を付与したワーカーの匿名化された作業者 ID

3.2.3 相互相関による音響イベントクラス内のクラスタリング

RWCP-SSD の各音響イベントクラスには約 100 音の環境音が含まれており，非常に類似した音も複数存在する。図 3.2 に RWCP-SSD に含まれる「笛の音」のスペクトログラムを示す。図中のクラス 1，クラス 2 はそれぞれ類似した音を分類したクラスを表す。このように音響的特徴が類似した音が多く含まれ，類似した音に対しては同じオノマトペが付与されると考えられる。そこで，各音響イベントクラス内の環境音に対し波形同士の相互相関を求め，類似していると判断された音に関しては同じ音と判断しクラスタリングを行った。音響イベントクラス内の環境音波形 $\boldsymbol{x} = \{x_1, x_2, ..., x_N\}$，$\boldsymbol{y} = \{y_1, y_2, ..., y_N\}$ の相互相関を求める式を以下に示す。

$$R_{xy} = \max_{0 \leq m < N} \left\{ \frac{\sum_{i=1}^{N} x_{i+m} y_i}{\sqrt{\{\sum_{i=1}^{N} x_{i+m} x_i\}\{\sum_{i=1}^{N} y_{i+m} y_i\}}} \right\} \tag{3.1}$$

N は各環境音波形の総数，下付き文字 m は環境音波形のシフト長，R_{xy} は x と y の相互相関を表す。なお，比較対象の環境音波形の系列長が異なる場合は，短い方の環境音波形の末尾に 0 を付加して，もう一方の環境音波形と系列長を揃えて計算を行った。この式に従い各音響イベントクラス内で環境音同士の相互相関を用いてクラスタリングを行い，結果として 6,024 クラスに分類された。本章ではこの 6,024 の各クラスで，それぞれ 1 音をランダムに選び，6,024 音の環境音についてオノマトペの付与を行なった。

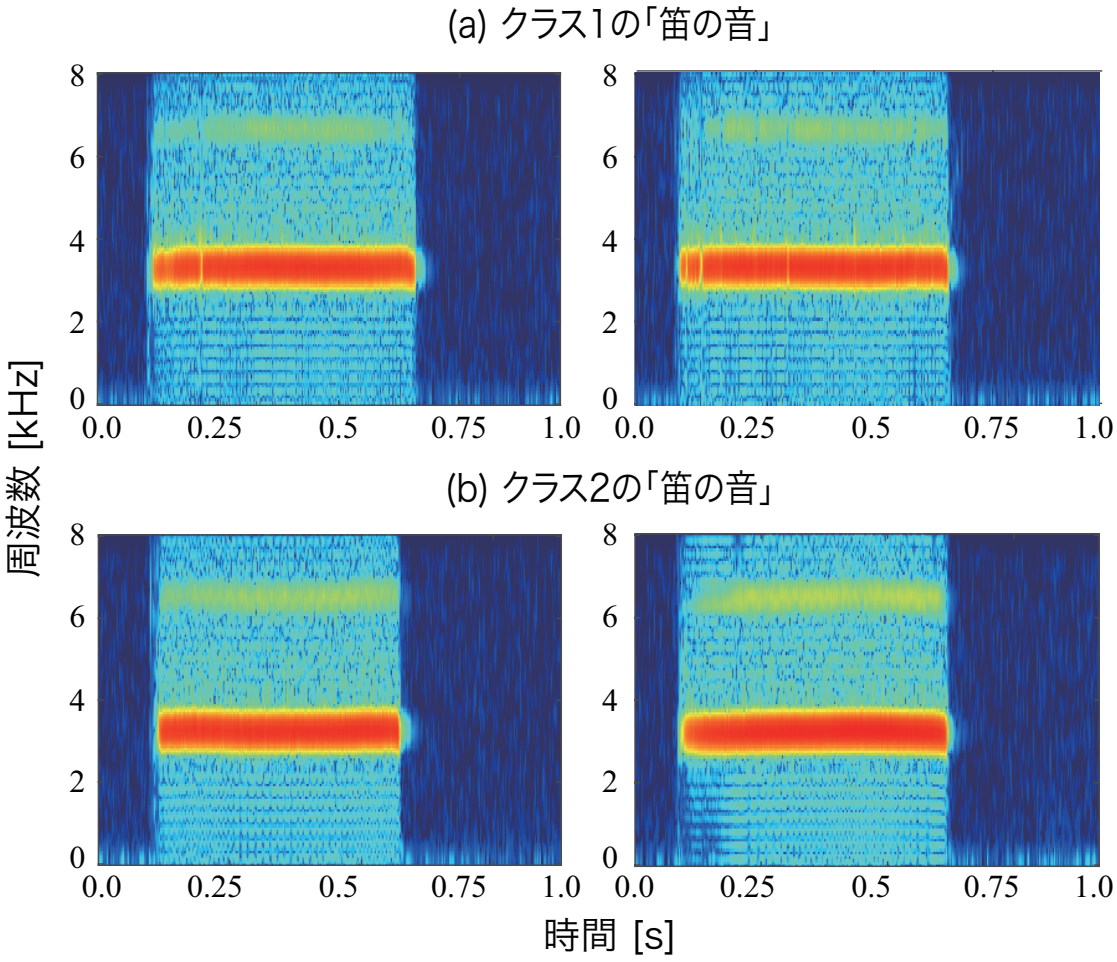


図 3.2: RWCP-SSD に含まれる笛の音のスペクトログラム

3.2.4 クラウドソーシングによるオノマトペおよび自信度，許容度の付与

環境音に対しオノマトペの付与を行う場合，その環境音に対して複数のオノマトペが想起される場合がある。予備実験として，「笛の音」という環境音に対しオノマトペの付与を実施した。その際，同一の音に対して「ピー」「ピイイイ」「ツィーッ」のような様々なオノマトペが付与された。このように一音の環境音に対して，付与されるオノマトペは大きくばらつくという傾向がある。また，これらのオノマトペは環境音を表すものとして適さないものが混在する可能性も懸念される。これらの課題に対処するため，本データセットでは，付与した本人によるオノマトペに対する適切度合い，他者が付与したオノマトペに対する許容度合いを付与した。

クラウドソーシングサービスを利用して以下の 2 つのタスクを実施した。

- タスク 1：環境音に対するオノマトペと自信度の付与

表 3.1: 付与されたオノマトペの例

音響イベント名	音 ID	付与された オノマトペ	自信度	許容度	音の説明
whistle1	000	ピッ ピィ ピィッ	5 4 4	4.9 5.0 4.6	笛の音
	064	フィー ピー ヒーッ	1 5 2	4.1 4.5 4.2	
shaver	071	ジー ビィィイイイイッ ブーン	1 4 4	4.5 3.9 4.5	電気カミソリの動作音
	080	ブィーン ビュイーン ウィーン	4 5 4	3.4 3.0 4.2	
file	054	ヒュンッ ミーッ サッー	4 5 3	3.5 1.9 3.3	金ヤスリで金属棒を擦る音
	095	シャッ シュワッ シュィッ	4 4 4	3.8 3.6 4.8	

音のみを作業者に提示して，その音に対して想起されるオノマトペを 3 個とそれぞれのオノマトペに対する自信度を 1（非常に自信がない）〜5（非常に自信がある）の 5 段階で付与した。なお，1 音に対して 5 人の作業者からオノマトペを収集した。

- **タスク 2：付与されたオノマトペに対する許容度の付与**
  タスク 1 で付与されたオノマトペと音をペアとして作業者に提示した。作業者は提示された音に対して，オノマトペがどの程度許容できるかを 1（非常に許容できない）〜5（非常に許容できる）の 5 段階で付与した。タスク 2 は，タスク 1 で付与された 1 個のオノマトペを他の 5 人の作業者に提示し，許容度の付与を行った。

本データセットにおいては，タスク 1 にて自信度が 4 以上のオノマトペにのみ許容度を付与した。

3.3 付与されたオノマトペの分析

3.3.1 付与されたオノマトペの結果

付与された一部のオノマトペ，自信度，許容度の平均を表 3.1 に示す。許容度は１個のオノマトペに対し複数付与されているため，表には平均値を示す。表より，

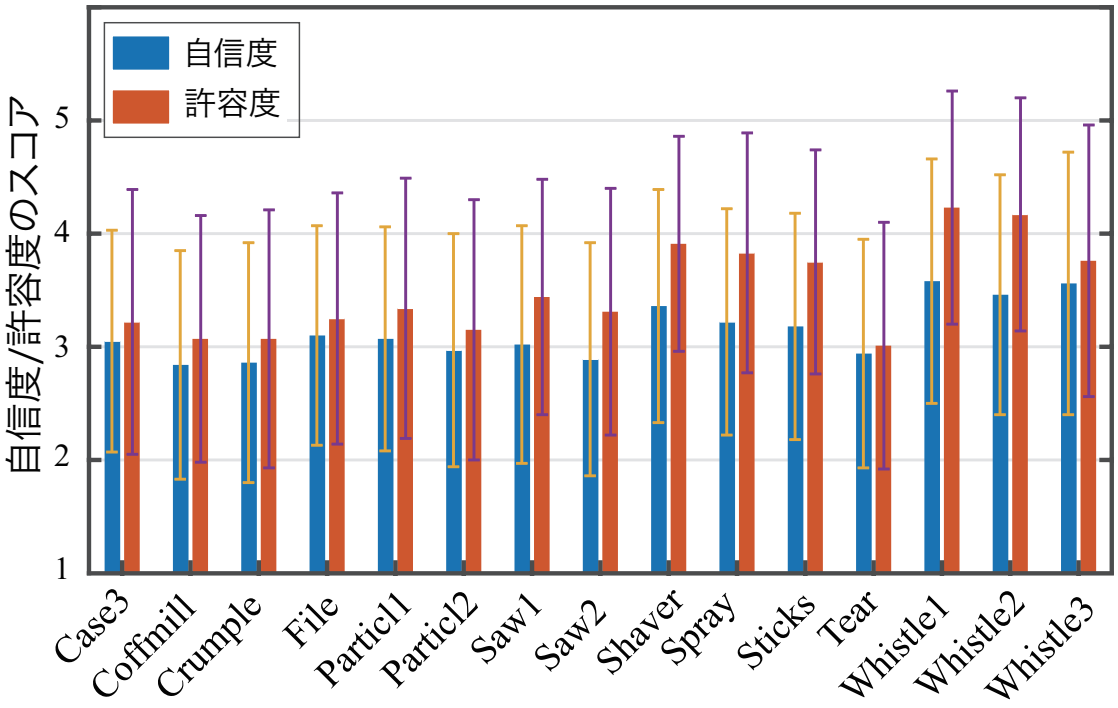


図 3.3: 音響イベントごとの自信度/許容度の平均

それぞれの音響イベントに対し多様なオノマトペが付与されていることがわかる。また，同一の環境音に対して収集されたオノマトペは，音の長さに関する表現は長音記号，同じ文字を繰り返すなどによってある程度類似したオノマトペが付与される傾向も確認された。

3.3.2 付与された自信度と許容度の分析

各音響イベントクラスの環境音に付与されたオノマトペに対する自信度，許容度のそれぞれの平均，分散を図 3.3 に示す。図より，付与された自信度と許容度を比較すると，オノマトペを付与した本人によるオノマトペに対する自信度より，他者がオノマトペに対して付与した許容度の方が高い値をとる傾向がある。

図 3.4 に付与された自信度と許容度の平均の散布図を示す。この結果からも，付与された自信度が高い場合には許容度も高い値が付与される傾向が確認できる。このことより，付与された自信度が高いオノマトペに関しては他者から見た場合も比較的許容されやすいオノマトペであることがわかる。しかし，表 3.1 で示す 電気カミソリの動作音に付与された「ジー」というオノマトペのように，自信度が低く付与されたにも関わらず許容度が高く付与されるオノマトペも存在する。一方，金ヤスリで金属棒を擦る音に付与された「ミーッ」というオノマトペのよう

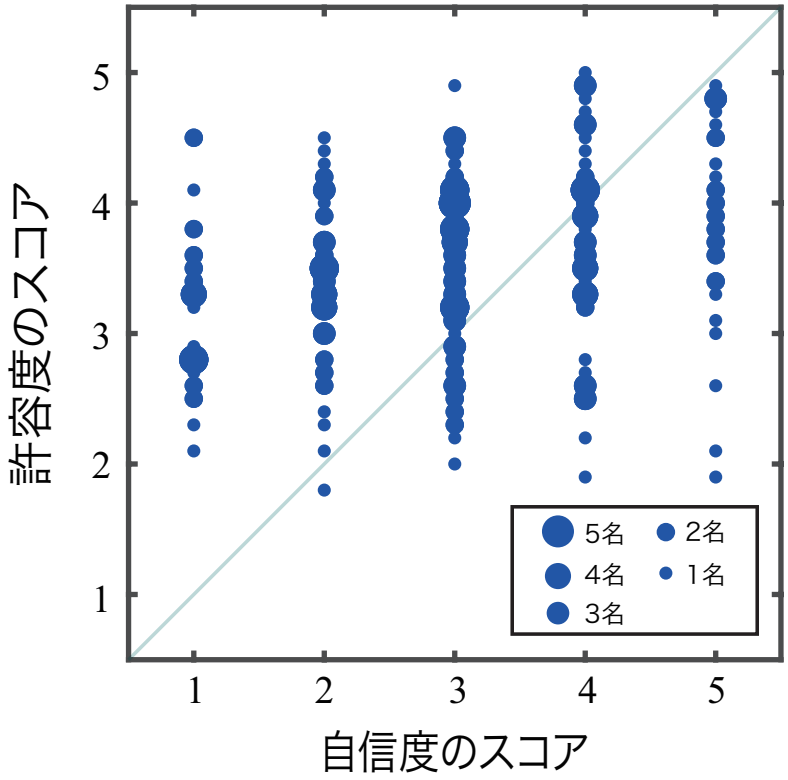


図 3.4: オノマトペに対する自信度/許容度の平均の散布図

に，自信度は高いが許容度が低く付与されたオノマトペも存在した。このようにオノマトペに対する自信度の付与のみでは，環境音を表すオノマトペとして適したオノマトペを選び出すことは困難であると考えられる。そのため，環境音に対して適切なオノマトペを選択する際，自信度と共に許容度を用いることは有効であると考えられる。

図 3.5 および図 3.6 に各音響イベントごとに付与された自信度，許容度の割合の一例を示す。図 3.5 より，笛の音以外の音響イベントでは自信度 3 が最も多く付与される結果となった。一方，笛の音という音響イベントに関しては，自信度 4 が最も多く付与された。その理由として，笛の音は了解性が高く，オノマトペが付与しやすかったのではないかと考えられる。また図 3.6 より，いずれの音響イベントに対しても許容度 3 以上が付与された割合が高く，自信度と同じく了解性の高い笛の音に関しては許容度 5 が最も高くなる傾向が確認できる。

これらのことより，環境音に対するオノマトペ付与では，自信度と許容度を付与することにより，環境音を表すオノマトペとして妥当なオノマトペを選び出す手助けになることが確認できる。

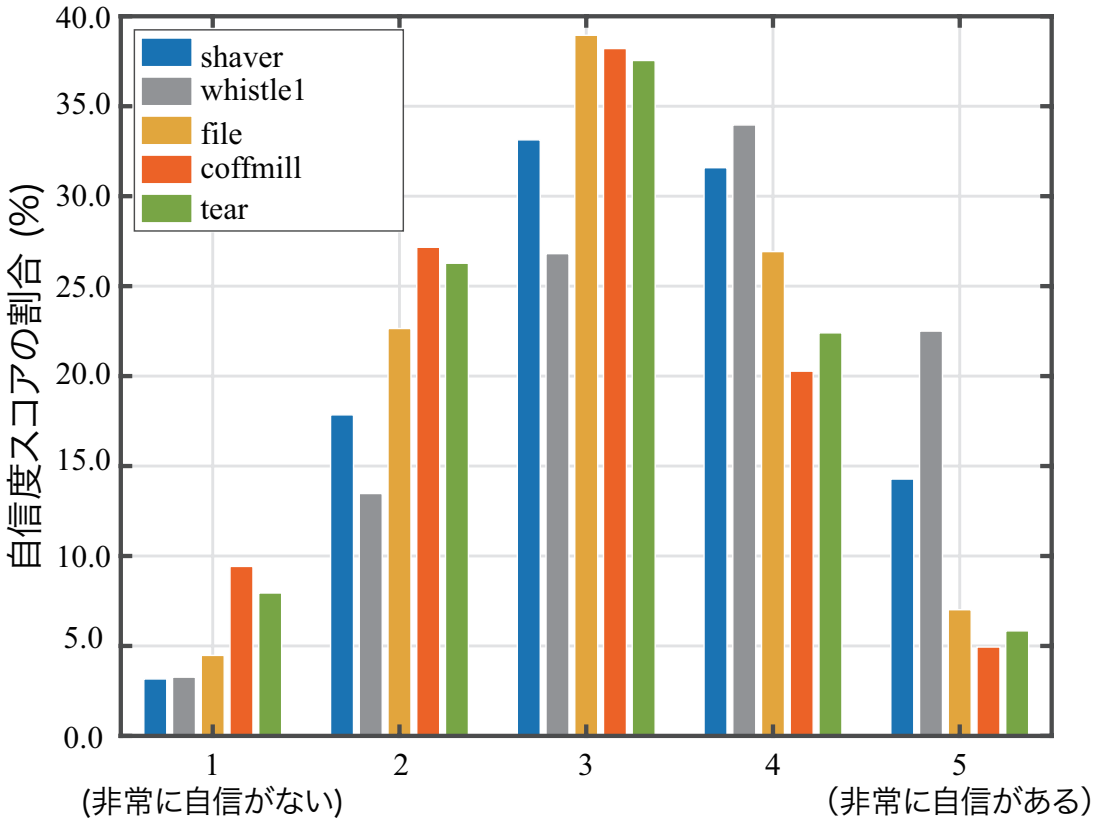


図 3.5: 音響イベントごとの自信度の平均

3.4 3章のまとめ

本章では，オノマトペを用いた環境音合成のためのデータセットとして，RWCP-SSD に含まれる音響イベントに対してオノマトペを付与して，その付与されたオノマトペを評価するための指標とし自信度と許容度の付与を行いデータセットを構築した。結果より，環境音に対して適切なオノマトペ付与を行うためには，自信度と許容度を付与することにより，環境音を表すオノマトペとして適切なオノマトペを選び出せることが確認された。

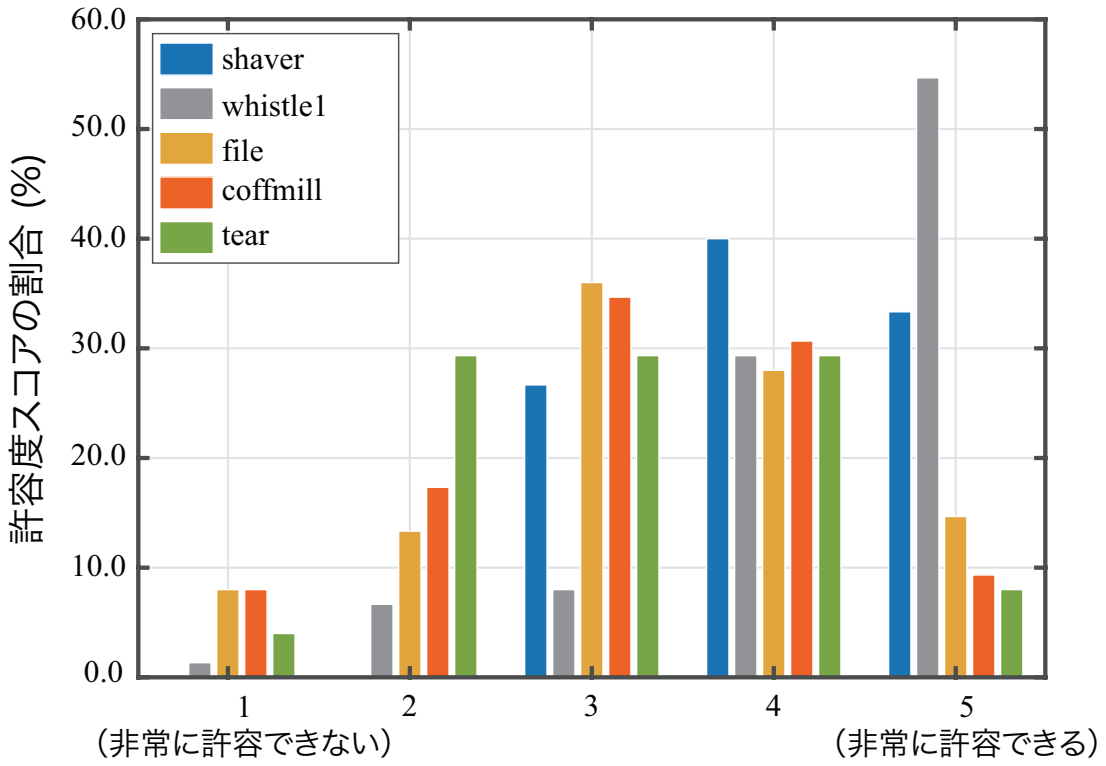


図 3.6: 音響イベントごとの許容度の平均

第4章　環境音合成のための評価手法

4.1　はじめに

合成された環境音を評価するため，主観・客観評価の方法がいくつか提案されている [5, 32, 33]。いくつかの評価方法が提案されている一方，どのように合成された環境音を評価するべきであるかについて確立された方法論は存在しない。そのため，客観評価でよいのか，主観評価でよいのか，またどのような観点で合成された環境音の評価を行えばよいのかが明らかになっていない。従来，評価にかかるコストの観点から，客観評価が多く行われる傾向がある。しかし，合成音の主観的な品質と完全に対応する客観評価手法は存在しない。現に，統計的手法によって合成された音声の品質を競う Blizzard Challenge [34] では，人間の評価者による主観評価によって順位を決定する。そのため，合成された環境音を評価する際にも，主観的な評価が必要であると考える。

本章においては，合成された環境音を主観的に評価するための方法論を提案する。具体的には，合成に用いられた入力情報に対して合成された環境音がどの程度妥当であるかを評価する手法を提案する。なお本章では，環境音合成モデルに音響イベントラベル，オノマトペ，その両方が入力された場合の合成音を対象として実験的評価を行う。さらに，従来使用されている客観評価手法と本章にて提案する主観評価手法の結果を比較することで，環境音合成における主観評価手法の必要性を示す。

4.2　環境音合成における従来の評価手法

4.2.1　従来の客観評価手法

Liu らは，音響イベントラベルを入力とした環境音合成の評価として，音響イベント分類器を用いて合成音を客観的に評価する手法を提案している [32]。図 4.1 に評価の流れを示す。まず，環境音合成モデルに対して合成したい音響イベントの音響イベントラベルを入力する。そして合成された音を音響イベント分類器に入力して，合成時に入力された音響イベントラベルと同じクラスに分類されるかど

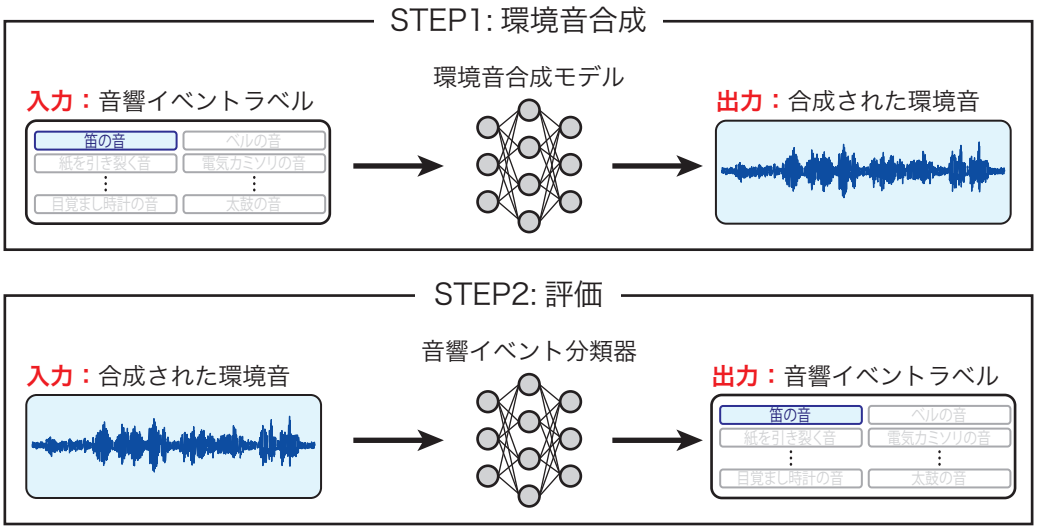


図 4.1: 音響イベント分類器を用いた環境音の客観評価の概要

うかで，合成音の品質を評価する。なお，環境音合成のモデル学習に使用した環境音データと同一のものを使用して音響イベント分類器を学習させる。このような客観評価手法は，環境音分析などといったコンピュータが音を識別するタスクにおいて，合成音が学習データとして使用できるかどうかの一つの判別方法としては有効であると考えられる。しかしながら，正しい音響イベントクラスに分類されたからといって，人間がその合成音を聴いた際に高品質と感じるとは限らない。例えば，メディアコンテンツに使用される環境音は，作品の印象などを特徴づける一つの要因であるため，高品質な音でなければならない。そのため，音響イベント分類器を使用した客観評価のみでは，環境音合成の評価手法として不十分であると考えられる。

4.2.2 従来の主観評価手法

合成された環境音の品質を主観的に評価する方法もこれまでいくつか取り組まれてきた [35, 36]。これらの評価では，被験者に音のみを提示して，5 段階のスコアを付ける絶対評価が採用されている。しかしながら，従来研究の評価結果より，「コーヒーミルで豆を挽く音」のような環境音は，高品質な音の場合でも評価スコアが低くなる傾向があることが確認されている。また，絶対評価によって音の評価を行うことは被験者間での一貫した評価を得ることが非常に難しいことも懸念される。さらに，従来実施されてきた音のみを被験者に提示して評価する方法では，合成音のみが提示されるため，どのような入力情報（音響イベントラベルやオノマトペなど）によって合成された音であるか被験者は把握できない。よって，

合成された音が入力となった情報を反映した音であるかを評価できていない。そのため，品質だけでなく，合成音が環境音合成モデルへの入力情報を反映した音であるかも評価する必要がある。

合成音が環境音合成モデルへの入力情報を反映した音であるかどうかの主観評価も取り組みが行われている [36]。この評価では，環境音と合成モデルへの入力となった情報の両方を被験者に提示して，5 段階の絶対評価を行う。この評価は，前述した課題の一つを解決するものであり，入力情報に対する合成音の妥当性を評価可能である。しかしながら，この方法は，絶対評価のみに限定されており，前述した絶対評価の不十分さを依然として解決できていない。

4.3 環境音合成のための主観評価手法の提案

本章では，合成された環境音が合成モデルの入力情報を反映した音であるかを主観的に評価する方法論を提案する。4.2.2 項にて述べた絶対評価の不十分さについて議論するため，本提案手法では絶対評価と相対評価の両方を行う。これらの評価結果を比較することで，どのように環境音を主観的に評価するべきかについて議論する。本章では，(1) 音響イベントラベルのみ，(2) オノマトペのみ，(3) 音響イベントラベルとオノマトペの両方の 3 種類の入力情報は扱う。本章では，入力情報に対する環境音の「妥当性」を評価指標として利用する。次項以降では，各入力情報の側面から環境音を評価する手法を提案する。

4.3.1 音響イベントラベルから合成された環境音の主観評価手法

音響イベントラベルから合成された環境音の評価手法として，環境音と音響イベントラベルを同時に提示する絶対評価と相対評価の 2 種類の主観評価手法を提案する。

- **評価手法 I-1：音響イベントラベルに対する環境音の妥当性の絶対評価**
  1 音の環境音と音響イベントラベルを被験者に同時に提示する。被験者は，提示された音響イベントラベルを表す音として，提示された環境音がどの程度妥当であるかを 1（非常に妥当でない）〜5（非常に妥当である）の 5 段階で回答。

- **評価手法 I-2：音響イベントラベルに対する環境音の妥当性の相対評価**
  2 音の環境音と音響イベントラベルを被験者に提示する。被験者は，提示された音響イベントラベルを表す音として，提示された 2 音の環境音のうちど

ちらの音のほうが妥当であるかを回答。

4.3.2　オノマトペから合成された環境音の主観評価手法

オノマトペから合成された環境音の評価手法として，環境音とオノマトペを同時に提示する絶対評価と相対評価の 2 種類の主観評価手法を提案する。

- **評価手法 II-1：オノマトペに対する環境音の妥当性の絶対評価**
  1 音の環境音とオノマトペを被験者に同時に提示する。被験者は，提示されたオノマトペを表す音として，提示された環境音がどの程度妥当であるかを 1（非常に妥当でない）〜5（非常に妥当である）の 5 段階で回答。

- **評価手法 II-2：オノマトペに対する環境音の妥当性の相対評価**
  2 音の環境音とオノマトペを被験者に提示する。被験者は，提示されたオノマトペを表す音として，提示された 2 音の環境音のうちどちらの音のほうが妥当であるかを回答。

4.4　評価実験

提案する主観評価手法と従来取り組まれている音響イベント分類器による合成音の評価実験を実施する。また，主観評価と客観評価の結果を比較する。本実験にて用いた環境音合成手法を表 4.1 に示す。WaveNet を用いた合成手法は 5 章，sequence-to-sequence (seq2seq) [37] を用いた手法は 6 章，Transformer [10] を用いた手法は付録 A にて詳細を述べる。なお，データセットに含まれる自然音（natural sound) に対しても評価を実施した。主観評価にはクラウドソーシングサービスを利用した。評価手法 I-1 では，各評価に 100 名の被験者が参加し，各被験者が 25 音の合成音を評価した。評価手法 I-2 では，各評価に 150 名の被験者が参加し，各被験者が 2 音 1 組の計 25 組の合成音を評価した。評価手法 II-1 では，各評価に 100 名の被験者が参加し，各被験者が 30 音の合成音を評価した。評価手法 II-2 では，各評価に 300 名の被験者が参加し，各被験者が 2 音 1 組の計 25 組の合成音を評価した。各主観評価にて被験者に提示した情報は，表 4.2 に示す。

表 4.1: 各主観/客観評価に使用した環境音合成手法の一覧

合成手法	モデルへの入力情報		主観評価手法				客観評価手法
	オノマトペ	音響イベントラベル	評価手法 I-1	評価手法 I-2	評価手法 II-1	評価手法 II-2	
WaveNet		✓	✓	✓			✓
Seq2seq	✓				✓	✓	
Seq2seq + event label	✓	✓	✓	✓	✓	✓	✓
Transformer	✓				✓	✓	
Transformer + event label	✓	✓	✓	✓	✓	✓	✓

表 4.2: 各主観評価実験において被験者に提示した情報

評価手法	提示する音の数	音響イベントラベル	オノマトペ
Okamoto et al. [35]	1		
Liu et al. [38]	1		
Yang et al. [39]	1		
Okamoto et al. [36]	1		✓
評価手法 I-1 （提案手法）	1	✓	
評価手法 I-2 （提案手法）	2	✓	
評価手法 II-1 （提案手法）	1		✓
評価手法 II-2 （提案手法）	2		✓

表 4.3: 評価実験で使用した音響イベントと学習・評価サンプル数

音響イベント名	学習サンプル数	評価サンプル数	音の説明
Coffee grinder	95	5	コーヒー豆をミルで挽く音
Cup	95	5	カップを叩く音
Clock	95	5	目覚まし時計の音
Whistle	95	5	笛の音
Maracas	95	5	マラカスの音
Drum	95	5	ドラムを叩く音
Shaver	95	5	ひげ剃りの動作音
Trash box	95	5	金属製のゴミ箱を叩く音
Tearing paper	95	5	紙を引き裂く音
Bell	95	5	ベルを鳴らす音

4.4.1 実験条件

各環境音合成モデルの学習には，RWCP-SSD に収録されている 10 種類の音響イベントを使用する。使用した音響イベントと学習，評価に用いたデータ数を表 4.3 に示す。各環境音に対応するオノマトペデータは，3 章にて構築したデータセットを用いる。各環境音に対して，3 章にて構築したデータセットから 15 個のオノマトペ，計 14,250 個のオノマトペを使用する。各環境音合成モデルのネットワークパラメータは表 4.4 に示す。

表 4.4: 各合成手法のパラメータ設定及び使用した音響特徴量

音の長さ	1–2 s
サンプリング周波数	16,000 Hz
音波形の圧縮形式	16-bit linear PCM
音響特徴量	対数メルスペクトログラム（80 次元）
フレーム長	0.128 s（2,048 サンプル）
フレームシフト	0.032 s（512 サンプル）
Seq2seq / Seq2seq + event label	
Encoder の LSTM 層の数	1
Encoder の LSTM 層のユニット数	512
Decoder の LSTM 層の数	2
Decoder の LSTM 層のユニット数	512, 512
音響イベントラベルの次元数	10
Teacher forcing 率	0.6
バッチサイズ	5
最適化手法	RAdam [40]
Transformer / Transformer + event label	
Encoder 層の数	3
Decoder 層の数	3
Multi-head の数	4
バッチサイズ	32
音響イベントラベルの次元数	10
最適化手法	RAdam

表 4.5: 音響イベント分類器のパラメータ設定および使用した音響特徴量

音響特徴量	対数メルスペクトログラム（64 次元）
フレーム長	0.04 s（640 サンプル）
フレームシフト	0.02 s（320 サンプル）
CNN 層の数	3
CNN のチャネル数	32, 64, 64
CNN のフィルタサイズ	3×5
プーリング	3×3, 3×3（最大プーリング）
全結合層の数	2
全結合層のユニット数	64, 128

4.2.1 項で紹介した客観評価を行うため，音響イベント分類器を構築した。この分類器は、環境音合成モデルの学習に使用した環境音を用いて学習させた。音響イベント分類器のネットワークは，3 層の CNN と 2 層の全結合層 (FC) から構成した．音響イベント分類器のネットワークパラメータは表 4.5 に示す。

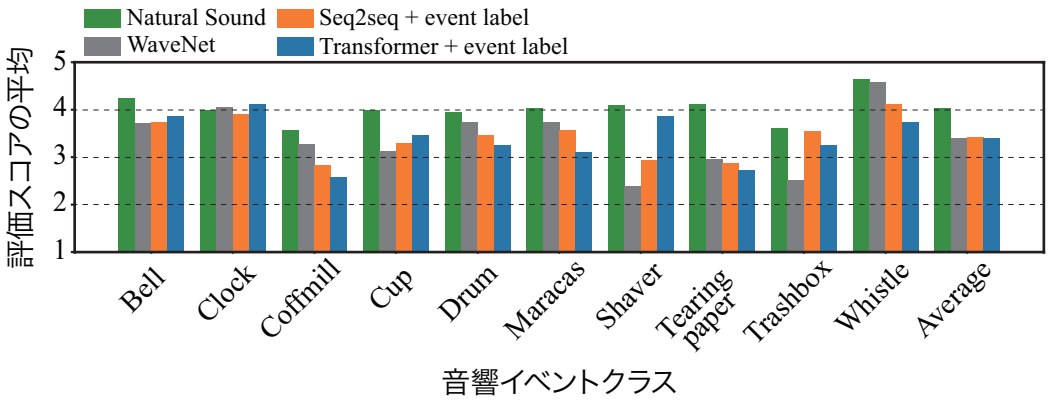


図 4.2: 音響イベントラベルに対する環境音の妥当性の絶対評価結果 (音と音響イベントラベルを提示)

4.4.2 音響イベントラベルから合成された環境音の主観評価結果

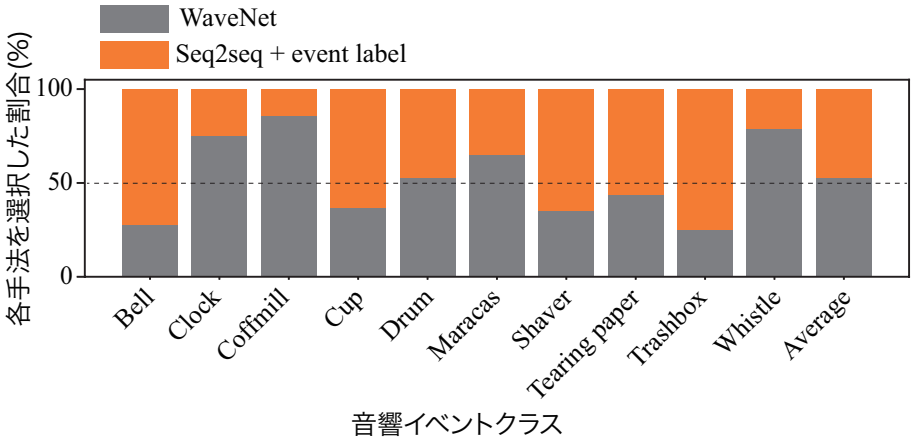
評価方法 I-1 の結果を図 4.2 に示す。図より，全体の平均スコア (Average) において，各合成手法間に顕著な差はないことがわかる。この結果から、各合成手法は、入力された音響イベントラベルを表現する音を同程度の品質で合成できることがわかる。

評価手法 I-2 の結果を図 4.3 に示す。図より，評価手法 I-1 と同様に，各手法の全体の平均スコア (Average) に顕著な差がないことがわかる。“Maracas” の音に着目すると，図 4.2 の絶対評価では，seq2seq + event label による合成音が Transformer + event label による合成音よりも高いスコアを獲得している。しかし，図 4.3 (c) では，seq2seq + event label による合成音と Transformer + event label による合成音は同程度のスコアを獲得している (p 値 = 0.76)。これらの結果より，音響イベントラベルを用いた合成音の評価では，絶対評価だけでなく相対評価も必要であることがわかる。

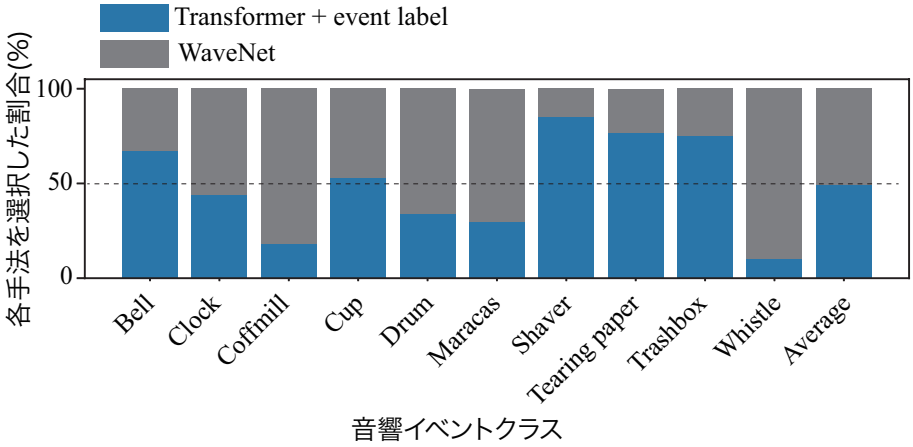
4.4.3 オノマトペから合成された環境音の主観評価結果

評価手法 II-1 の結果を図 4.4 に示す。図より，入力に使用したオノマトペに対する合成音の妥当性は，各手法間において顕著な差がないことがわかる。よって，絶対評価の結果からは，合成時にオノマトペと同時に音響イベントラベルを入力とすることは合成音の品質に影響を与えないことが確認できる。

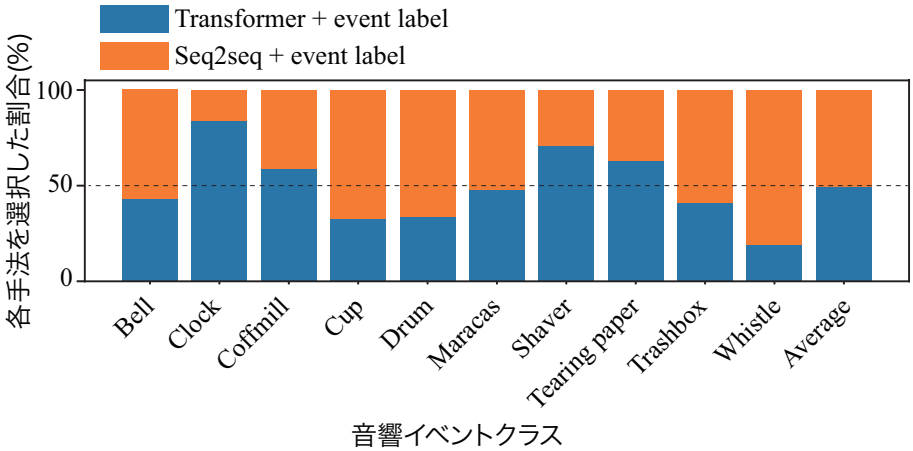
評価手法 II-2 の結果を図 4.5 に示す。図 4.5 (a) および (d) の結果より，オノマトペと音響イベントラベルを入力とする手法による合成音と比較して，オノマトペのみを入力とする手法による合成音のほうが入力となったオノマトペに対して妥



(a) WaveNet vs Seq2seq + event label



(b) Transformer + event label vs WaveNet



(c) Transformer + event label vs Seq2seq + event label

図 4.3: 音響イベントラベルに対する環境音の妥当性の相対評価結果

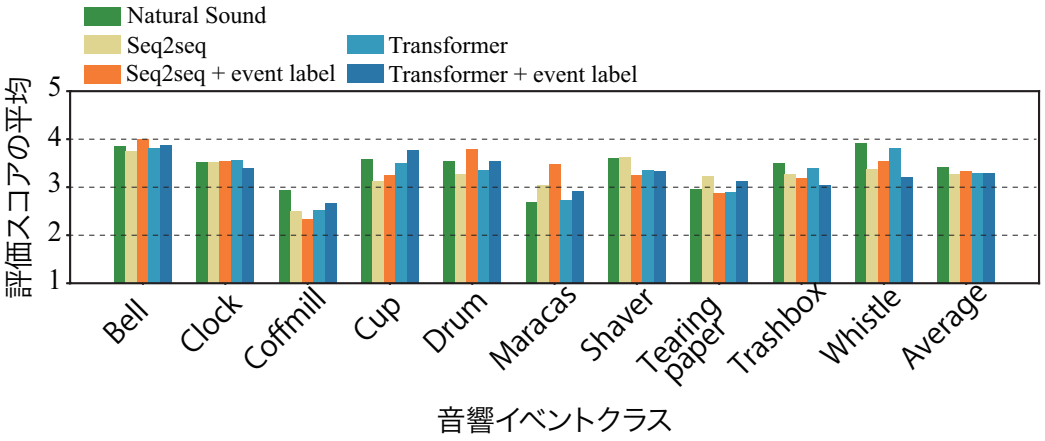


図 4.4: オノマトペに対する環境音の妥当性の絶対評価結果

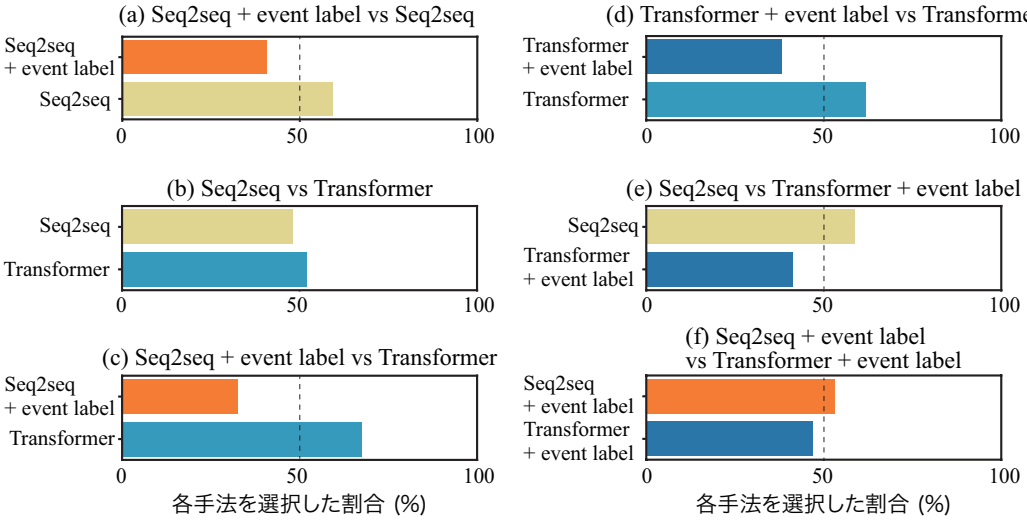


図 4.5: オノマトペに対する環境音の妥当性の相対評価結果

当であるという結果になった。評価手法 II-1 の絶対評価では，これら手法間において顕著な品質差がなかったことから，オノマトペから合成された環境音の評価においても，絶対評価のみならず相対評価も実施する必要があることがわかった。

4.4.4 音響イベント分類器による合成音の客観評価結果

図 4.6 に音響イベント分類器を用いた合成音の客観評価結果を示す。図 4.6 (a) より，WaveNet による合成音のほとんどが分類器によって正しいクラスに分類されていないことがわかる。一方，図 4.6 (b) および (c) の結果より，seq2seq と Transformer によって合成された音は，音響イベント分類器によって概ね正しいク

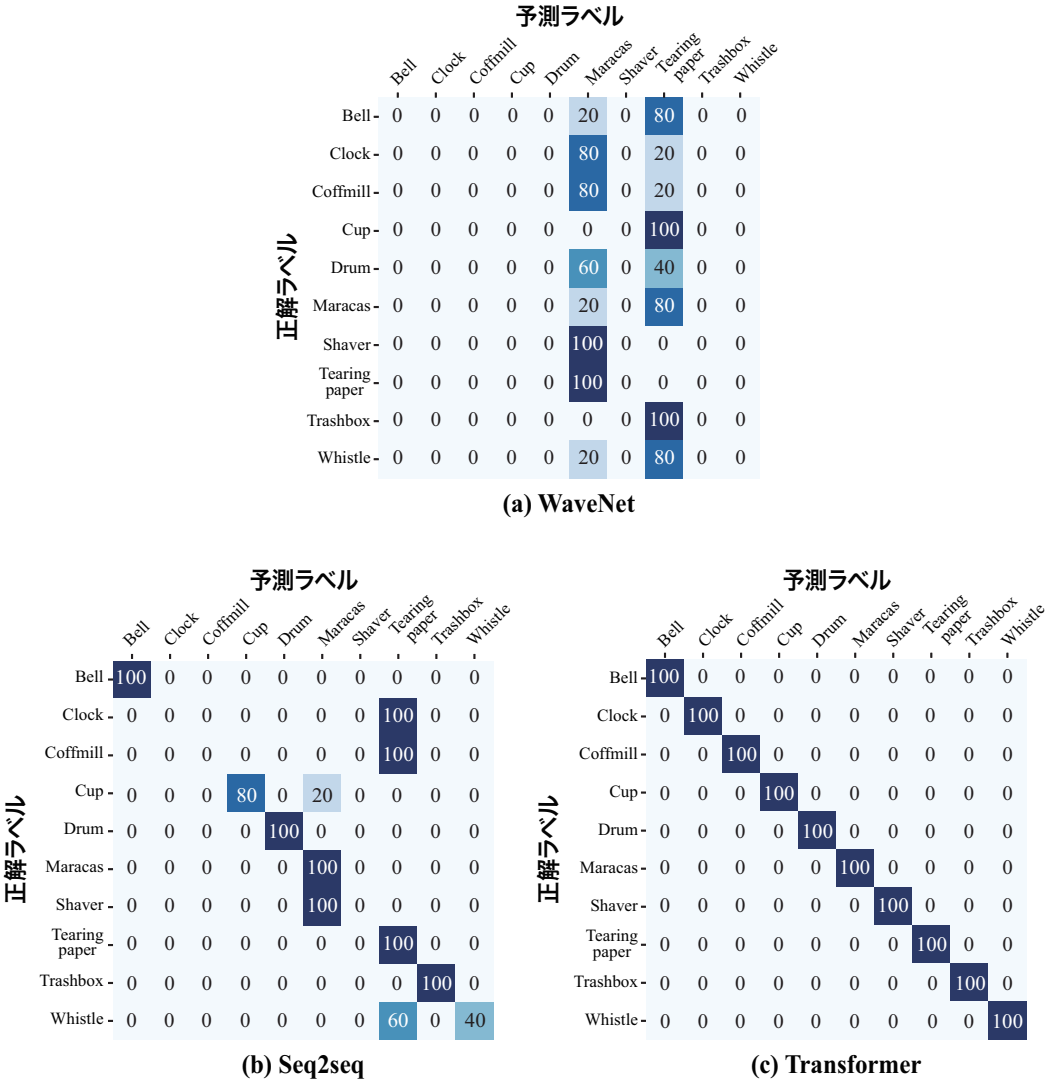


図 4.6: 音響イベント分類器による合成音の分類結果

ラスに分類されていることがわかる。

4.4.5 主観評価と客観評価結果の比較

図 4.7 に主観評価スコアと音響イベント分類器による客観評価の比較結果を示す。図より，音響イベント分類器によって合成音が正しい音響イベントクラスに分類された場合においても，主観評価のスコアが高くならない場合もあることが確認できる。一方，客観評価で正しい音響イベントクラスに分類されない場合でも，主観評価において高いスコアが得られる合成音も存在する。音響イベント分類器を用いた客観評価では，学習データに類似した音をより高い精度で正しいク

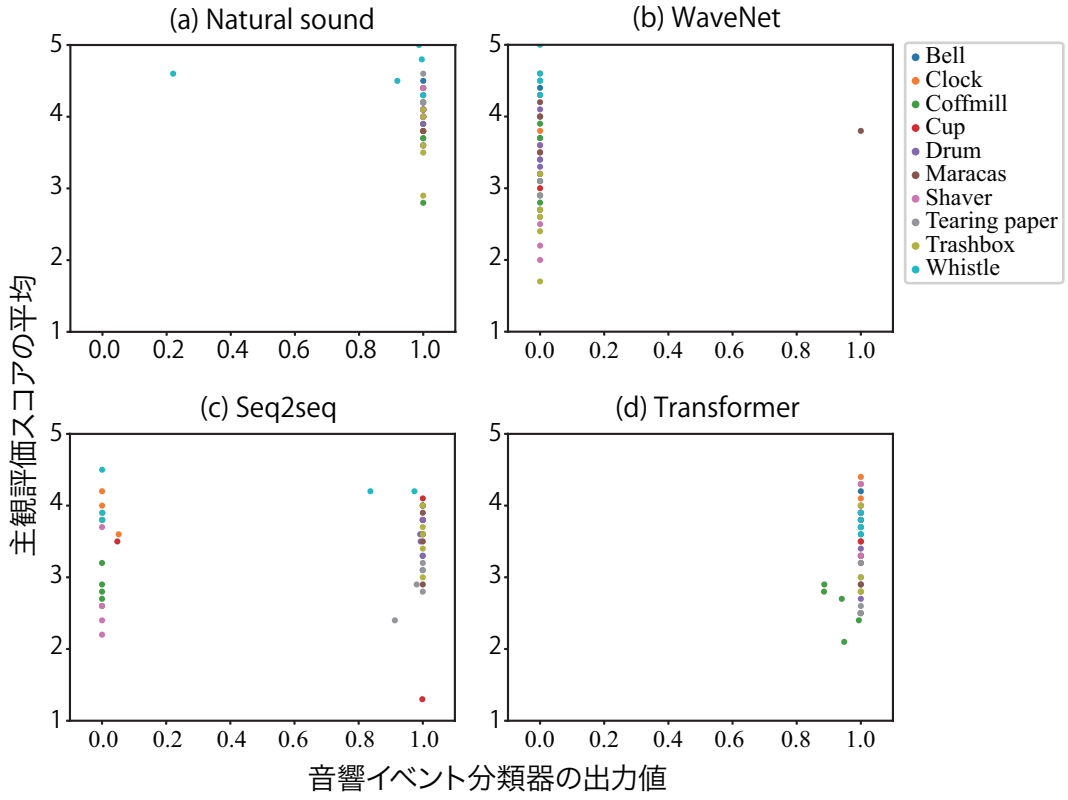


図 4.7: 主観評価と客観評価の比較結果

ラスに分類することができる。そのため，例えば，音響イベントラベルを反映した合成音が合成されている場合でも，学習データに使用した環境音の特徴と異なる場合，正しい音響イベントクラスに分類されない傾向がある。よって，メディアコンテンツなど，人が聴取するために合成する環境音の評価には，客観的な評価だけでなく，主観的な評価も必要であると言える。

4.5　4章のまとめ

本章では，合成された環境音をどのように評価するべきかについて検討した。環境音合成モデルの入力に使用された情報に対する合成音の妥当性を評価する主観評価手法を提案して，客観評価の結果と比較を行った。比較実験の結果，主観評価と客観評価の間では異なる傾向が見られることが確認された。よって，環境音合成の評価においては，客観評価のみならず，主観評価も必要であることが明らかとなった。今後，オノマトペのみを環境音合成モデルの入力に使用した場合の合成音の客観評価手法を検討して，主観評価の結果と比較を行う必要がある。

第5章　音響イベントラベルからの環境音合成

5.1　はじめに

動画やアニメなどの映像コンテンツへ環境音を付与する際，大量のデータベースから音響イベントを検索クエリとして使用する場合がある。また，環境音が多く含まれるデータセットなどには音響イベントラベルが付与されていることが多い [41, 42, 43]。そのため，目的の環境音を得るために音響イベントラベルを使用することは効果的であると考えられる。

本章では，統計的手法による音響イベントラベルからの環境音合成手法を提案する。環境音は音声と異なり，音素やアクセントなどの言語情報を利用してモデル学習を行うことが不可能である。そのため，大量の環境音データよりf音の特徴パターンをモデル学習することによって，環境音と音響イベントラベル間の対応関係の獲得が期待できる。本章では，音声合成において高い自然性を獲得しているWaveNetを用いることで，環境音と音響イベントラベル間の対応関係を獲得することを目指す。そして，音響イベントラベルを入力とすることで，合成音の音源の種類を制御する。評価実験において，入力となった音響イベントラベルの特徴を反映した環境音が合成可能であることを示す。

5.2　統計的手法による音響イベントラベルからの環境音合成

5.2.1　提案手法の概要

図 5.1 に音響イベントラベルからの環境音合成手法の概要を示す。提案手法は学習部と合成部から構成される。学習部において環境音 $\boldsymbol{x} = \{x_1, x_2, ..., x_T\}$ とそれに対応づいた音響イベントラベル $\boldsymbol{c}$ を用いて環境音合成モデル $\boldsymbol{\lambda}$ を学習する。T は波形サンプルの総数を表す。なお，本章においてはone-hot表現された音響イベ

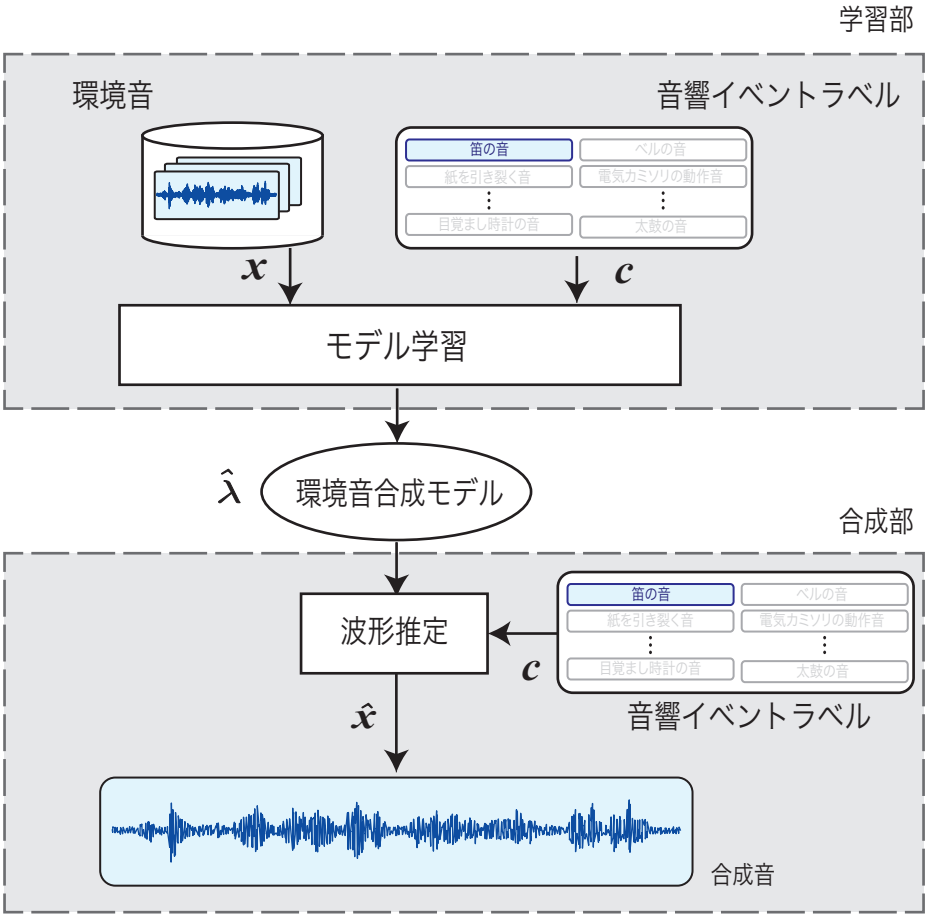


図 5.1: 統計的手法による音響イベントラベルを用いた環境音合成の概要

ントラベルを用いた。

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}} P(\boldsymbol{x} \mid \boldsymbol{c}, \boldsymbol{\lambda}) \tag{5.1}$$

環境音合成モデルの構築方法に関しては，5.2.2 項にて詳細に述べる。合成部では，合成したい音響イベントの音響イベントラベルと学習部にて構築した環境音合成モデル $\hat{\boldsymbol{\lambda}}$ を使用して環境音の波形 $\hat{\boldsymbol{x}}$ を推定する。

$$\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} P(\boldsymbol{x} \mid \boldsymbol{c}, \hat{\boldsymbol{\lambda}}) \tag{5.2}$$

5.2.2 音響イベントラベルを用いた WaveNet によるモデル構築

図 5.2 に音響イベントラベルを用いた WaveNet による環境音合成モデル $\hat{\boldsymbol{\lambda}}$ の構築手法の概要を示す。WaveNet は主に音声合成で用いられる深層学習モデルで，

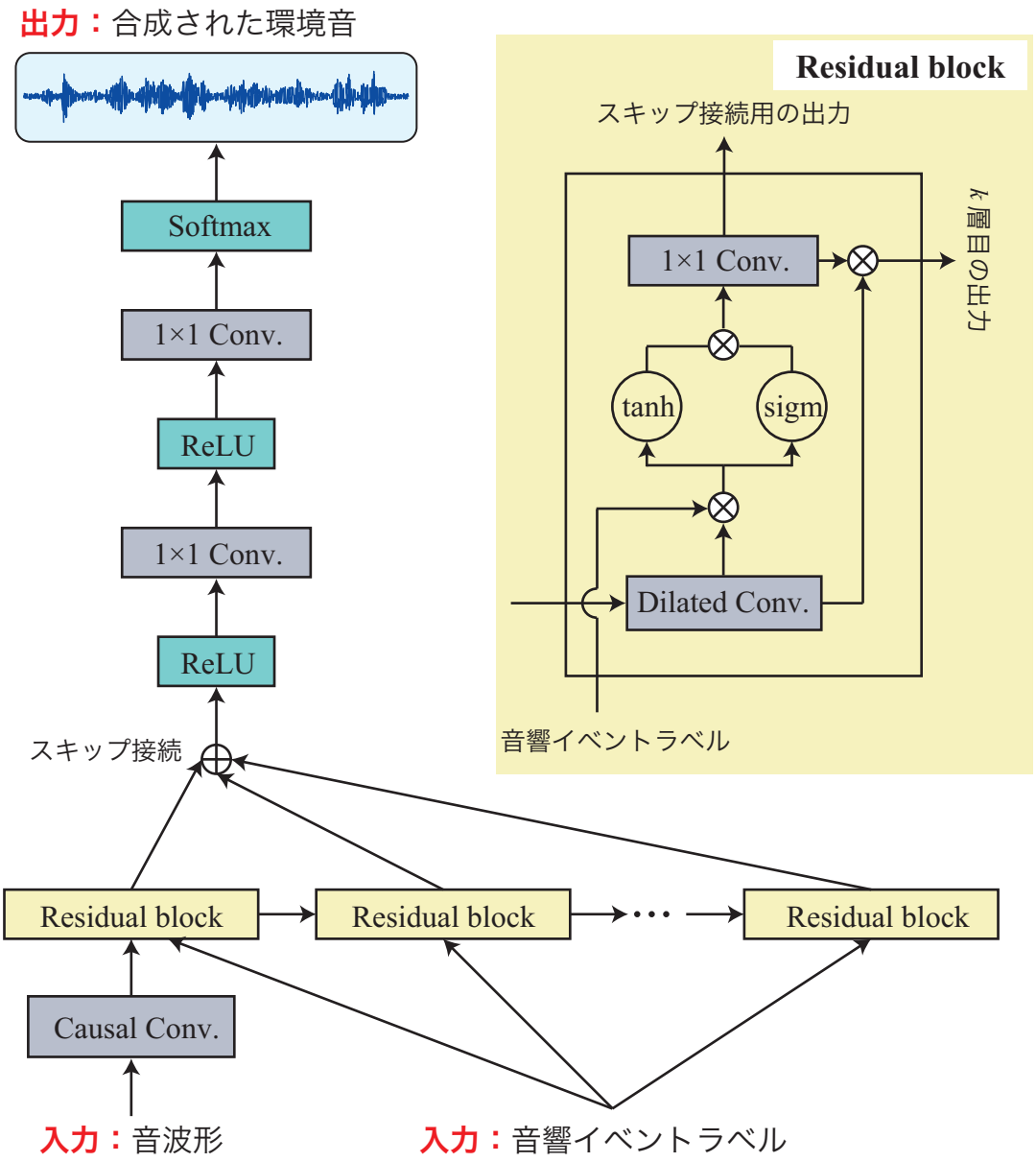


図 5.2: WaveNet を用いた環境音合成モデルの構築

causal convolution 層と dilated causal convolution 層と呼ばれる 1 次元畳み込み層の積み重ねにより構成される。入力された環境音はこれらの畳み込み層とゲート付き活性化関数を通り，ソフトマックス関数により波形サンプルの事後確率が出力される。事後確率の対象は μ-law アルゴリズム [44] によって量子化された波形値である。図 5.2 に示す residual block 内のゲート付き活性化関数の設計は次式のように行われる。

$$\boldsymbol{z} = \tanh(\boldsymbol{W}_{f,k} * \boldsymbol{x}) \odot \sigma(\boldsymbol{W}_{g,k} * \boldsymbol{x}) \tag{5.3}$$

*は畳み込み演算，$\odot$ はアダマール積を表し，$\tanh(\cdot)$ は双曲線正接関数，$\sigma(\cdot)$ はシグモイド関数，$\boldsymbol{W}$ は畳み込み重みを表す。下付き文字 f，g，k はそれぞれ，フィルタ，ゲートを表す添字，活性化関数に関する通し番号である。

波形サンプル $\boldsymbol{x} = \{x_1, x_2, ..., x_T\}$ を確率変数とみなしたとき，causal convolution 層による過去の波形サンプルを考慮した線形演算と，活性化関数による非線形演算の繰り返しにより次の時刻のサンプルを推定する。なお，本章においては音響イベントラベル $\boldsymbol{c}$ による条件付けを行うため，以下の条件付き確率により次の時刻のサンプル予測を行う。

$$p(\boldsymbol{x} \mid \boldsymbol{c}) = \prod_{t=1}^{T} p(x_t \mid x_1, ..., x_{t-1}, \boldsymbol{c}) \tag{5.4}$$

実際，ネットワークのサイズは有限であるため，無限個の波形サンプルを考慮することは不可能である。そこで，WaveNet は式 (5.5) の同時確率を近似的に計算する。

$$p(\boldsymbol{x} \mid \boldsymbol{c}) \approx \prod_{t=1}^{T} p(x_t \mid x_{t-R}, ..., x_{t-1}, \boldsymbol{c}) \tag{5.5}$$

R は過去に遡ることができる時間幅であり，受容野 (receptive field)[1] と呼ばれる。

環境音の波形を合成する際は，WaveNet からの波形サンプリングを所望の回数繰り返すことにより行われる。その際のネットワークへの入力は，過去に自身が合成した波形サンプルである。

5.3 評価実験

映画やゲーム，VR などのメディアコンテンツ作品における背景音や効果音として環境音合成を用いる場合，再生された音が何の音か分かることや，自然音のように感じること，環境音として自然に感じることなどが重要と考えられる。そこで本章では，(I)：複数の環境音をどの程度区別可能な音として合成し得るか，どのような誤りが発生するのか，(II)：自然音と区別できない音がどの程度合成できるか，(III)：どの程度自然性の高い音が合成可能か，の 3 つの観点で主観評価実験を実施した。

5.3.1 実験条件

4 章で述べたように，合成された環境音の評価には主観評価が必要である。そのため，提案手法により合成した環境音に対して主観評価実験を実施した。本実験

表 5.1: WaveNet のパラメータ設定

音の長さ	1–2 s
サンプリング周波数	16,000
音波形の圧縮形式	16-bit linear PCM（自然音） 8-bit μ-law（合成音）
フィルタサイズ	2
学習率	0.001
バッチサイズ	5
受容野の大きさ	64 ms
Dilation の数	$2^0 - 2^9$
Residual のチャネル数	32
Dilation のチャネル数	32
量子化のチャネル数	256
スキップ接続のチャネル数	512

では RWCP 実環境音声・音響データベース (RWCP-SSD) [26] の中から，表 4.3 に示す 10 種類の音響イベントを計 1,000 音用いた。それぞれの音響イベントのうち，95 音をモデル学習用に，5 音を評価用に用いた。なお，RWCP-SSD に含まれる環境音はクラス内の音が非常に類似しており，比較的高品質に音の合成が行えると考えられる。合成音は，上記の音響イベントラベルを one-hot 表現したものを conditional WaveNet へ入力することで得た。WaveNet に利用したパラメータを表 5.1 に示す。また，合成した環境音のうち品質が極めて悪いもの（無音のサンプルなど）は評価実験に利用しないこととした。合成された環境音のサンプルは [45] より聴取可能である。

各主観評価実験は 24 名の被験者数に対して実施し，5.3.2 項の合成された環境音の了解性に関する評価では 24（人）× 10（音響イベントの種類）× 5（音）× 2 = 2,400 サンプル，5.3.3 項の実在する音としての自然性の評価では 24（人）× 10（音響イベントの種類）× 2（順序入れ替え）= 480 サンプル，5.3.4 項の環境音としての自然性の評価では 24（人）× 10（音響イベントの種類）× 2（音）× 2 = 960 サンプルの合成音を用いた。また，評価実験にはオーディオインターフェースとして Roland QUAD-CAPTURE UA-55，ヘッドホンとして SONY MDR-CD900ST を使用した。評価実験は静音環境にて実施した。また，RWCP-SSD に含まれる環境音は 16 ビット量子化されているものを利用した一方，WaveNet による合成音は μ-law アルゴリズムにより量子化されているため 8 ビット量子化である点に注意が必要である。

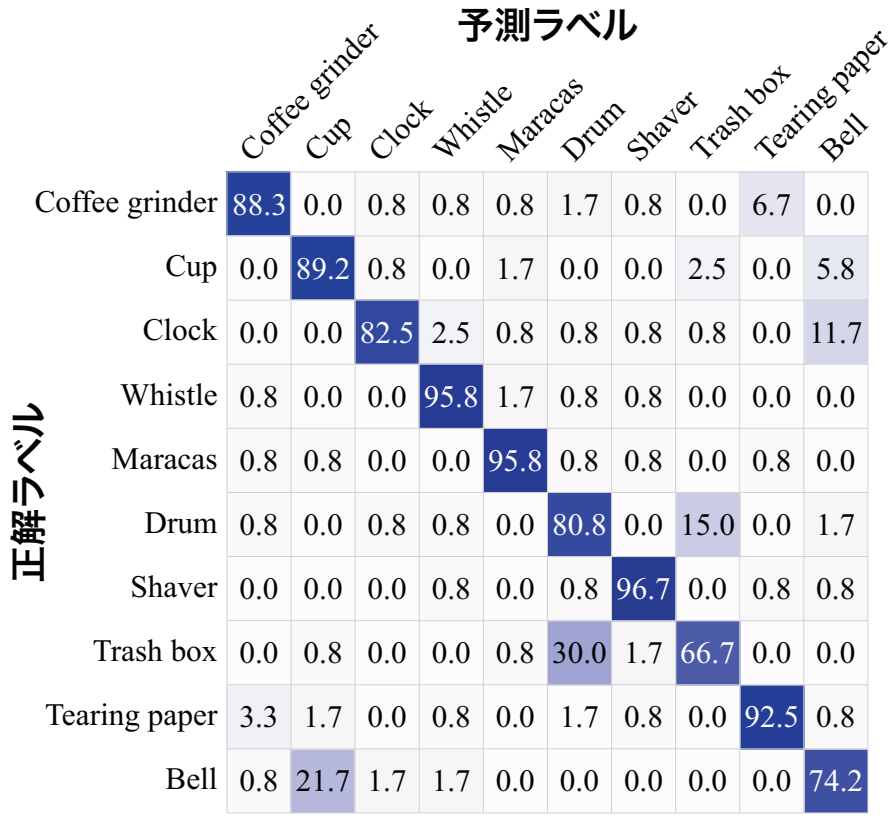


図 5.3: 自然音に対して被験者が回答した音響イベントラベルの正解率

5.3.2 合成された環境音の了解性に関する評価

提示された合成音に対して，学習に使用した音響イベントラベルのうち，どのラベルに一番適しているか選択させ，合成された音に対する認識性能を評価した。比較として自然音に対する評価も実施した。

自然音と合成された環境音の認識性能（再現率）をそれぞれ図 5.3 および図 5.4 に示す。また，すべての音響イベントに対する再現率の平均は，自然音では 86.22%，合成音では 76.30%となった。結果より，ドラムを叩く音などの音響イベントでは，自然音と合成音で同等の認識性能が得られた一方で，カップを叩く音をベルを鳴らす音に誤認識する例や，ひげ剃りの動作音を紙を引き裂く音に誤認識する例が多く見られた。図 5.5 に，自然音と合成された環境音のスペクトログラムを示す。図より，合成されたひげ剃りの動作音では微細なスペクトル構造が再現されておらず，紙を引き裂く音とよく似たスペクトル構造になっていることが確認でき，誤認識の原因になっていることが分かる。これらの結果より，現状の環境音合成手法を用いることで，比較的了解性の高い環境音が合成可能である一方，微細なスペクトル構造の再現には至っておらず，今後の合成手法の改善が必要と言える。ま

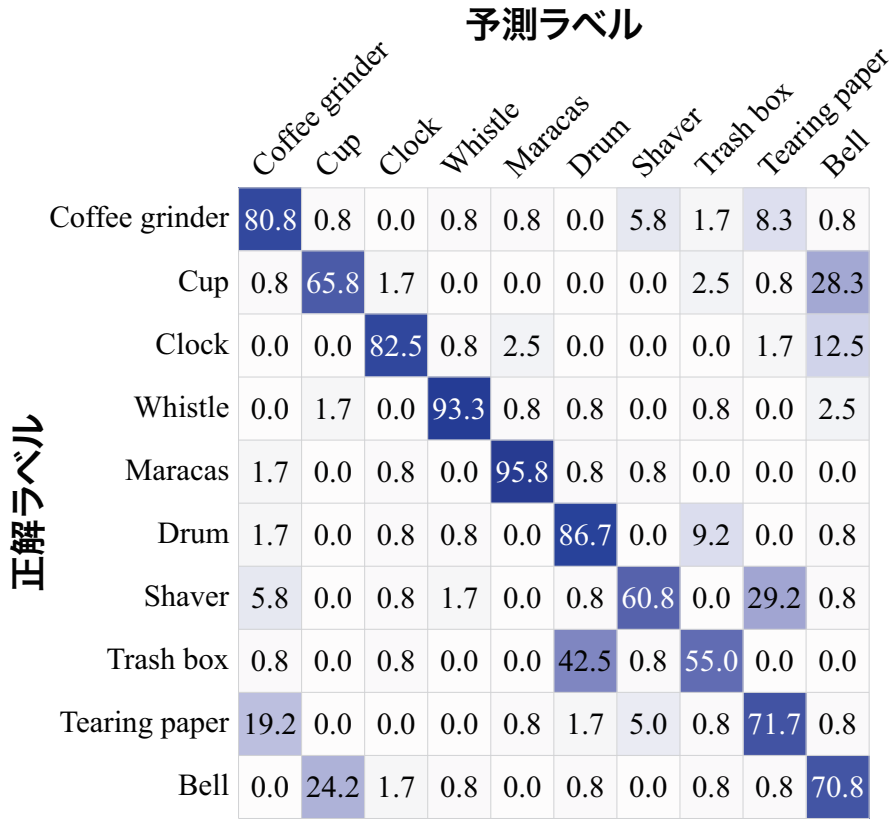


図 5.4: 合成音に対して被験者が回答した音響イベントラベルの正解率

た，環境音の了解性評価は，類似した複数の環境音を区別可能な音として合成し得るか，を評価する指標として特に有効と言える。

5.3.3 実在する音としての自然性の評価

自然音，合成音の 2 音をペアとして，ランダムな順番で被験者に聴かせ，自然音と感じる方を選択するプリファレンス AB テストを実施した。

各音響イベントの音に対して被験者が自然音を認識した割合を図 5.6 に示す。結果より，被験者は 82.71%の比較対について自然音を識別することができた。このことより，音声の合成で非常に高い音声品質を実現している WaveNet においてさえ，自然音と区別できない品質の環境音を合成するには至っていないことが確認できる。

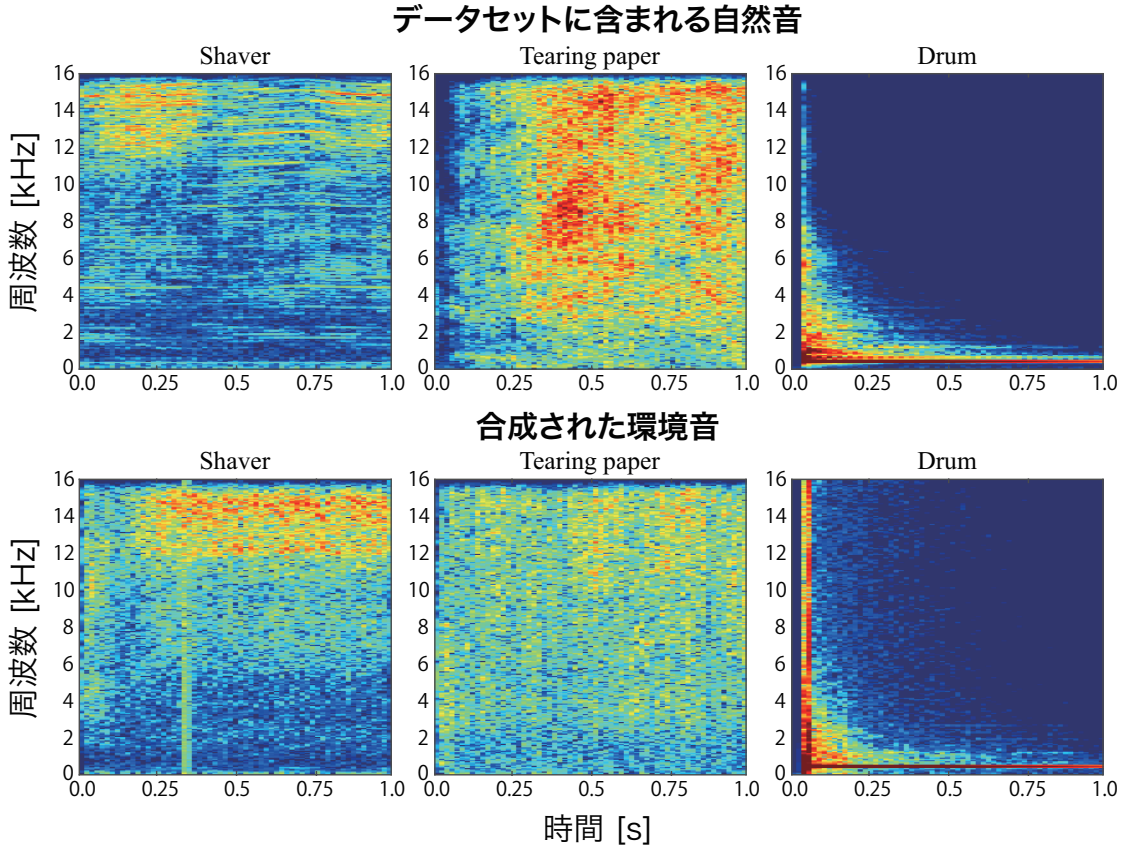


図 5.5: 自然音と合成音のスペクトログラム

5.3.4 環境音としての自然性の評価

自然音と合成音をランダムに被験者に聴かせ，1（環境音として非常に不自然である）から 5（環境音として非常に自然である）の 5 段階で音の自然性（実際に存在しそうな音であるか，音として違和感がないか）について，評価した。

自然音および，合成された環境音の自然性に関する評価の平均スコアと 95%信頼区間を図 5.7 に示す。結果より，コーヒー豆をミルで挽く音や目覚まし時計の音，マラカスの音では自然音と合成音で同程度の自然性が得られた一方で，ひげ剃りの動作音や金属製のゴミ箱を叩く音では自然性の評価結果に大きな開きが見られた。その理由として，合成されたひげ剃りの動作音などでは，微細なスペクトル構造が再現されていないことが原因と考えられる。また，笛の音のように，音の了解性は高いにも関わらず自然性が低いと判断される環境音もあることから，環境音の品質評価については，環境音の了解性の評価のみでなく，自然性に関する評価も行うべきと言える。

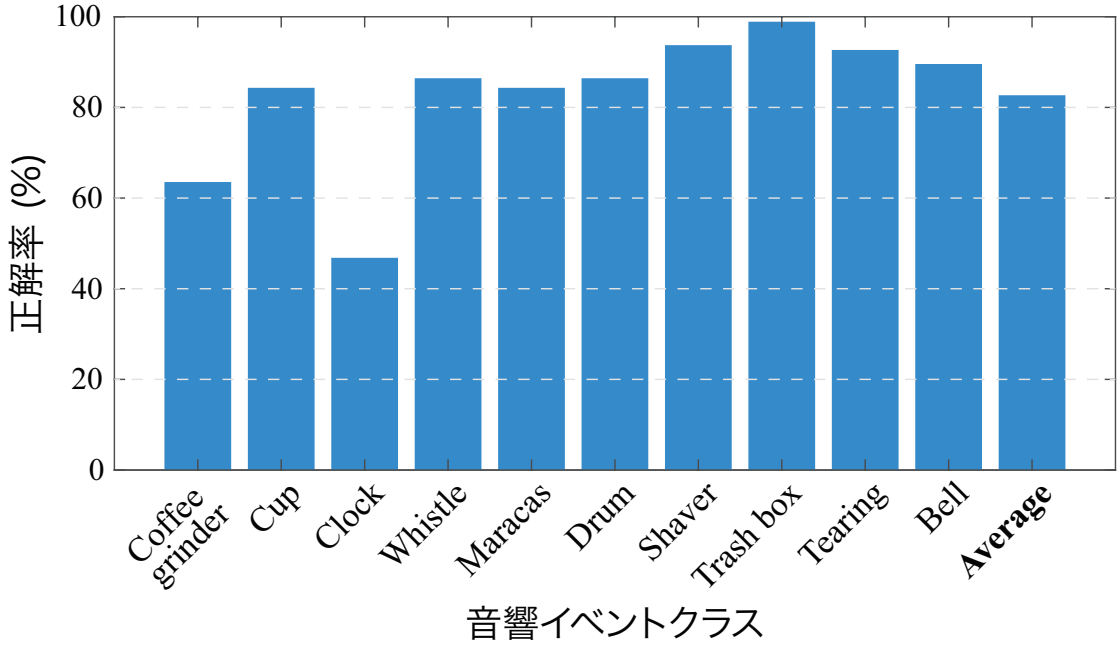


図 5.6: 各音響イベントの音に対して被験者が自然音を認識した割合

5.4　5章のまとめ

本章では，音響イベントラベルを入力とする環境音合成手法を提案した。主観評価実験より，コーヒーミルで豆を挽く音，目覚まし時計の音，マラカスの音の合成音では自然音と同程度の自然性を得られた。このことより，統計的手法を用いて環境音を合成可能であることを示唆した。一方，自然音と合成音を識別する主観評価実験においては，82.71%の比較対で自然音を識別できるという結果となった。そのため現状では，自然音と区別できないほど高い品質で環境音を合成するには至っていない。今後，高品質な環境音合成技術を実現するための手法の改良が必要である。

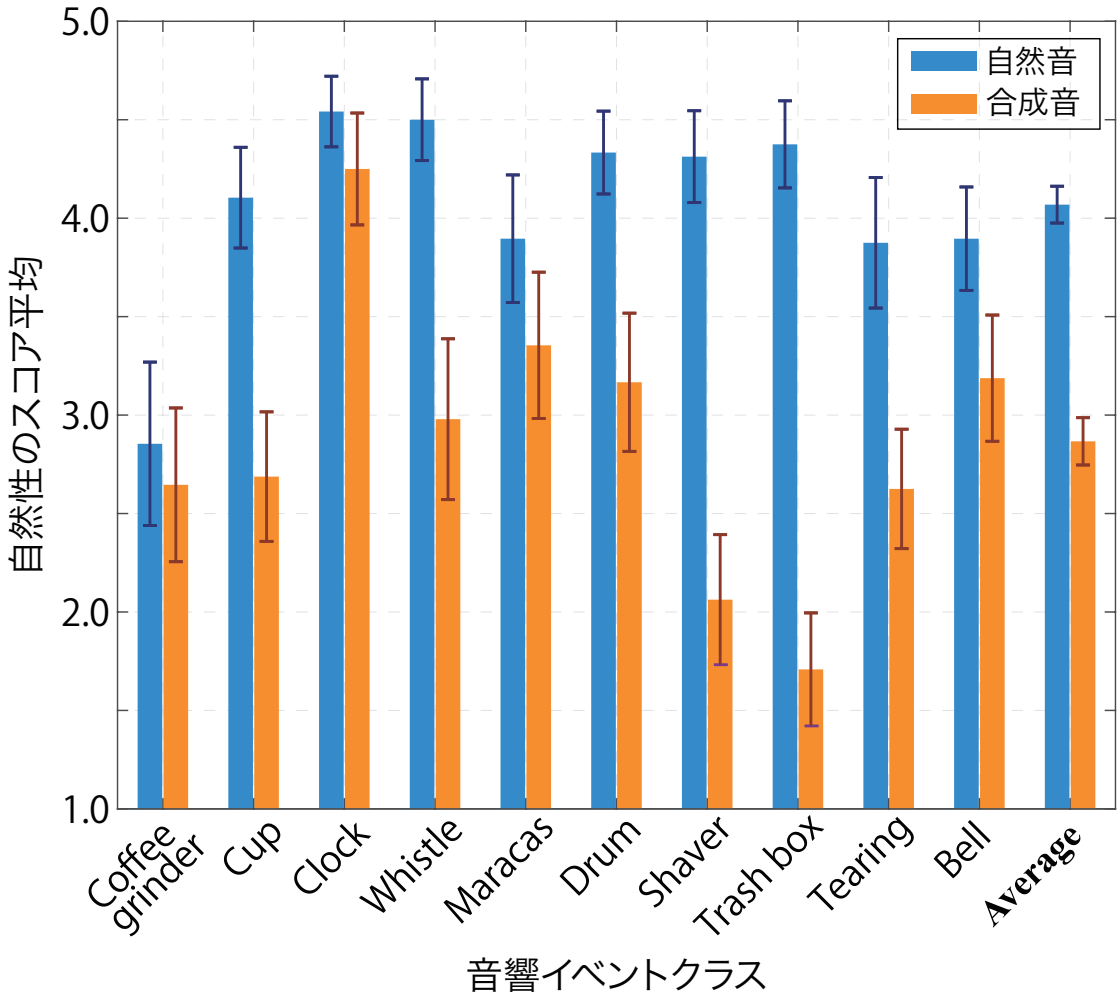


図 5.7: 自然音と合成音に対する自然性のスコア平均

第6章　オノマトペからの環境音合成

6.1　はじめに

本章では，3 章で構築した環境音にオノマトペが対応付いたデータセットを用いて，統計的手法によるオノマトペからの環境音合成手法を提案する。音のオノマトペ表現の特徴として，「ピュ」や「ピイイイ」のように文字列の長さや繰り返しによって音の継続長や繰り返し回数などといった時間的変化を表現可能であることが挙げられる。そのため，オノマトペを利用した環境音合成を実現することで，5 章で提案した音響イベントラベルのみを入力とする手法では制御困難であった音の時間的な変化をより柔軟に表現した環境音の合成が期待できる。オノマトペから環境音を合成する従来手法として，KanaWave[46] が存在する。しかしながら，KanaWave の合成音はオノマトペと音が 1 対 1 で対応づいたデータ内より，入力されたオノマトペに応じて音を接続して合成するため，実環境で発生する音としては自然性，多様性に欠ける。そのため，より自然かつ多様な音を合成するための合成手法が必要となる。

本章では，オノマトペからの環境音合成を行うため，系列変換モデル [37](以降 seq2seq と省略する) を利用する。seq2seq モデルを基本とした様々な手法は系列間変換を目的とする多くの研究において高い性能を示している [47, 48, 49, 50]。従来手法では，１つのオノマトペに対して決まった１音が対応付いているため，オノマトペ系列中の前後の音素の関係を考慮できていない。一方，seq2seq は再帰的構造により入力オノマトペの時系列情報をモデリングできるため，前後の出現音素に柔軟に対応した自然かつ，多様な音の合成が期待できる。また，オノマトペに加えて音響イベントラベルも入力として使用することで，音の時間的な変化だけでなく音源の種類も同時に制御する手法も提案する。

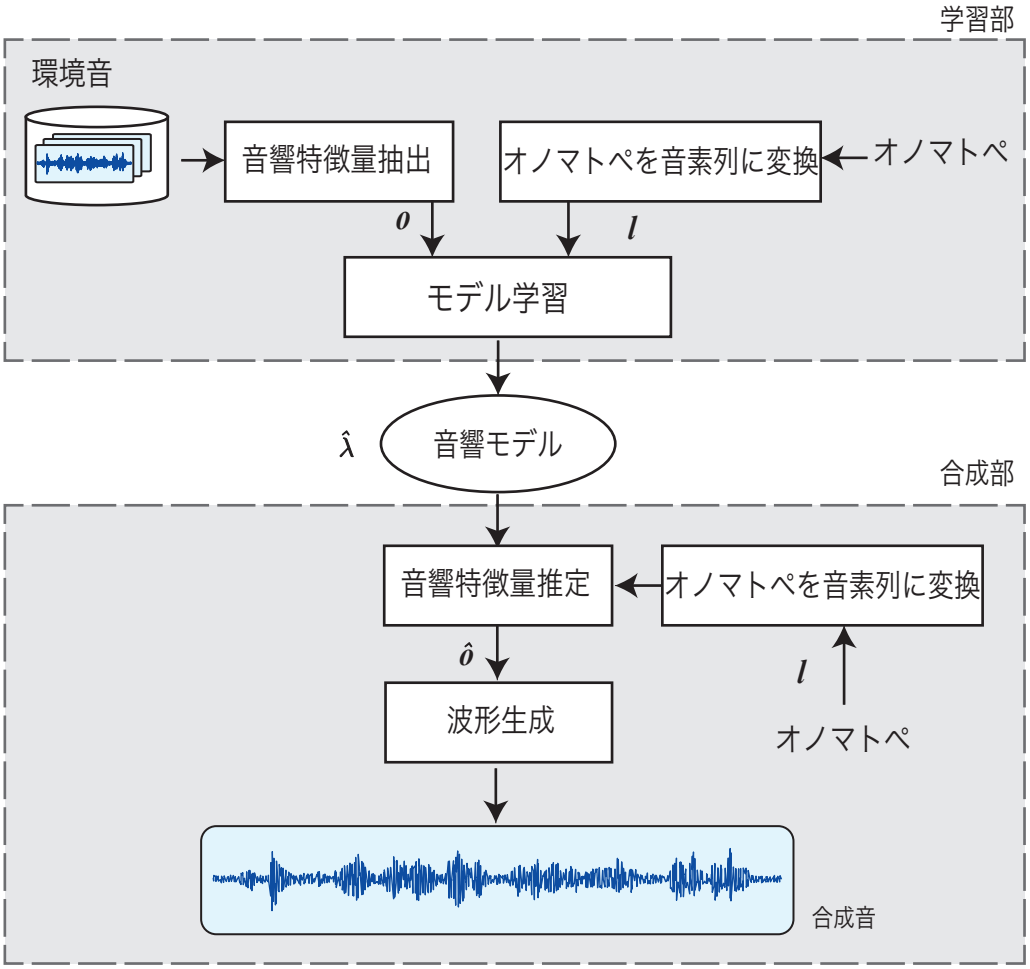


図 6.1: 統計的手法によるオノマトペからの環境音合成の概要

6.2 統計的手法によるオノマトペからの環境音合成手法

6.2.1 提案手法の概要

図 6.1 に提案するオノマトペからの環境音合成手法の概要を示す。提案手法は，学習部と合成部から構成される。学習部において環境音から音響特徴量系列 $\boldsymbol{o}$，オノマトペから音素系列 $\boldsymbol{l}$ を抽出し，それらの特徴量を用いて音響モデル $\boldsymbol{\lambda}$ を学習する。

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}} P(\boldsymbol{o} \mid \boldsymbol{l}, \boldsymbol{\lambda}) \tag{6.1}$$

音響モデル $\hat{\boldsymbol{\lambda}}$ の構築方法に関しては 6.2.2 項，6.2.3 項にて詳細を述べる。合成部では，合成に使用するテキスト表記されたオノマトペを音素系列 $\boldsymbol{l}$ に変換し，対応

する音響特徴量系列 $\boldsymbol{o}$ を音響モデルより統計的に推定する。

$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} P(\boldsymbol{o} \mid \boldsymbol{l}, \hat{\boldsymbol{\lambda}}) \tag{6.2}$$

最後に，式 (6.2) によって推定した音響特徴量系列 $\hat{\mathbf{o}}$ を環境音の波形に変換する。本章では Griffin-Lim 法 [51] を用いて環境音の波形に変換する。

6.2.2 オノマトペのみを入力とする音響モデルの構築

図 6.2 にオノマトペのみを入力とする音響モデル $\hat{\boldsymbol{\lambda}}$ の構築手法の概要を示す。利用した seq2seq モデルは 1 層の BiLSTM による encoder と，2 層の LSTM による decoder からなる。まず, 音素表記されたオノマトペの系列 $\boldsymbol{l} = \{l_1, ...l_T\}$ が encoder に入力され, 特徴ベクトル $\boldsymbol{\nu}$ が抽出される。Encoder に BiLSTM を利用することにより，オノマトペ系列の長さが長い場合でも全体の特徴を捉えたような特徴ベクトル $\boldsymbol{\nu}$ の抽出が期待できる。そして，抽出された特徴ベクトル $\boldsymbol{\nu}$ に基づき，decoder において各時刻における音響特徴量 $\boldsymbol{o} = \{\boldsymbol{o}_1, ..., \boldsymbol{o}_{T'}\}$ の推定を行う。なお本章においては，5 章で紹介した波形を直接推定する WaveNet とは異なり，音声合成にて高い品質を獲得している Tacotron [48] のモデル構造を参考に，音響特徴量を推定するモデル構築を行う。

$$p(\boldsymbol{o}_1, ..., \boldsymbol{o}_{T'} \mid l_1, ..., l_T) = \prod_{t=1}^{T'} p(\boldsymbol{o}_t \mid \boldsymbol{\nu}, \boldsymbol{o}_1, ..., \boldsymbol{o}_{t-1}) \tag{6.3}$$

Decoder の各時刻において推定された音響特徴量系列と教師データの時刻に対応する音響特徴量系列の L1 ノルムを誤差関数とすることでモデル学習を行う。

6.2.3 オノマトペと音響イベントラベルを入力とする音響モデルの構築

6.2.2 項で述べたオノマトペのみを入力とする手法では，音の継続長や繰り返し回数など時間的変化を制御することが期待できる。一方で，「パン (/p a N/)」というオノマトペのみを入力とした場合,「銃を撃つ音」,「風船が割れる音」など複数の音響イベントに対応するため，音源の種類といった周波数特性を制御することが困難であると考えられる。そこでオノマトペだけでなく，音響イベントラベルも入力として用いることで音源の種類の制御も同時に期待できる。

図 6.3 にオノマトペと音響イベントラベルを入力とする音響モデル $\hat{\boldsymbol{\lambda}}$ の構築手

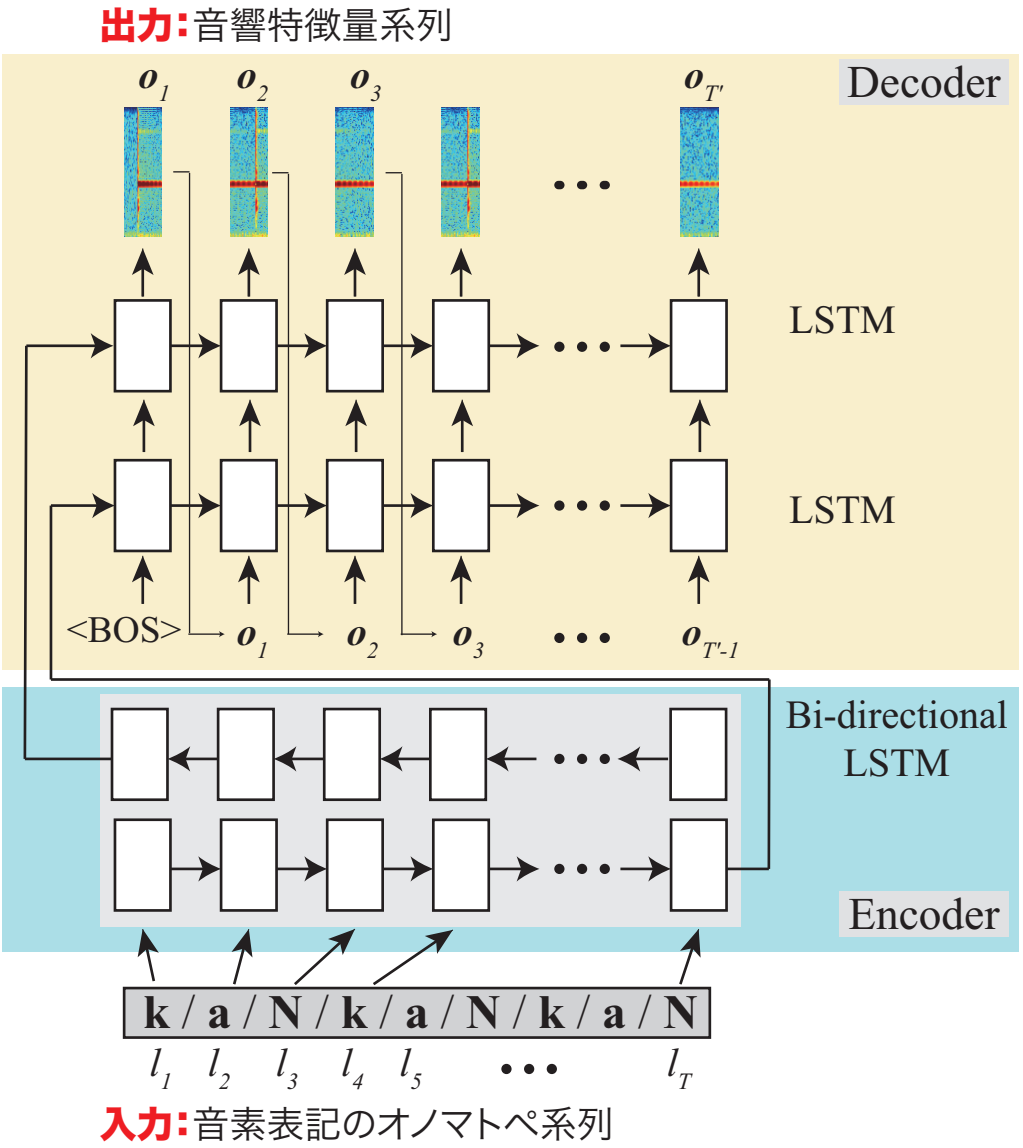


図 6.2: オノマトペのみを入力とするモデル学習

法の概要を示す。利用した seq2seq モデルはオノマトペのみを入力する手法と同じく，1 層の BiLSTM による encoder と，2 層の LSTM による decoder からなる。seq2seq による系列間変換では decoder からの出力を制御するため，decoder への条件付けを行うことがある [52, 53]。そこで本手法では，encoder において抽出したオノマトペの時間的変化を表す特徴ベクトルに対し，one-hot 表現された音響イベントラベルを連結し decoder の初期状態として与える。それによって音響イベントの情報も考慮したモデル学習を行うことができると考える。Encoder で抽出した特徴ベクトル $\boldsymbol{\nu}$ と音響イベントラベル $\boldsymbol{c}$ に基づき，decoder において各時刻に

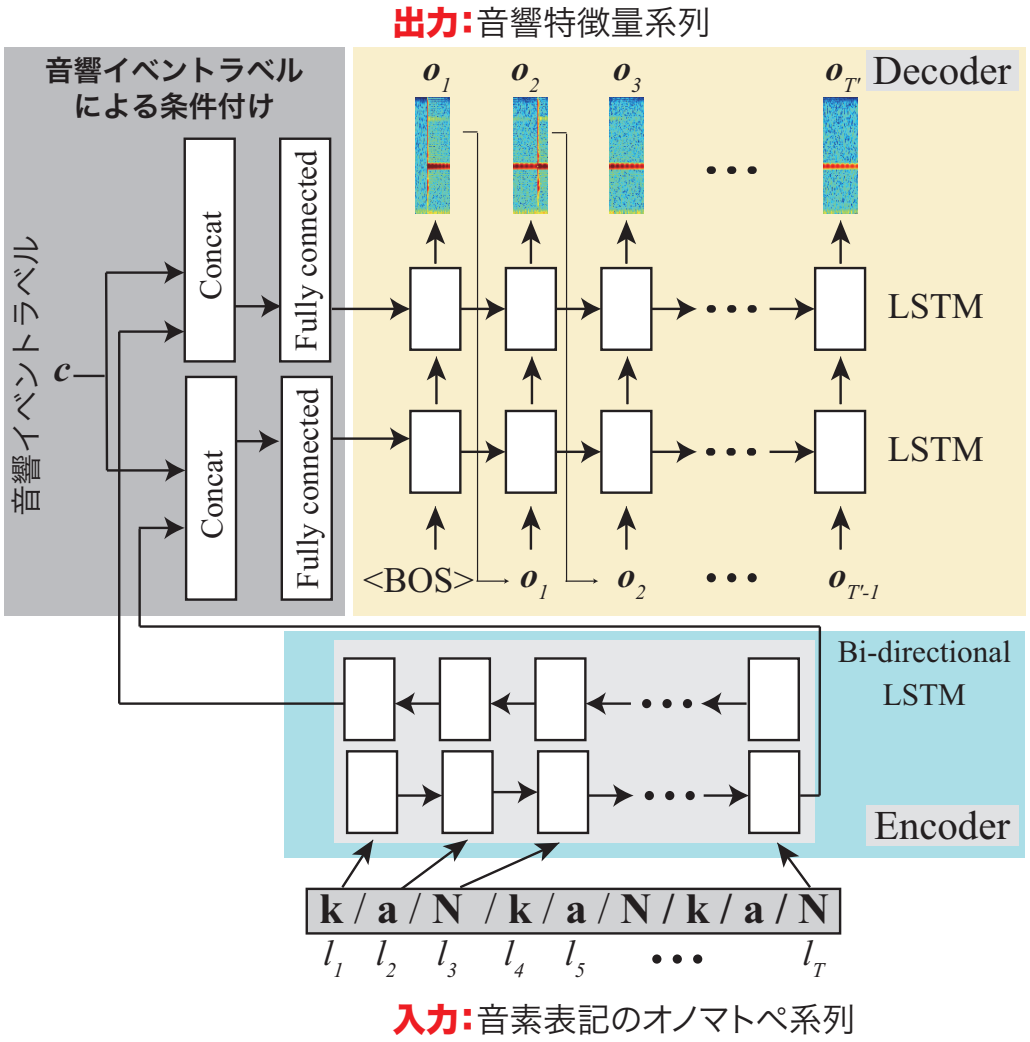


図 6.3: オノマトペと音響イベントラベルを入力とするモデル学習

おける音響特徴量 $\boldsymbol{o} = \{\boldsymbol{o}_1, ..., \boldsymbol{o}_{T'}\}$ の推定を行う。

$$p(\boldsymbol{o}_1, ..., \boldsymbol{o}_{T'} \mid l_1, ..., l_T) = \prod_{t=1}^{T'} p(\boldsymbol{o}_t \mid \boldsymbol{\nu}, \boldsymbol{o}_1, ..., \boldsymbol{o}_{t-1}, \boldsymbol{c}) \tag{6.4}$$

Decoder の各時刻において推定された音響特徴量系列と教師データの時刻に対応する音響特徴量系列の L1 ノルムを誤差関数とすることでモデル学習を行う。

表 6.1: seq2seq のパラメータ設定および利用した音響特徴量

音の長さ	1–2 s
サンプリング周波数	16,000
音波形の圧縮形式	16-bit linear PCM
音響特徴量	対数振幅スペクトログラム
フレーム長	0.128 s（2,048 サンプル）
フレームシフト	0.032 s（512 サンプル）
Encoder の LSTM 層の数	1
Encoder の LSTM 層のユニット数	512
Decoder の LSTM 層の数	2
Decoder の LSTM 層のユニット数	512, 512
バッチサイズ	5
音響イベントラベルの次元数	10
Teacher forcing の比率	0.6
最適化手法	RAdam

表 6.2: 各評価実験に使用した合成音のサンプル数と被験者数

実験	音響イベント数	音響イベントごとの音の数	評価者数	合計の音の数
実験 I-1	10	10	30	3,000
実験 I-2	10	10	30	3,000
実験 I-3	10	5	30	1,500
実験 II-1	5	5	30	750
実験 II-2	10	2-3	50	1,300

6.3 評価実験

映画やゲームなどのメディアコンテンツにおける背景音・効果音として使用される環境音は，高い自然性と多様性が求められる。よって，(I) 自然性，(II) 多様性の 2 つの観点から主観評価実験を実施した。

6.3.1 実験条件

本実験では，環境音として RWCP 実環境音声・音響データベース (RWCP-SSD)[26] の中から，表 4.3 に示す 10 種類の音響イベントを計 1,000 音用いた。それぞれの音響イベントのうち，95 音をモデル学習用，5 音を評価用に用いた。オノマトペのデータは，3 章にて作成したデータセットを使用した。モデル学習には，1 音につき 15 個，計 14,250 個（15 個 $\times$ 950 サンプル）のオノマトペを使用した。seq2seq によるモデル学習のパラメータは表 6.1 に示す。また，本章においては，音響特徴量として対数振幅スペクトログラムを用いる。

表 6.3: 各実験にて評価した環境音合成手法の一覧

合成手法	実験 I-1	実験 I-2	実験 I-3	実験 II-1	実験 II-2
Natural sound	✓	✓	✓		
WaveNet			✓	✓	
KanaWave	✓	✓	✓		
Seq2seq （提案手法）	✓	✓	✓		✓
Seq2seq + event label （提案手法）	✓	✓	✓	✓	✓

各実験は，クラウドソーシングサービスを用いて実施した。表 6.2 に各評価実験における合成音のサンプル数と被験者数を示す。3 章にて構築したデータセットは，日本語話者から収集したオノマトペのみが含まれている。環境音に対して付与されるオノマトペは母国語によって異なるため，合成音の評価は日本話者によってのみ評価する。

表 6.3 に各実験にて評価した環境音合成手法の一覧を示す。提案手法との比較のために，データセットに含まれる自然音 (natural sound) 5 章にて提案した音響イベントラベルのみから合成された環境音，KanaWave [46] によって合成された環境音に対しても評価を実施する。KanaWave は，オノマトペのみから環境音を合成する非統計的な手法である。入力されたオノマトペの各文字に対応する音を，事前に用意された音データから選択して接続することで音を合成する。また、音の高さを変更するためのいくつかのパラメータを設定することが可能である。しかし，あらかじめ用意された複数の音を接続して音を合成するため，音同士の接続箇所が不自然になったりと，音としての自然性は低い。

6.3.2 環境音の自然性に関する評価

環境音の自然性の評価においては、(i) 環境音として違和感がない，(ii) 入力されたオノマトペを表現できている，という 2 つの観点から合成音を評価する。まず，実験 I-1 と I-2 では，環境音と入力に使用したオノマトペを被験者に提示して，オノマトペに対する環境音の許容度と表現性を評価する。そして，実験 I-3 では，環境音として違和感がないかを評価するために，音のみを提示して，環境音自体の品質に関して評価する。

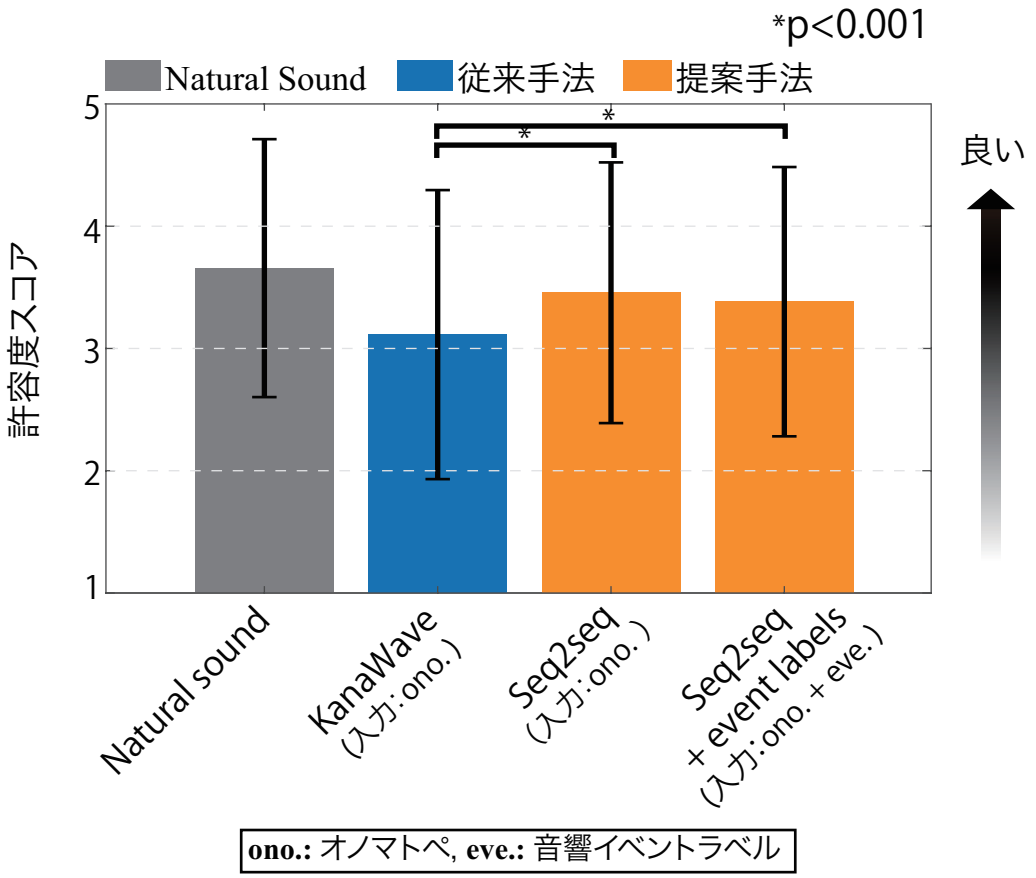


図 6.4: オノマトペに対する環境音の許容度に関する評価結果

- **実験 I-1：オノマトペに対する環境音の許容度**
  環境音とオノマトペを被験者に提示する。被験者は，オノマトペに対する環境音の許容度を 1（非常に許容できない）〜5（非常に許容できる）の 5 段階で評価する。
- **実験 I-2：オノマトペに対する環境音の表現性**
  環境音とオノマトペを被験者に提示する。被験者は，オノマトペに対する環境音の表現性を 1（非常に表現できていない）〜5（非常に表現できている）の 5 段階で評価する。
- **実験 I-3：環境音の自然性**
  環境音のみを被験者に提示する。被験者は，環境音の自然性について 1（非常に不自然である）〜5（非常に自然である）の 5 段階で評価する。

実験 I-1，I-2 の許容度並びに表現性に関する平均スコアと標準偏差をそれぞれ図 6.4 と 6.5 に示す。これらの結果より，提案手法は KanaWave で合成される音よ

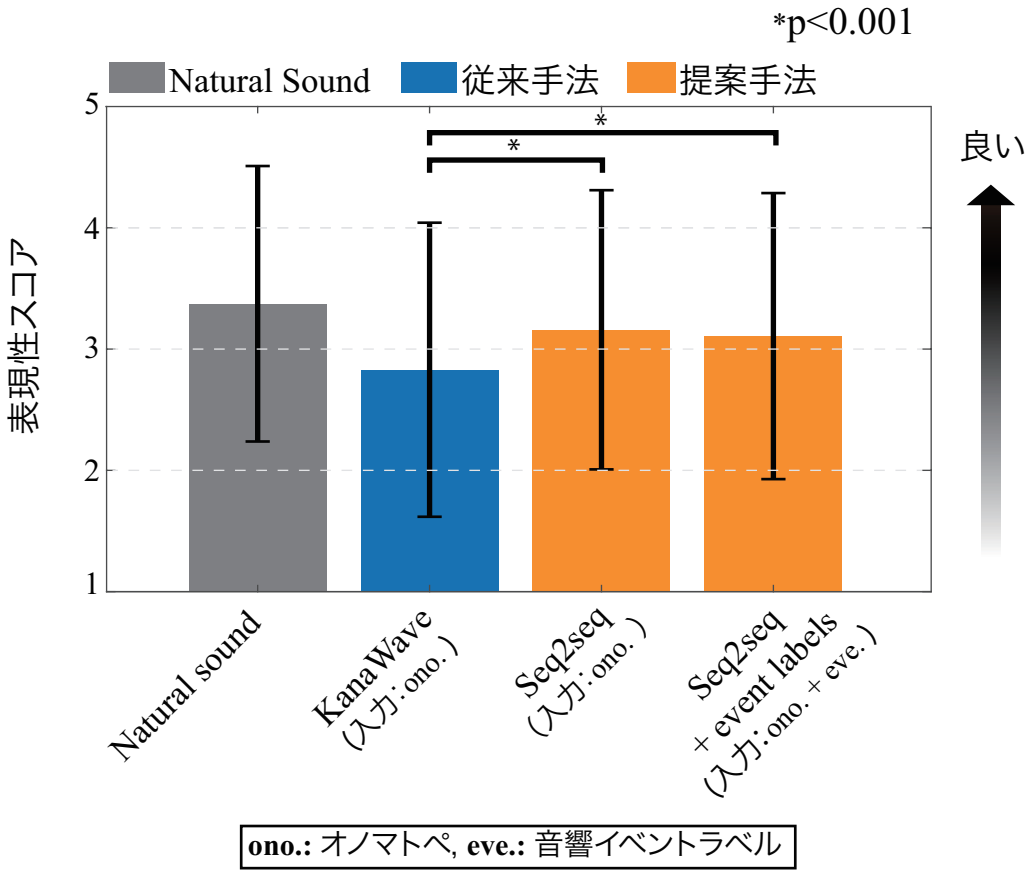


図 6.5: オノマトペに対する環境音の表現性に関する評価結果

りも入力となったオノマトペを表現した環境音を合成できることがわかる。

図 6.6 にオノマトペのみを用いた提案手法によって合成した環境音のスペクトログラムを示す。図に示すように，提案手法はオノマトペを変化させることで多様な環境音を合成することができる。また，オノマトペ系列の長さによって，合成される環境音の長さを変化させることが可能であることもわかる。このように，環境音合成の入力情報としてオノマトペを利用することは，音の長さなど時間的変化を制御するために有効であることが確認できる。

図 6.7 に KanaWave 並びにオノマトペと音響イベントラベルを入力とする提案手法による合成音のスペクトログラムを示す。図における各合成手法には「ビイイイ /b i i i i/」というオノマトペを入力した。オノマトペと音響イベントラベルを入力とする提案手法では，音響イベントラベルとして「笛の音」,「ひげ剃りの動作音」,「紙を引き裂く音」を用いた。なお，KanaWave は同一のオノマトペからは 1 種類の環境音しか合成されない。そのため，同一オノマトペを入力して KanaWave によって合成される音には多様性がない。一方，提案手法は，オノマ

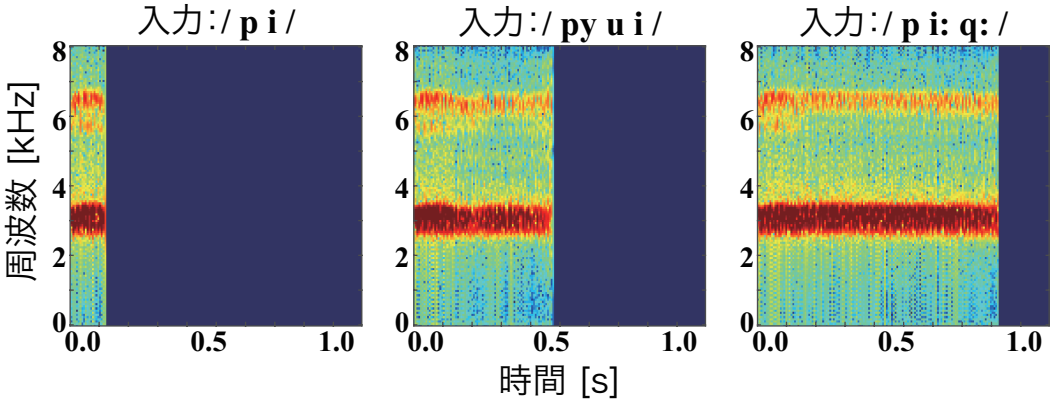


図 6.6: オノマトペのみを用いた提案手法によって合成した環境音のスペクトログラム

トペに加えて音響イベントラベルも入力することで，同一のオノマトペからでも多様な環境音を合成することが可能である。

実験 I-3 の自然性の平均スコアと標準偏差を図 6.8 に示す。図より，提案手法によって合成された音は，KanaWave によって合成された音と比較して高い自然性を獲得していることがわかる。また，統計的有意差検定の結果，提案手法によって合成された音は WaveNet によって合成された音と同程度の自然性であることが示された。これらの結果より，提案手法は従来手法と比べて音質を劣化させることなくオノマトペから環境音を合成可能であることがわかる。一方，natural sound は提案手法よりも高い自然性を獲得している。よって，今後 natural sound と同等の自然性を実現できる環境音合成手法を実現する必要がある。

6.3.3 環境音の多様性に関する評価

環境音の多様性を評価するために，以下に示す 2 つの実験を実施した。

- **実験 II-1：音響イベントクラス内における環境音の多様性**
  各音響イベントクラスごとに同一の手法によって合成された 2 音の環境音を被験者に提示する。被験者は，提示された 2 音がどの程度類似していないかを 1（非常に似ている）〜5（非常に似ていない）の 5 段階で評価する。なお，提案手法では，各音響イベントごとにデータセットからランダムに選択されたオノマトペを入力として音を合成した。
- **実験 II-2：同一オノマトペから合成された環境音の多様性**
  環境音と 10 個の音響イベントラベルを被験者に提示する。被験者は，提示

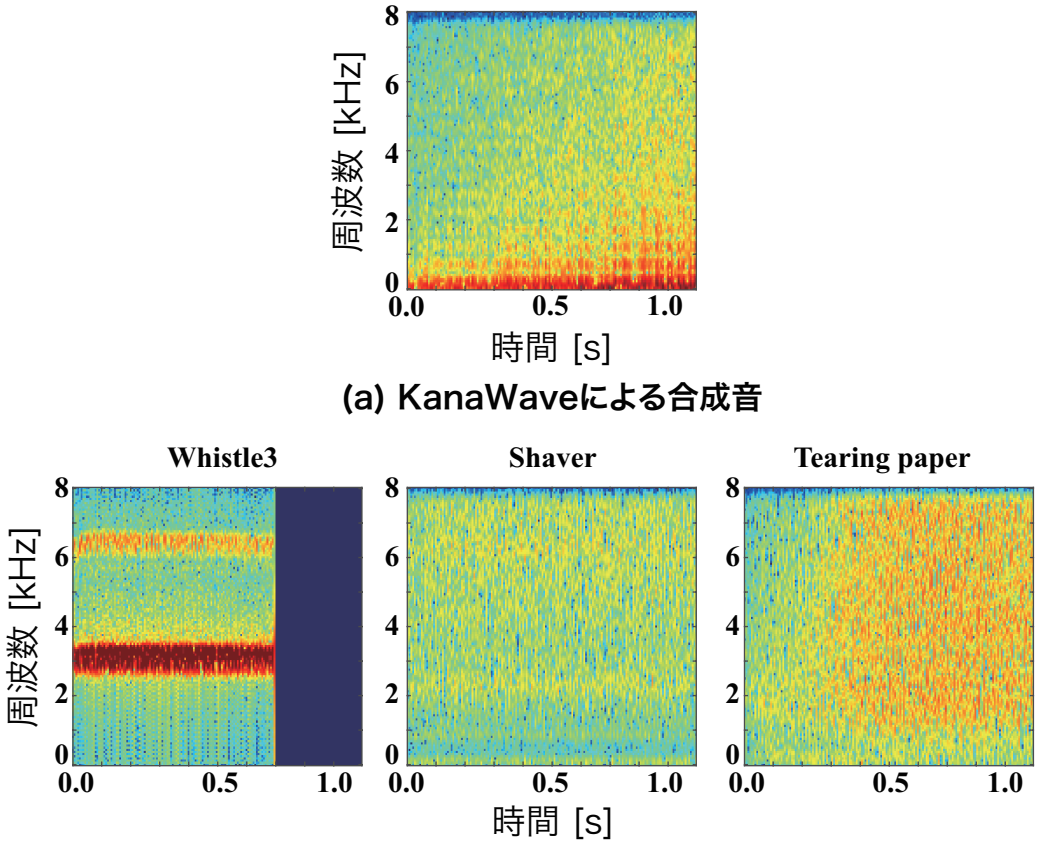


図 6.7: KanaWave 並びにオノマトペと音響イベントラベルを同時入力とする提案手法による合成音のスペクトログラム

された環境音を最もよく表すと感じる音響イベントラベルを 1 個回答する。なお，提示した 10 個の音響イベントラベルは，表 4.3 に示すラベルである。

実験 II-1 の各音響イベントごとの平均スコアを図 6.9 に示す。本実験においてスコアが高いことは，同一音響イベントクラス内における合成音が多様であることを意味する。結果より，音響イベントラベルとオノマトペを入力とする提案手法 (seq2seq + w/ event labels + onomatopoeiac words) は，音響イベントラベルのみを提案手法と同一のモデルに入力とする合成手法 (seq2seq + w/ event labels) よりも，多様な環境音を合成できることがわかる。よって，オノマトペを利用することで多様な環境音が合成可能である。また，音響イベントラベルのみを入力とする WaveNet を用いた手法と音響イベントラベルとオノマトペを入力とする提案手法を比較すると，提案手法のほうが「ドラムを叩く音」,「ひげ剃りの動作音」の音響イベントにおいて多様な環境音を合成できることがわかる。一方,「カップ

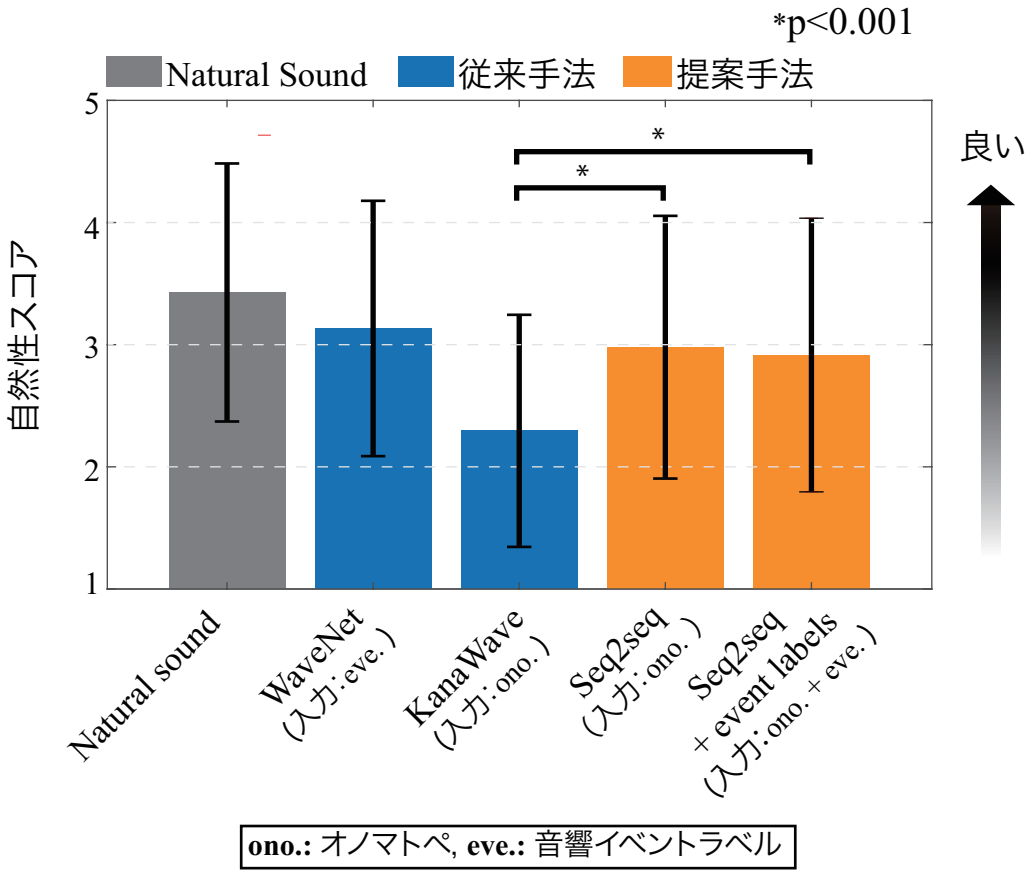


図 6.8: 環境音に対する自然性の評価結果

を叩く音」と「マラカスの音」の音響イベントでは WaveNet によって合成された音と同程度の多様性であることがわかった。WaveNet を用いた合成手法では，合成音にノイズが含まれる傾向にあった。そのため，そのノイズによって WaveNet による合成音のスコアが高くなる傾向があったと考えられる。一方，提案手法による合成音はノイズが少なく，多様で比較的明瞭な音を合成することが可能である。このように，オノマトペを環境音合成に用いることで，より多様な環境音を合成可能であるわかった。

実験 II-2 の結果を図 6.10 に示す。図 6.10 (a) の結果より，オノマトペのみを用いる提案手法で合成した環境音には，1 種類の音響イベントラベルに回答が集まる傾向が確認された。一方，図 6.10 (b) の結果より，オノマトペと音響イベントラベルの両方用いる提案手法で合成された環境音に対しては，様々な音響イベントラベルに回答がバラつく傾向が確認された。回答された音響イベントラベルの分布のエントロピーは，オノマトペのみを用いる手法では 1.70 bit，オノマトペと音響イベントラベルの両方を用いる手法では 1.82 bit となった。なお，本実験では，

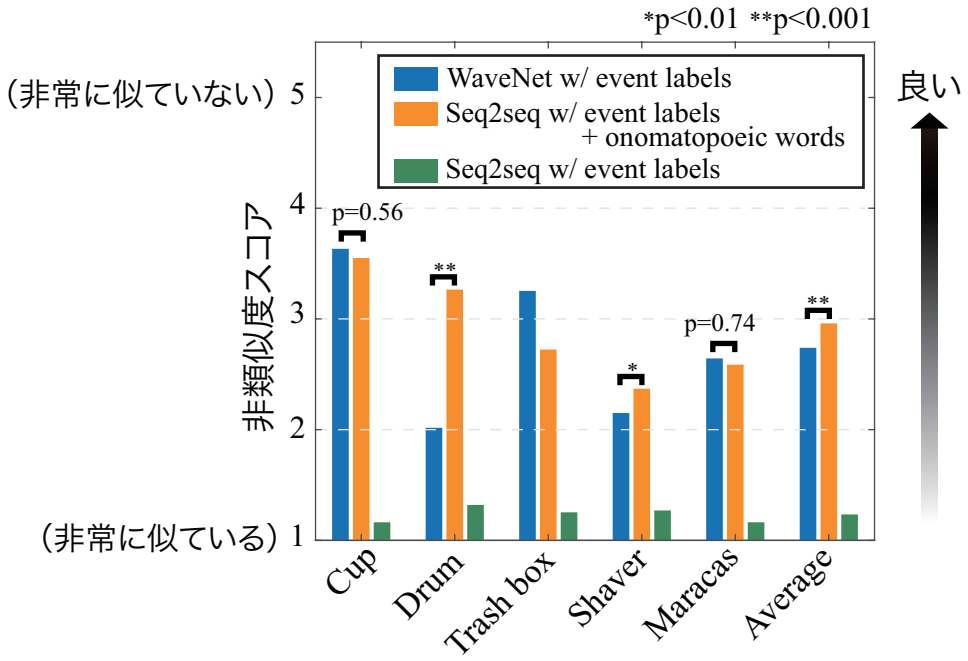


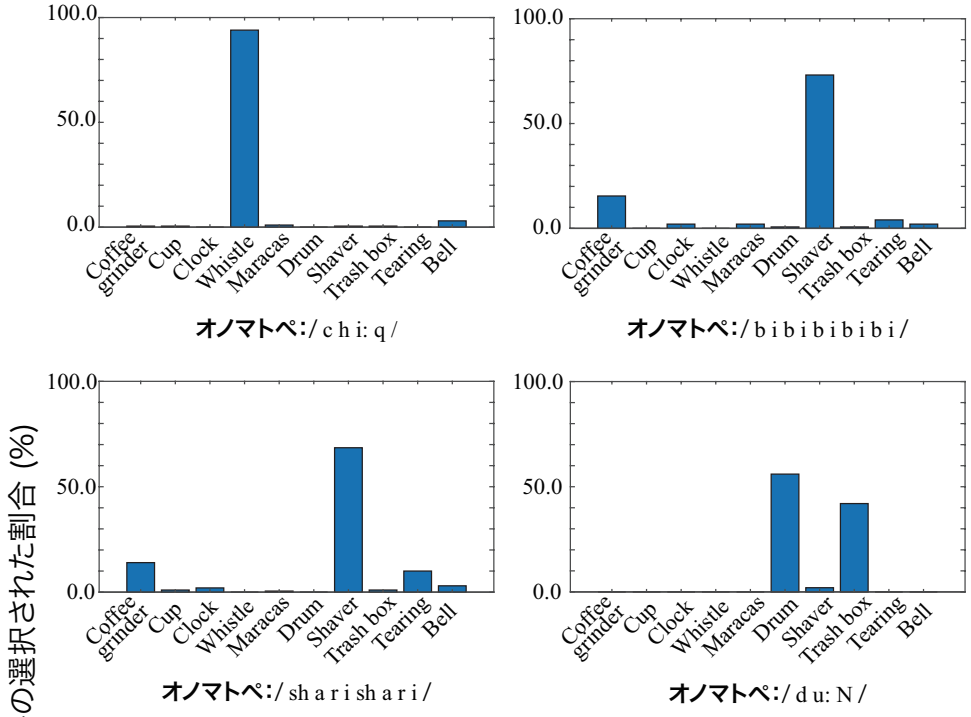
図 6.9: 音響イベントクラス内における環境音の多様性の評価結果

各合成音に対して 10 種類の音響イベントが偏りなく出現する場合，最大でエントロピーは 3.02 bit となる。この結果より，オノマトペと音響イベントラベルの両方を使用することで，同一のオノマトペに対しても多様な音響イベントの音を表現した環境音を合成可能であることがわかる。

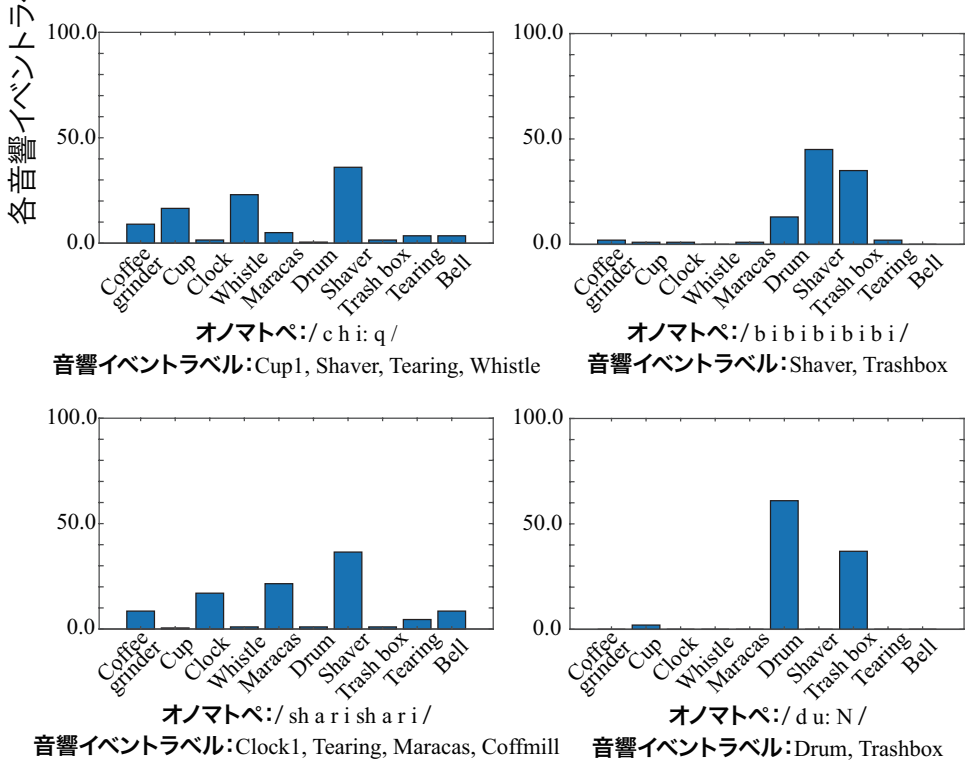
図 6.11 に「ビイッ (/ ー b i: i q /)」というオノマトペから合成された環境音のスペクトログラムを示す。オノマトペと音響イベントラベルを両方使用する提案手法では，音響イベントラベルとして「笛の音」，「ひげ剃りの動作音」，「紙を引き裂く音」を使用した。図に示すように，オノマトペのみ使用する提案手法によって環境音を合成する場合，合成時におけるモデルパラメータの初期値を変更しても，類似した特徴を持つ音ばかりが合成される。一方，オノマトペと音響イベントラベルの両方を使用すると，入力された音響イベントラベルに応じて，それぞれの音響イベントとオノマトペの特徴を捉えた環境音を合成できる。これらの結果より，音響イベントラベルを用いることで，オノマトペから合成される環境音の音源の種類を制御できることがわかる。

6.4　6章のまとめ

本章では，統計的手法によるオノマトペからの環境音合成手法を提案した。音響モデルの学習方法とし，オノマトペのみを入力とする手法とオノマトペと音響イベントラベルを入力とする手法を提案した。環境音の自然性に関する主観評価実験として，入力に使用したオノマトペに対する合成音の許容度と表現性に関して評価した。結果より、従来手法である KanaWave と比較して提案手法は許容度，表現性共に高いスコアを獲得した。また，環境音自体の自然性に関する結果より，提案手法は音響イベントラベルのみを入力とする WaveNet による手法と同程度の自然性を獲得した。一方，natural sound と比較すると提案手法による合成音は自然性が劣るため，今後さらに高品質な環境音合成手法の実現が必要である。環境音の多様性における主観評価実験として，音響イベントクラス内における合成音の多様性と同一オノマトペから合成される環境音の多様性について評価した。音響イベントクラス内における合成音の多様性の結果より，オノマトペを用いることで，音響イベントのみを使用した場合よりも多様な環境音が合成可能であることが明らかになった。また，同一オノマトペから合成される環境音の多様性の結果より，音響イベントラベルとオノマトペの両方を使用することで，同一のオノマトペを入力とした場合でも各音響イベントの特徴を表現した多様な環境音が合成可能であることがわかった。今後，合成音の品質向上を目指すと共に，より多くの音響イベントに対してオノマトペを収集して環境音合成を行う予定である。

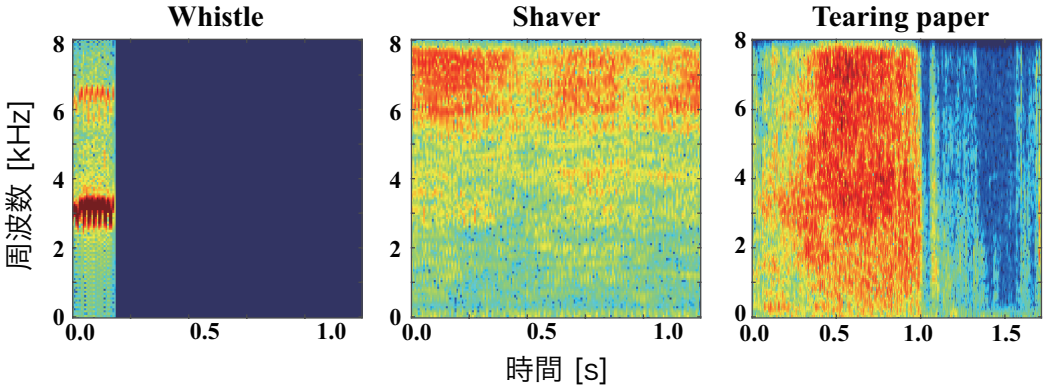


(a) オノマトペのみ入力とした提案手法による合成音

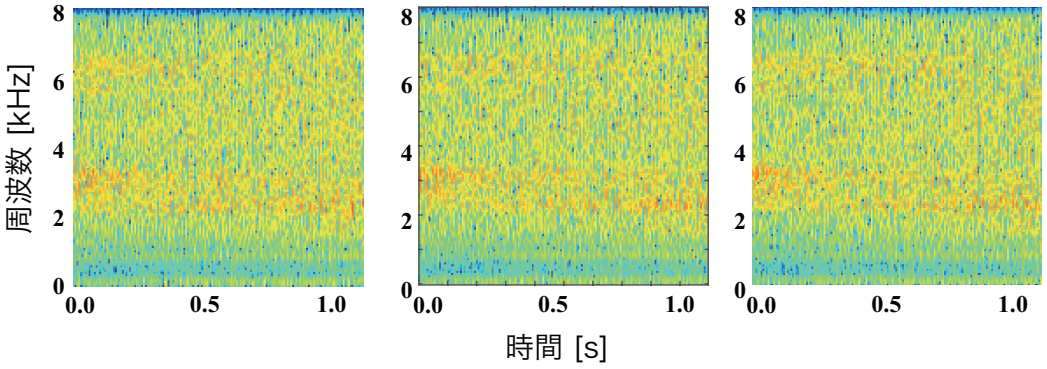


(b) オノマトペと音響イベントラベルを入力とした提案手法による合成音

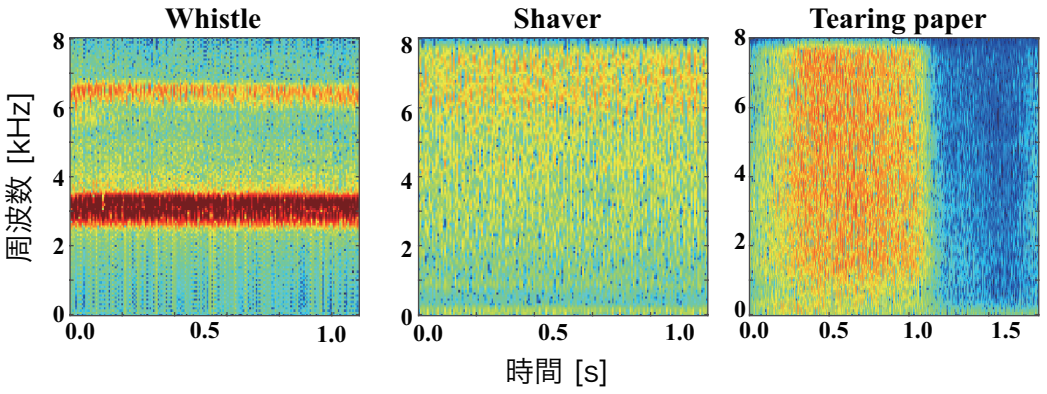
図 6.10: 同一オノマトペから合成された環境音の多様性の評価結果



(a) データセットに含まれる自然音 (Natural sound)



(b) オノマトペのみを入力とした提案手法による合成音



(c) オノマトペと音響イベントラベルを入力とした提案手法による合成音

図 6.11: ビイッ (/ ー b i: i q /) というオノマトペを各手法において入力した際の合成音のスペクトログラム

第7章　環境音を模倣した音声を用いた環境音合成

7.1　はじめに

5，6 章で提案した環境音合成手法では，音響イベントラベルによって合成音の音源の種類，オノマトペによって音の繰り返し回数などといった合成音の時間的な変化を制御することが可能となった。しかしながら，合成音の音高やリズムといった情報は音響イベントラベルやオノマトペでは表現することができない。例えば，ある音がオノマトペによって「カンカン」のように表現された場合，音の繰り返し回数は 2 回と断定することができるが，どのようなリズムの環境音かは表現できない。このように，音響イベントラベルやテキスト表記のオノマトペでは，合成したい音の特徴を表現しきれない場合もある。

環境音の音高やリズムを表現する方法として，環境音を声によって模倣した音声（以下，模倣音声）の利用が挙げられる [54]。模倣音声は，音の特徴を直感的に表現することが可能であるため，環境音の検索などの用途でも使用されている [55, 56, 57]。そのため，環境音合成においても模倣音声を利用することで，合成音の音高やリズムなど、音響イベントラベルやオノマトペでは表現困難であった特徴の制御が期待できる。

本章では，音響イベントラベルと模倣音声を同時入力とする環境音合成手法を提案する。音響イベントラベルに加えて模倣音声も利用することで，音響イベントラベルでは合成音の音源の種類，模倣音声によって音高やリズムの制御を行う。合成モデルには，ecoder–decoder 型の深層学習モデルを使用する。Encoder 部は特徴量抽出器とクラスタリングモデルからなる量子化器，decoder 部には音声合成にて高品質を達成している Tacotron2 [58] の decoder を使用することで，入力となる音響イベントラベルと模倣音声に対する環境音の対応関係をモデル化する。

7.2 関連研究

音声を利用した環境音合成の関連研究として，Takizawa らによる音声からの爆発音合成手法 [59] が挙げられる。この手法は，音声によって表現したニュアンスを反映した爆発音を合成することが可能である。しかしながら，爆発音のみを対象としており，その他の音響イベントクラスの音も合成可能であるかどうか明らかでない。また，合成モデルへの入力が音声のみであるため，1 つのモデルで複数の音響イベントに対応する音を合成するためには，クラスごとにモデルを学習させる必要がある。

従来手法 [59] では，Transformer モデルの encoder を用いて抽出したフレーム単位の連続表現を用いていた。しかしながら，入力された音声が雑音などによって劣化している場合，音声に含まれる雑音が合成音の品質に影響を与える可能性が懸念される。入力音声の劣化の影響を低減するために，音声合成・強調などにおいては，フレーム単位の離散表現を用いる手法が提案されている [60]。環境音合成においても，離散表現を用いることで，入力音声の劣化の影響を低減することが期待できる。

7.3 統計的手法による模倣音声を利用した環境音合成

提案手法は，合成モデル $\mathsf{Synthesizer}(\cdot)$ によって模倣音声 $\boldsymbol{x}$ と音響イベントラベル $\boldsymbol{c}$ から環境音 $\hat{\boldsymbol{y}}$ を推定する。

$$\hat{\boldsymbol{y}} = \mathsf{Synthesizer}(\boldsymbol{x}, \boldsymbol{c}) \tag{7.1}$$

合成モデルのモデルパラメータは，正解の環境音 $\boldsymbol{y}$ と $\{\boldsymbol{x}, \boldsymbol{c}\}$ のペアデータを用いて推定される。

図 7.1 に音響イベントラベルと模倣音声を入力とする環境音合成のモデル構造を示す。提案手法は，encoder と decoder から構成される。Encoder は，特徴量抽出器と k-menas クラスタリングモデル [61] から構成されており，模倣音声を量子化した音響特徴量を抽出する。特徴量抽出器には，Freesound Dataset 50K [62] によって事前学習された Bootstrap Your Own Latent for Audio (BYOL-A) [63] を用いた。BYOL-A を用いて抽出された音響特徴量の各時間フレームは，k-means クラスタリングモデルによって対応する環境音の特徴量にマッピングして，連続表現から離散表現に変換した。7.2 節で述べたように，入力された模倣音声を離散表現に変換して利用することで，入力音声が劣化していた場合でも，その影響を低減することが期待できる。本章では、クラスタリングモデルの学習に ESC-50 デー

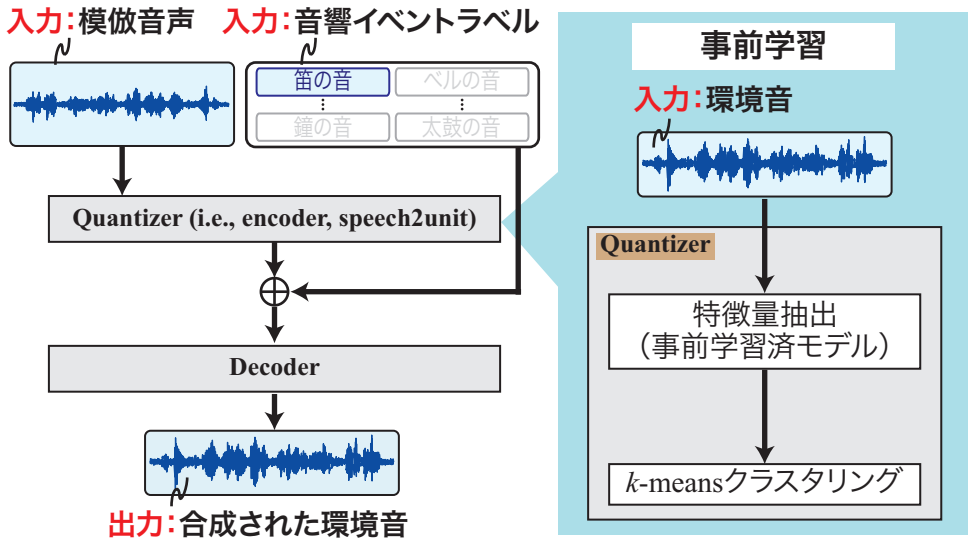


図 7.1: 音響イベントラベルと模倣音声を入力とする環境音合成の概要

タセット [64] に含まれる環境音を用いた。Encoder によって抽出された模倣音声の離散表現 $\boldsymbol{V} \in \mathbb{R}^{D \times T}$ に対して，音響イベントラベル $\boldsymbol{c}$ で以下のように条件づけを行う。

$$\boldsymbol{V}' = \mathrm{Linear}\left(\boldsymbol{V}, [\underbrace{\boldsymbol{c}, \boldsymbol{c}, \ldots, \boldsymbol{c}}_{T}] \right) \in \mathbb{R}^{D \times T} \tag{7.2}$$

ここで，D と T はそれぞれ特徴量の次元数と時間フレームを表す。$\mathrm{Linear}(\cdot)$ は線形変換を表す。そして，音響イベントラベルによって条件付けされた特徴量 $\boldsymbol{V}'$ を decoder に入力する。本章では，Tacotron2 の decoder を使用した。Decoder は，encoder によって変換された離散表現からメルスペクトログラムを推定するようにモデル学習される。Decoder のモデル学習では，各時間ステップにおいて推定されたメルスペクトログラムと正解のメルスペクトログラムの平均二乗誤差を損失関数として用いる。推論時には，ニューラルボコーダーを用いてメルスペクトログラムから時間波形に変換する。

7.4 模倣音声データセットの構築

環境音を声によっても模倣した音声データセットはいくつか公開されている [54, 55]。しかしながら，従来公開されているデータセットは，収録されている音響イベントのクラス数や 1 クラスあたりの環境音の数が少ない。環境音合成に使用するためには，1 個の音響イベントクラスにつき数十サンプルの音データが必要となる。そのため，本章において，環境音合成に利用可能な模倣音声データセットの構築を行う。

表 7.1: 実験に使用した ESC-50 データセットの音響イベントとサンプル数

音響イベント名	音の数	音の説明
Airplane	40	飛行機の音
Brushing teeth	40	歯を磨く音
Can opening	40	缶を開ける音
Car horn	40	車のクラクション
Cat	40	猫の鳴き声
Chainsaw	40	チェーンソーの音
Church bells	40	教会の鐘の音
Clapping	40	拍手の音
Clock alarm	40	目覚まし時計の音
Clock tick	40	時計の針の音
Cow	40	牛の鳴き声
Crackling fire	40	火がパチパチする音
Dog	40	犬の鳴き声
Door wood creaks	40	木のドアが軋む音
Door wood knock	40	木のドアを叩く音
Engine	40	エンジン音
Fireworks	40	花火の音
Footsteps	40	足音
Frog	40	カエルの鳴き声
Glass breaking	40	ガラスが割れる音
Hand saw	40	ノコギリの音
Hen	40	めんどりの鳴き声
Keyboard typing	40	キーボードのタイピング音
Mouse click	40	マウスのクリック音
Pig	40	豚の鳴き声
Pouring water	40	水を注ぐ音
Rooster	40	おんどりの鳴き声
Sheep	40	羊の鳴き声
Siren	40	サイレンの音
Vacuum cleaner	40	掃除機の音
Water drops	40	水滴

本章では，ESC-50 に含まれる環境音に対して模倣音声を収録した。収録方法としては，まず被験者に環境音を 1 音ずつ提示する。そして，被験者は提示された環境音を模倣した音声を発話する。これらの作業を繰り返し行って，各環境音に対する模倣音声の収録を実施した。なお，発話者の環境音の聴取並びに音声の収録は，何度でも可能とした。

1 個の音ファイル内で複数の環境音が重なっている場合，どの音に着目して発話すれば良いか決めることができず，模倣音声の収録が困難である。そのため，ESC-50 に含まれる音響イベントクラスの中から，各音クリップ内での音同士の重なりが比較的少ない 31 種類の音響イベントクラスを選択して，それらの環境音に対して模倣音声を収録した。各音響イベントクラスには 40 音含まれており，合計

1,240 音（40 音 × 31 クラス）の環境音を用いて模倣音声を収録した。その結果，8 名（男女各 4 名）の発話者から合計 9,920 音の音声を収録した。

音声収録には，マイクロフォンに SHURE MX150B O-XLR，オーディオインターフェースに Roland Rubix24 を使用した。なお，各音声は標本化周波数 48 kHz，量子化ビット数 16 bit Linear PCM で保存した。合成に使用する際には，22.05 kHz にダウンサンプリングして使用した。

7.5 評価実験

合成音をメディアコンテンツなどに利用するためには，環境音として自然である必要がある。そのため，本章においては，提案手法によって合成された音の自然性に関する主観評価をする。また，模倣音声を入力とすることで合成音の音高とリズムが制御可能であるかについても主観・客観評価を実施する。7.5.2 節では，合成音の自然性を評価する。7.5.3 項，7.5.4 項では，合成音が入力となった音声の音高とリズムを適切に反映できているかを評価する。7.5.5 項，7.5.6 項では，入力音声の音高とリズムの変化に提案手法が追従しているかどうかを評価する。

7.5.1 実験条件

環境音と模倣音声のペアには，7.4 節にて構築したデータセットを使用した。モデル学習には，各音響イベントクラスにつき 35 音の環境音とそれに対応する 4 名分（男女各 2 名）の模倣音声データを使用した。評価には，各音響イベントクラスにつき学習に使用していない 5 音の環境音とそれに対応する，学習に使用してない話者 2 名分（男女各 1 名）の模倣音声データを使用した。

Encoder 部にて使用した BYOL-A は，デフォルトのパラメータ設定 [65] を使用した。Encoder 部の k-means クラスタリングモデルは，クラスタ数 200，イテレーション数 100 でモデル学習した。なお，本章におけるクラスタ数は，encoder 部にてベクトル量子化を用いた音声合成・強調タスクの従来研究 [66] を参考に設定した。提案手法の入力に使用する音響イベントラベルは，31 次元の one-hot 表現を使用した。その他の提案手法に使用した音声，環境音データの条件並びに，モデルパラメータは表 7.2 に示す。また，波形復元のためのニューラルボコーダーには，HiFi-GAN [67] を使用した。なお，HiFi-GAN は，UrbanSound8K [68] データセットによって事前学習済みのモデル [69] を使用した。

各主観評価実験では，クラウドソーシングサービスを利用して，1 音あたり 10 名の評価者によって評価した。各評価では，音響イベントラベルのみから合成す

表 7.2: 環境音を模倣した音声からの環境音合成の実験条件

環境音/模倣音声	
音の長さ	5 s
サンプリング周波数	22.05 kHz
音波形の圧縮形式	16-bit linear PCM
音響特徴量	メルスペクトログラム（80 次元）
フレーム長	1,024 サンプル
フレームシフト	256 サンプル
Decoder のパラメータ	
LSTM 層の数	2
LSTM 層のユニット数	1024
Pre-net の層数	2
Pre-net の隠れ層の次元数	256
モデル学習のパラメータ	
バッチサイズ	64
学習率	0.0001
最適化手法	RAdam

る手法 (“Label”)，本章での提案手法 (“Label+vocal”)，正解のメルスペクトログラムをニューラルボコーダーによって再合成した音 (“Reconstructed”) の 3 種類の手法に対して評価を実施した。なお，“Label” には DCASE 2023 Challenge Task7 のベースライン手法 [70] を使用した。

7.5.2 合成音の自然性の評価

合成された環境音の自然性を評価するため，5 段階評価を実施した。評価者は，環境音と音響イベントラベルを同時に提示され，提示された音の自然性を 1（非常に不自然）〜5（非常に自然）の 5 段階で評価した。

表 7.3 に各音響イベントクラスごとの平均スコアと標準偏差を示す。表より，模倣音声と音響イベントラベルを使用する提案手法は，音響イベントラベルのみを入力する場合と比べて，全体的な音の自然性がわずかに劣ることが確認された。

図 7.2 (a) に “Label+vocal” と “Lable” の評価スコアの分布を示す。統計的有意差検定の結果より，31 種類の音響イベントのうち，5 種類の音響イベントにおいて “Label+vocal” が “Label” よりも高い自然性を獲得できたことがわかった。31 種類のうち 11 種類の音響イベントにおいては，統計的有意差が確認されなかった。よって，半数近くの音響イベントにおいては，提案手法は従来手法と同程度の自然性であると言える。また，提案手法によって合成された “fireworks” や “door

表 7.3: 合成音の自然性の評価結果

Sound event	airplane	brushing teeth	can opening	car horn	cat
Reconstructed	3.22 ± 1.09	4.16 ± 1.02	3.88 ± 1.04	3.50 ± 1.22	3.84 ± 1.09
Label	3.50 ± 1.22	2.40 ± 1.6	3.00 ± 1.54	2.04 ± 1.37	2.36 ± 1.38
Label + vocal	2.06 ± 0.98	2.00 ± 1.01	3.20 ± 1.34	1.76 ± 0.96	2.56 ± 1.03

Sound event	chainsaw	church bells	clapping	clock alarm	clock tick
Reconstructed	4.38 ± 0.83	2.94 ± 1.35	3.52 ± 1.16	3.84 ± 1.00	4.08 ± 0.83
Label	3.24 ± 1.19	1.82 ± 1.02	2.06 ± 1.35	1.20 ± 0.49	2.24 ± 1.57
Label + vocal	2.70 ± 1.09	1.66 ± 1.02	1.40 ± 0.78	2.74 ± 1.24	2.30 ± 1.30

Sound event	cow	cracking fire	dog	door wood creaks	door wood knock
Reconstructed	3.96 ± 0.92	3.18 ± 1.40	4.48 ± 0.79	3.84 ± 1.09	3.90 ± 0.97
Label	3.52 ± 1.13	2.88 ± 1.12	3.40 ± 1.75	2.04 ± 1.05	3.42 ± 1.62
Label + vocal	2.34 ± 1.14	2.20 ± 0.83	3.54 ± 1.07	1.62 ± 0.90	3.92 ± 1.03

Sound event	engine	fireworks	footsteps	frog	glass breaking
Reconstructed	3.56 ± 1.07	3.64 ± 1.05	3.64 ± 1.19	3.78 ± 1.20	2.94 ± 1.10
Label	3.58 ± 0.99	2.90 ± 1.42	2.04 ± 0.92	1.86 ± 1.01	1.90 ± 0.97
Label + vocal	2.52 ± 1.05	3.66 ± 1.14	1.64 ± 0.98	2.36 ± 1.17	1.56 ± 0.81

Sound event	hand saw	hen	keyboard typing	mouse click	pig
Reconstructed	4.38 ± 0.83	4.06 ± 1.00	3.84 ± 1.15	3.36 ± 1.32	4.10 ± 1.18
Label	2.80 ± 1.34	2.48 ± 0.97	1.86 ± 1.40	1.94 ± 1.36	3.02 ± 1.12
Label + vocal	2.14 ± 0.99	3.04 ± 1.19	1.72 ± 0.78	2.64 ± 1.34	2.28 ± 1.18

Sound event	pouring water	rooster	sheep	siren	vacuum cleaner
Reconstructed	4.08 ± 0.80	3.96 ± 1.21	3.92 ± 0.97	3.90 ± 1.09	3.26 ± 1.19
Label	2.28 ± 1.26	3.28 ± 1.47	2.70 ± 1.50	2.40 ± 1.50	2.94 ± 1.06
Label + vocal	1.52 ± 0.79	3.12 ± 1.02	2.20 ± 1.16	2.62 ± 1.12	1.84 ± 0.87

Sound event	water drops	average all events
Reconstructed	4.10 ± 0.95	3.80 ± 1.07
Label	2.12 ± 1.04	2.52 ± 1.25
Label + vocal	2.12 ± 1.00	2.36 ± 1.04

wood knock" の音などは，"Reconstructed" と同程度の自然性を獲得している。一方，"Label" によって合成された音は自然性が低下していることが確認できる。これらの結果より，提案手法は，時間的にスパースな音に対して比較的高い自然性を持つ音を合成可能であることが確認された。

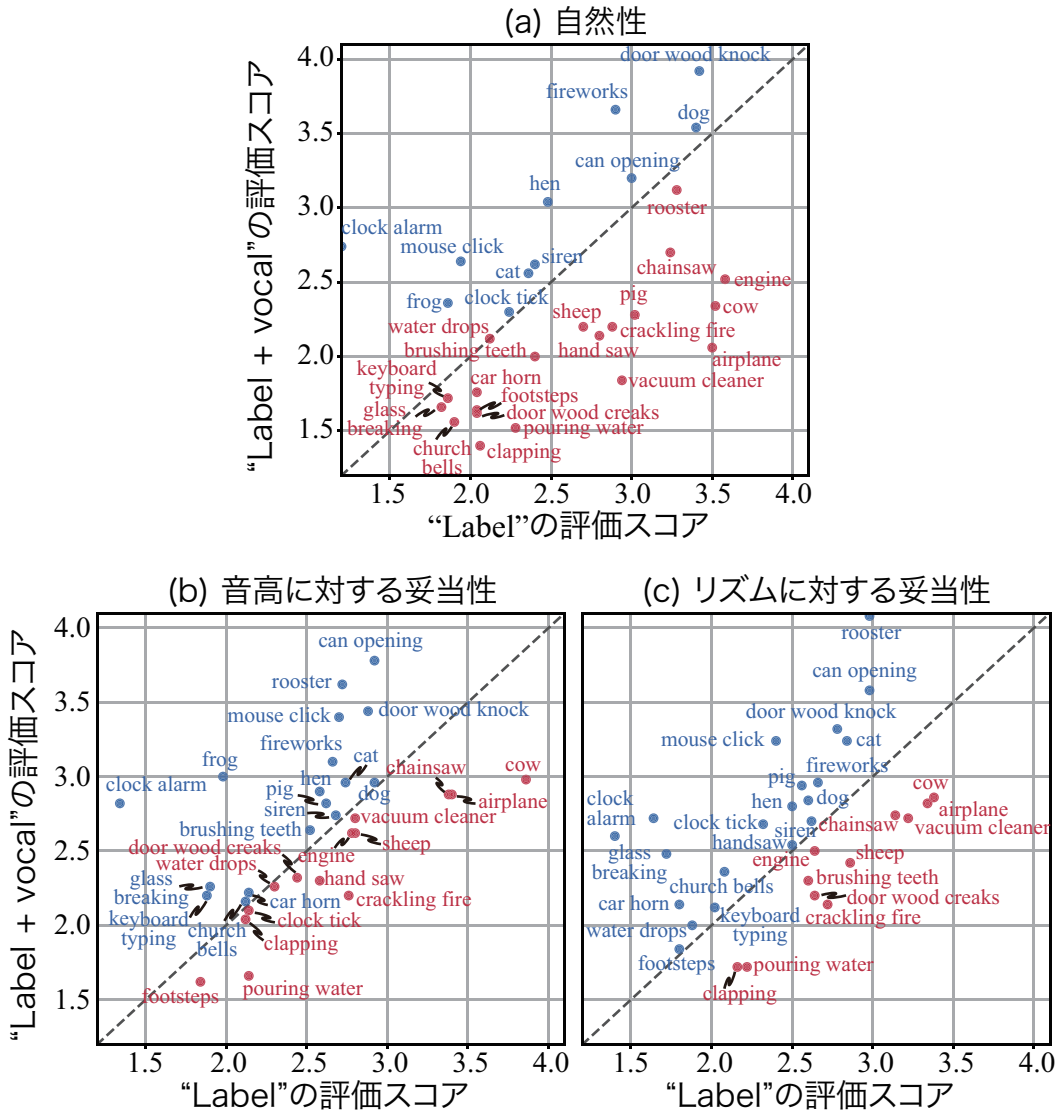


図 7.2: “Label+vocal” と “Label” の評価スコアの分布

7.5.3 入力音声の音高に対する合成音の妥当性の評価

入力された模倣音声の音高を反映した音として，合成音が妥当であるかどうか主観評価実験を実施した。評価者には音響イベントラベル，模倣音声，合成音を提示して評価を実施した。評価者は，提示された模倣音声の音高に対して合成音がどの程度妥当であるかどうかを 1（非常に妥当でない）〜5（非常に妥当である）の 5 段階で評価した。

表 7.4 に各音響イベントクラスにおける模倣音声の音高に対する合成音の妥当性の平均スコアと標準偏差を示す。全体の平均スコアに着目すると，“Label” と “Label+vocal” の間には統計的な有意な差が確認された。よって，提案手法は，音響イベントラベルのみを入力とする従来手法と比較して，入力となった模倣音声

表 7.4: 入力音声の音高に対する合成音の妥当性の評価結果

Sound event	airplane	brushing teeth	can opening	car horn	cat
Reconstructed	3.44 ± 1.20	4.10 ± 0.71	3.82 ± 1.00	3.98 ± 0.96	4.14 ± 0.83
Label	3.40 ± 0.97	2.52 ± 1.39	2.92 ± 1.19	2.14 ± 1.18	2.74 ± 1.44
Label + vocal	2.88 ± 1.14	2.64 ± 1.35	3.78 ± 1.04	2.22 ± 1.11	2.96 ± 0.73

Sound event	chainsaw	church bells	clapping	clock alarm	clock tick
Reconstructed	3.70 ± 1.20	3.62 ± 1.05	3.32 ± 1.20	4.24 ± 0.94	3.82 ± 1.14
Label	3.38 ± 0.99	2.12 ± 1.32	2.12 ± 1.04	1.34 ± 0.69	2.30 ± 1.13
Label + vocal	2.88 ± 1.06	2.16 ± 1.04	2.04 ± 1.05	2.82 ± 1.14	2.26 ± 0.96

Sound event	cow	cracking fire	dog	door wood creaks	door wood knock
Reconstructed	4.22 ± 0.86	3.50 ± 1.22	3.88 ± 1.06	3.94 ± 0.96	3.90 ± 1.05
Label	3.86 ± 0.83	2.76 ± 0.98	2.92 ± 1.37	2.44 ± 1.13	2.88 ± 1.19
Label + vocal	2.98 ± 0.96	2.20 ± 0.99	2.96 ± 0.99	2.32 ± 1.15	3.44 ± 0.91

Sound event	engine	fireworks	footsteps	frog	glass breaking
Reconstructed	3.52 ± 0.93	3.50 ± 1.20	3.84 ± 1.08	4.06 ± 0.89	3.04 ± 1.35
Label	2.78 ± 1.15	2.66 ± 1.21	1.84 ± 0.93	1.98 ± 1.13	1.90 ± 0.93
Label + vocal	2.62 ± 0.95	3.10 ± 0.97	1.62 ± 0.75	3.00 ± 1.09	2.26 ± 1.17

Sound event	hand saw	hen	keyboard typing	mouse click	pig
Reconstructed	3.82 ± 1.02	3.96 ± 0.95	3.90 ± 0.99	3.88 ± 1.12	3.76 ± 1.02
Label	2.58 ± 1.30	2.58 ± 0.97	1.88 ± 1.17	2.70 ± 1.56	2.62 ± 1.12
Label + vocal	2.30 ± 0.97	2.90 ± 0.97	2.20 ± 1.09	3.40 ± 1.16	2.82 ± 0.80

Sound event	pouring water	rooster	sheep	siren	vacuum cleaner
Reconstructed	3.30 ± 1.09	3.86 ± 0.83	4.10 ± 0.93	3.94 ± 1.19	3.72 ± 0.95
Label	2.14 ± 1.09	2.72 ± 1.33	2.80 ± 1.34	2.68 ± 1.15	2.80 ± 1.18
Label + vocal	1.66 ± 0.82	3.62 ± 1.14	2.62 ± 1.29	2.74 ± 1.26	2.72 ± 1.14

Sound event	water drops	average all events
Reconstructed	4.14 ± 0.93	3.81 ± 1.03
Label	2.14 ± 1.25	2.54 ± 1.15
Label + vocal	2.10 ± 1.18	2.65 ± 1.04

の音高を表現した音を合成可能であることがわかる。

図 7.2 (b) に “Label+vocal” と “Lable” の評価スコアの分布を示す。統計的有意差検定の結果から，31 種類の音響イベントのうち，6 種類において提案手法は従来手法よりも高い評価スコアを獲得した。特に，“rooster” や “frog” のような動物

の鳴き声や，“mouse click” のような時間的にスパースな音に関して，提案手法は従来手法と比較して有意な差を示した。

7.5.4　入力音声のリズムに対する合成音の妥当性の評価

入力された模倣音声のリズムを反映した音として，合成音が妥当であるかどうか主観評価実験を実施した。評価者には音響イベントラベル，模倣音声，合成音を提示して評価を実施した。評価者は，提示された模倣音声の音高に対して合成音がどの程度妥当であるかどうかを 1（非常に妥当でない）〜5（非常に妥当である）の 5 段階で評価した。

表 7.5 に各音響イベントクラスにおける模倣音声のリズムに対する合成音の妥当性の平均スコアと標準偏差を示す。全体の平均スコアに着目すると，“Label” と “Label+vocal” の間には統計的な有意な差が確認された。よって，提案手法は，音響イベントラベルのみを入力とする従来手法と比較して，入力となった模倣音声のリズムを表現した音を合成可能であることがわかる。

図 7.2 (c) に “Label+vocal” と “Lable” の評価スコアの分布を示す。統計的有意差検定の結果から，31 種類の音響イベントのうち，7 種類において提案手法は従来手法よりも高い評価スコアを獲得した。特に，動物の鳴き声や時間的にスパースな音に関して，提案手法は従来手法と比較して有意な差を示した。また、7.5.3 節と 7.5.4 節の評価では、“engine” のような断続的に鳴り続ける音のスコアが低い傾向にあった。今後、“engine” のような合成がうまくいかなかった音の分析をさらに進める必要がある。

7.5.5　入力音声の音高変化による合成音の評価

入力する模倣音声の音高を変化させた場合，それに追従して合成音が変化するかどうかについて客観・主観評価を実施した。入力とする模倣音声の音高は，Python のパッケージである *librosa* の音高をシフトさせる関数[1]を用いて，元となる音声を基準に上限に半オクターブずつ変化させた。客観評価としては，音高シフトを行わない元の模倣音声を 1 とした場合の音の相対的なスペクトル重心の変化を評価した。主観評価は，7.5.2 項，7.5.3 項，7.5.4 項と同様の評価を実施した。

図 7.3 に入力音声の音高を変化させた場合の合成音の客観/主観評価の結果をそれぞれ示す。図 7.3 (a) の結果より，音高が上がるとそれに追従してスペクトル重

[1] `https://librosa.org/doc/main/generated/librosa.effects.pitch_shift.html`

表 7.5: 入力音声のリズムに対する合成音の妥当性の評価結果

Sound event	airplane	brushing teeth	can opening	car horn	cat
Reconstructed	3.54 ± 1.01	3.78 ± 0.84	4.10 ± 0.86	4.12 ± 0.85	4.32 ± 0.91
Label	3.34 ± 1.15	2.60 ± 1.40	2.98 ± 1.19	1.80 ± 0.86	2.84 ± 1.30
Label + vocal	2.82 ± 1.06	2.30 ± 1.22	3.58 ± 1.11	2.14 ± 1.20	3.24 ± 1.00

Sound event	chainsaw	church bells	clapping	clock alarm	clock tick
Reconstructed	3.82 ± 1.16	3.78 ± 1.17	3.40 ± 1.21	4.26 ± 0.92	4.10 ± 0.91
Label	3.14 ± 0.99	2.08 ± 1.05	2.16 ± 1.22	1.40 ± 0.70	2.32 ± 1.32
Label + vocal	2.74 ± 1.08	2.36 ± 1.01	1.72 ± 0.97	2.60 ± 1.21	2.68 ± 1.13

Sound event	cow	cracking fire	dog	door wood creaks	door wood knock
Reconstructed	4.28 ± 0.88	3.26 ± 1.07	4.08 ± 0.99	3.94 ± 0.98	4.08 ± 0.99
Label	3.38 ± 1.16	2.72 ± 1.16	2.60 ± 1.14	2.64 ± 1.17	2.78 ± 1.33
Label + vocal	2.86 ± 1.20	2.14 ± 1.14	2.84 ± 0.93	2.20 ± 1.28	3.32 ± 1.19

Sound event	engine	fireworks	footsteps	frog	glass breaking
Reconstructed	3.50 ± 1.20	3.68 ± 0.79	3.52 ± 1.13	4.16 ± 0.98	3.72 ± 1.47
Label	2.64 ± 1.21	2.66 ± 0.94	1.80 ± 0.90	1.64 ± 1.01	1.72 ± 1.01
Label + vocal	2.50 ± 1.20	2.96 ± 1.01	1.84 ± 0.93	2.72 ± 1.20	2.48 ± 1.16

Sound event	hand saw	hen	keyboard typing	mouse click	pig
Reconstructed	4.04 ± 1.03	4.22 ± 0.93	3.70 ± 0.86	3.74 ± 1.12	3.50 ± 1.09
Label	2.50 ± 1.25	2.50 ± 0.99	2.02 ± 1.20	2.40 ± 1.46	2.56 ± 1.03
Label + vocal	2.54 ± 1.01	2.80 ± 0.97	2.12 ± 1.00	3.24 ± 0.98	2.94 ± 1.10

Sound event	pouring water	rooster	sheep	siren	vacuum cleaner
Reconstructed	3.50 ± 1.20	4.02 ± 1.06	3.94 ± 0.87	4.08 ± 0.92	3.84 ± 0.89
Label	2.22 ± 1.28	2.98 ± 1.50	2.86 ± 1.28	2.62 ± 1.07	3.22 ± 1.13
Label + vocal	1.72 ± 0.86	4.08 ± 0.85	2.42 ± 1.31	2.70 ± 1.18	2.72 ± 1.25

Sound event	water drops	average all events
Reconstructed	4.02 ± 1.08	3.87 ± 1.01
Label	1.88 ± 0.94	2.48 ± 1.14
Label + vocal	2.00 ± 1.11	2.62 ± 1.09

心の値も大きくなることがわかる。一方，音高が下がった場合のスペクトル重心はわずかに値が小さくなっているが，大きな変化は確認されなかった。また，図 7.3 (b)–(c) の主観評価結果より，入力音声の音高を変化させるとわずかに各評価スコアが低下することが確認された。

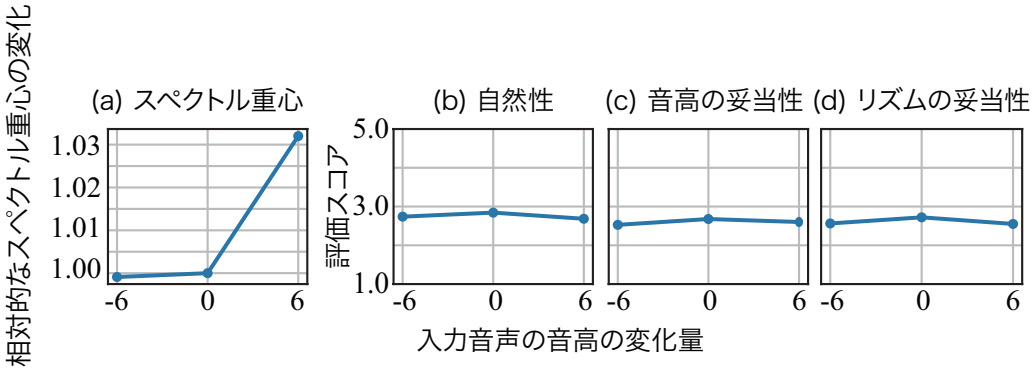


図 7.3: 入力音声の音高を変化させた場合の合成音の主観/客観評価結果

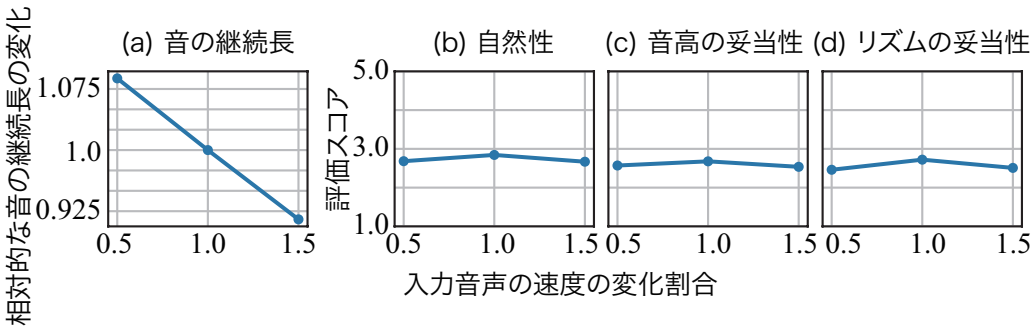


図 7.4: 入力音声のリズムを変化させた場合の合成音の主観/客観評価結果

7.5.6 入力音声のリズム変化による合成音の評価

入力する模倣音声のリズムを変化させた場合，それに追従して合成音が変化するかどうかについて客観・主観評価を実施した。入力とする模倣音声のリズムは，*librosa* の音を時間伸縮にさせる関数[2]を用いて，元となる音声を基準に 0.5 倍速，1.5 倍速の 2 段階で変化させた。客観評価としては，入力音声のリズムを変化させない場合の音の継続長を 1 とした場合の相対的な音の継続長の変化を評価した。主観評価は，7.5.2 項，7.5.3 項，7.5.4 項と同様の評価を実施した。

図 7.4 に入力音声のリズムを変化させた場合の合成音の客観/主観評価の結果をそれぞれ示す。図 7.4 (a) の結果より，入力音声の伸縮率が大きくなるにつれて合成音の継続長も長くなることがわかる。一方，図 7.3 (b)–(c) の主観評価結果より，入力音声のリズムを変化させるとわずかに各評価スコアが低下することが確認された。7.5.5 節の主観評価において各評価スコアが下がった結果と合わせて，今後原因を調べる必要がある。

[2] https://librosa.org/doc/main/generated/librosa.effects.time_stretch.html

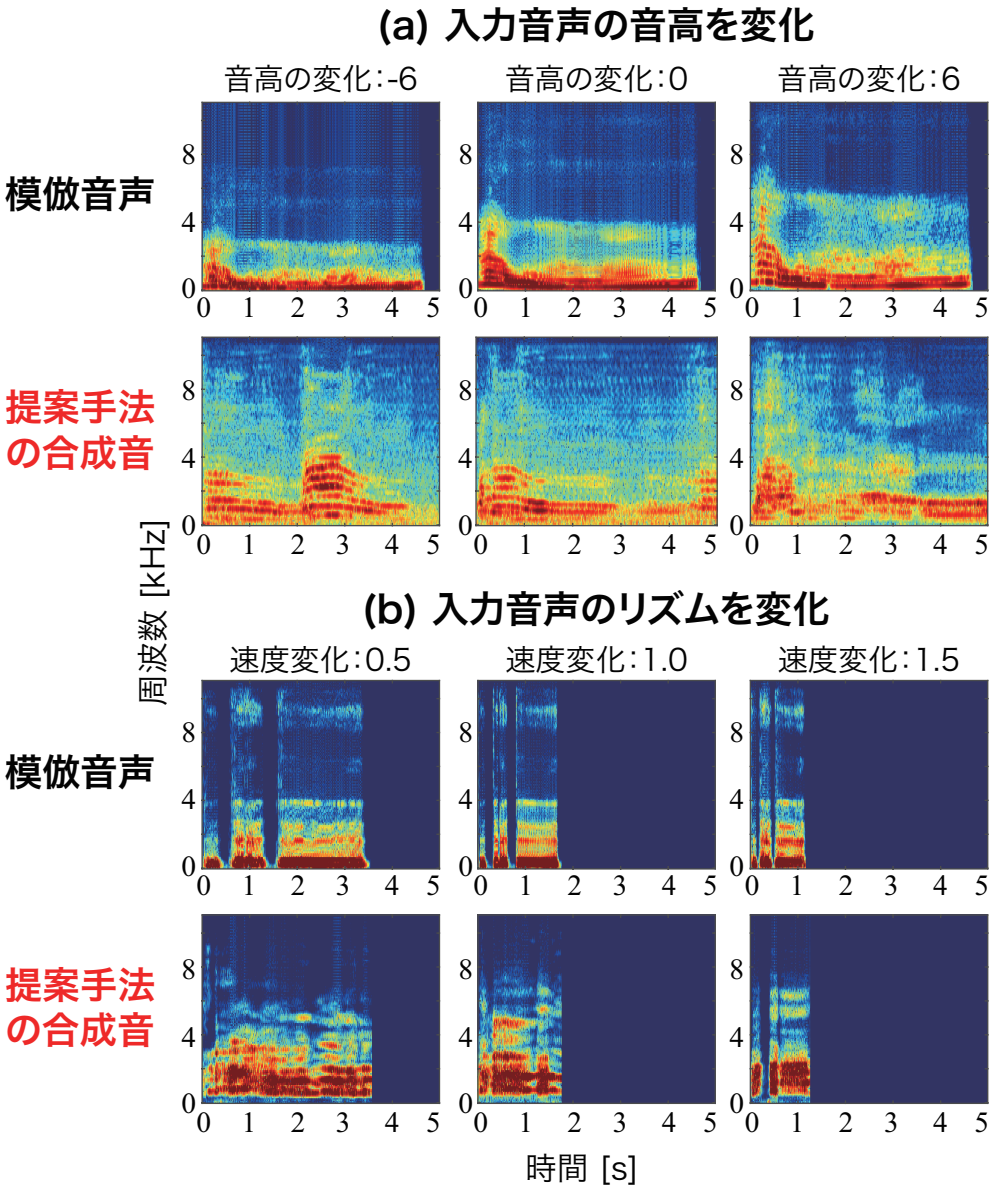


図 7.5: 入力音声の音高とリズムを変化させた場合の合成音のスペクトログラム

7.5.7　スペクトログラムによる合成音の変化の確認

図 7.5 に音高とリズムを変化させた模倣音声と提案手法による合成音のスペクトログラムを示す。図より，提案手法は，入力する模倣音声の音高とリズムを変化させた場合においても，各特徴に対応した環境音を合成可能であることがわかる。このように，合成音の音高とリズムを制御するためには，環境音合成モデルへの入力情報として音響イベントラベルに加えて模倣音声も利用することが有効であるとわかった。なお，各合成手法による合成音はウェブページにて聴取可能である [71]。

7.6　7章のまとめ

本章では，模倣音声を利用した環境音合成手法を提案した。また，環境音を模倣した音声データの収録も行い，データセット構築を行った。評価実験の結果より，音響イベントラベルに加えて模倣音声を利用することで，合成音の音の高さとリズムを制御可能であることが明らかになった。また，合成音の自然性の評価においては，半数近くの音響イベントにおいて，提案手法は従来手法と同程度の自然性を獲得した。一方，natural sound と比較すると提案手法の合成音は自然性にて劣るという結果となった。よって，さらに高品質な環境音を合成可能な手法を検討する必要がある。今後，模倣音声だけでなく画像や音の説明文なども同時に利用した環境音合成手法に取り組む。

第8章　結論

本論文では，個々の音である音響イベントに着目し，多様な環境音を生成するための環境音合成手法を提案した。本研究の学術的貢献は，これまで検討例が少ない環境音，特に個々の音を高い精度かつ柔軟に制御するための新たな合成手法を提案したことである。

第3章では，オノマトペからの環境音合成を実現するためのデータセット構築を行った。クラウドソーシングにてRWCP-SSD内の環境音に対しオノマトペの付与を実施して，合計155,568個のオノマトペ収集した。また，環境音を表現するのに適したオノマトペを選び出すために，オノマトペに対して付与した自信度と他者から付与された許容度を用いることが効果的であることも示した。

第4章では，合成された環境音をどのように評価するべきかについて検討した。環境音合成モデルの入力に使用された情報に対する合成音の妥当性を評価する主観評価手法を提案して，客観評価の結果と比較を行った。比較実験の結果，主観評価と客観評価の間では異なる傾向が見られることが確認された。よって，環境音合成の評価においては，客観評価のみならず，主観評価も必要であることが明らかとなった。

第5章では，音の種類を表す音響イベントラベルを入力とする環境音合成手法を提案した。環境音の自然性の評価においては，コーヒーミルで豆を挽く音，目覚まし時計の音，マラカスの音の合成音では自然音と同程度の自然性を得られた。一方，自然音と合成音を識別する主観評価実験においては，82.71%の比較対で自然音を識別できるという結果となった。これらの結果より，提案手法によって，個々の音を表す音波形を合成可能であることを確認し，主観評価実験にてデータセット内に存在する自然音と同程度の自然性を得る環境音が合成可能であることを示した。

第6章では，第4章にて構築したデータセットを用いて，オノマトペからの環境音合成手法を提案した。主観評価実験にて，波形接続のような仕組みでオノマトペから音を合成する従来手法であるKanaWaveよりも，提案手法がオノマトペをより表現できているという結果を示した。また，オノマトペと音響イベントラベル両方を入力とする手法の提案も行い，オノマトペだけでなく，音響イベントの制御も可能であることを示した。

第 7 章では，模倣音声を利用した環境音合成手法を提案した。評価実験の結果より，音響イベントラベルに加えて模倣音声を利用することで，合成音の音の高さとリズムを制御可能であることが明らかになった。また，合成音の自然性の評価においては，半数近くの音響イベントにおいて，提案手法は従来手法と同程度の自然性を獲得した。一方，natural sound と比較すると提案手法の合成音は自然性にて劣るという結果となった。

本研究では，音響イベントラベル，オノマトペ，模倣音声をそれぞれを入力とし，環境音を合成することは実現したが，合成音の品質に関しては自然音と劣ってしまう傾向にあった。さらなる品質向上に向け，比較的高い品質で合成できる音と上手く合成できない音の特徴を詳細に分析する必要がある。特に，模倣音声を利用した環境音合成ではその他の手法と比較して大きく品質が劣化した傾向にあった。そのため，今後，データ数の増強やモデルの改良など，さらなる高品質な環境音合成の実現への取り組みが必要である。

付 録A　Transformerを用いたオノマトペからの環境音合成

A.1　はじめに

環境音合成手法の 1 つとして，6 章にてオノマトペを用いた手法を提案した。6 章にて提案した手法では，seq2seq を用いてオノマトペからの環境音合成を実現している。seq2seq は RNN [72] などの再帰構造を利用しているため，短いオノマトペから比較的単調な音を合成するケースに優れている。一方，長いオノマトペから長期的な構造を持つ音を合成するケースに対しては，音の合成時に必要な特徴を十分に捕らえられない課題がある。

本章では，Transformer を用いたオノマトペからの環境音合成手法を提案する。Transformera は attention により長期的な特徴を捉えることが可能であり，系列長が長い場合においても，機械翻訳 [73] や音声合成 [74] などといった系列間変換を目的とする研究において高い変換性能を示している。そのため，従来の seq2seq を用いた手法では表現困難であった大局的な時間構造を持つ環境音に対して頑健な手法の実現が期待できる。

A.2　Transfomerを用いたオノマトペからの環境音合成手法

Fig. A.1 に本研究にて提案する Transformer を用いたオノマトペからの環境音合成手法のモデル学習の流れを示す。提案手法は, 6 章にて提案した手法と同じく, オノマトペから特徴ベクトルを抽出する encoder と特徴ベクトルに基づき音響特徴量を推定する decoder により構成されている。6 章では, encoder, decoder 共に LSTM を使用した。本章においては，encoder，decoder それぞれに対して Transformer を使用する。まず，オノマトペを音素系列に変換し 3 層の CNN により構成された encoder pre-net にて埋め込みベクトルを獲得する。そして得られた埋め込みベクトルを encoder へと入力する。Encoder 部の multi-head attention 部において入力

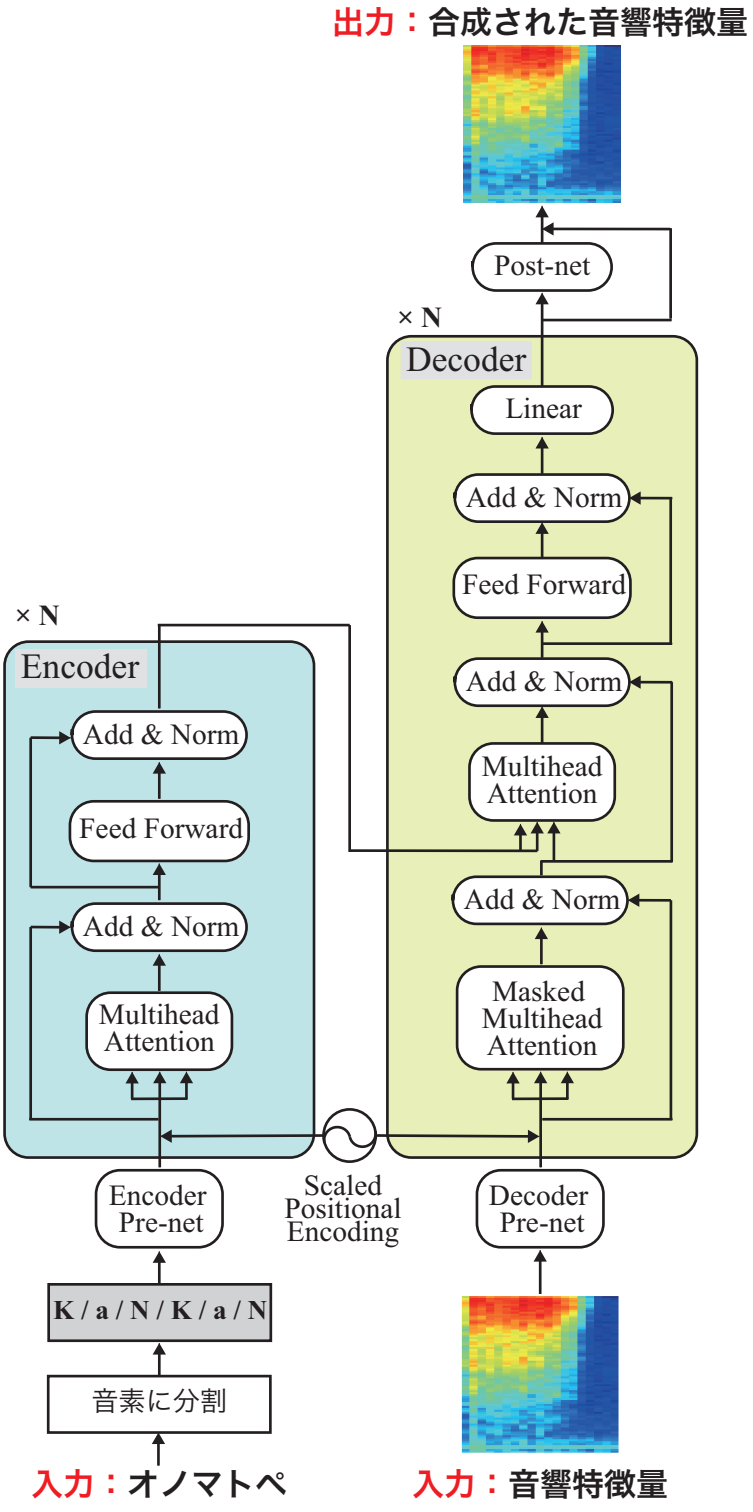


図 A.1: Transformer によるオノマトペからの環境音合成の概要

オノマトペ内における大局的な照応関係を学習することが可能となる。Decoder 部では，音データから抽出された音響特徴量を 2 層の全結合層によって構成された

表 A.1: Transformer を用いたオノマトペからの環境音合成の実験条件

音の長さ	1–2 s
サンプリング周波数	16,000 Hz
音波形の圧縮形式	16-bit linear PCM
Encoder 層の数	3
Encoder 層の数	3
Multi-head の数	4
Batch size	32
音響イベントラベルの次元数	10
最適化手法	RAdam
音響特徴量	メルスペクトログラム（80 次元）
フレーム長	0.128 s（2,048 サンプル）
フレームシフト	0.032 s（512 サンプル）

decoder pre-net へと入力する。Decoder pre-net は encoder pre-net における音素の埋め込み空間と同じ部分空間に音響特徴量を射影する役割を担っている。これにより，decoder 部の multi-head attention 機構にて音素と音響特徴量間の類似性を計算することが可能となる。そして，decoder 部において推定された音響特徴量と教師データの各時刻に対応する音響特徴量の L1 ノルムを誤差関数とすることでモデルを学習する。推論時には，推定した音響特徴量に対して位相復元のアルゴリズムを適用し環境音の波形へ変換する。本章においては，音響特徴量としてメルスペクトログラムを使用したため，3 層の全結合層から構成されたネットワークを介し振幅スペクトログラムに変換し，6 章と同様 Griffin-Lim 法により波形へと変換する。

A.3 評価実験

A.3.1 実験条件

本実験では，環境音データに RWCP-SSD の中から，表 4.3 に示す 10 種類の音響イベントを各 100 音ずつ，計 1,000 音用いた。オノマトペのデータには，本研究にて構築した RWCP-SSD の音データに対してオノマトペを付与したデータセットを使用した。モデル学習には，1 音あたり 15 個，計 14,250 個（15 個 × 950 サンプル）のオノマトペを使用した。Transformer を用いた提案手法のモデルパラメータは Table A.1 に示す。

評価実験はクラウドソーシングサービスにより実施した。実験 I-1，I-2 では 10（音響イベント数）× 10（オノマトペ数）× 30（人）= 3,000 サンプル，実験 II-1，

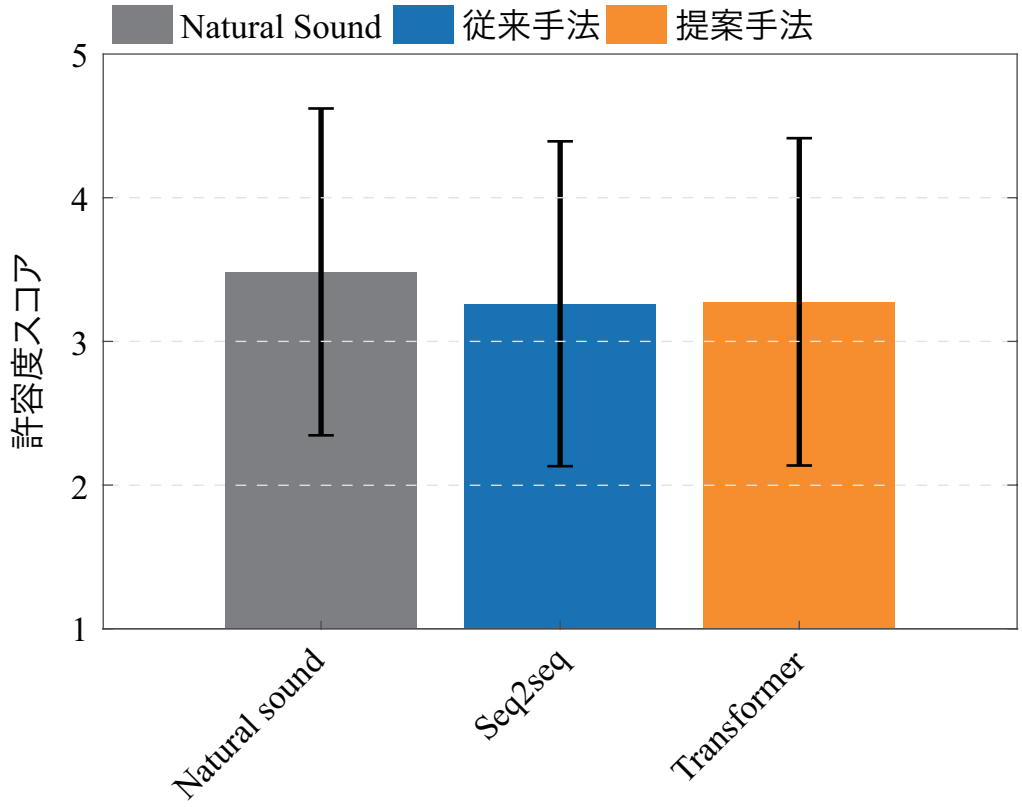


図 A.2: オノマトペに対する環境音の許容度の評価結果

II-2 では 10（音響イベント数）× 5（合成音数）× は 30（人）= 1,500 サンプルの合成音を用いた。比較として，実験 I-1，I-2 ではデータセットに含まれる自然音 (natural sound)，seq2seq を用いた従来手法による合成音に対しても同様の評価を実施した。実験 II-1，II-2 では実験 I-1，I-2 で比較した手法に加え，5.2 章で提案した WaveNet を用いた音響イベントラベルを入力として環境音を合成する手法による合成音に対しても同様に評価した。WaveNet の合成音の特徴として，入力に音響イベントラベルのみを使用しているため，学習時に使用した音に類似した音ばかりが合成される傾向にある。

A.3.2　オノマトペに対する環境音の評価

合成された環境音が入力となったオノマトペをどの程度表現できているか評価するため，音とオノマトペを提示し，以下の 2 つの評価を行った。

- **実験 I-1：オノマトペに対する環境音の許容度**
  提示されたオノマトペを表す音として提示された音がどの程度許容できるかについて 1（非常に許容できない）〜5（非常に許容できる）の 5 段階で評価。

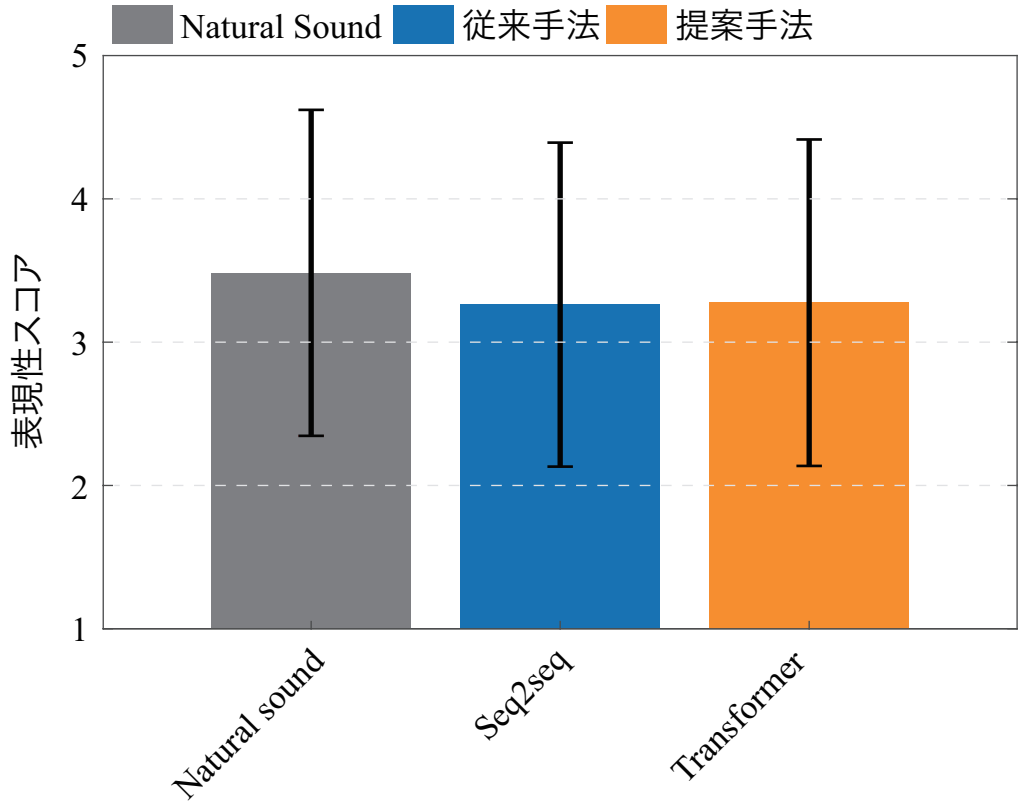


図 A.3: オノマトペに対する環境音の表現性の評価結果

- **実験 I-2：オノマトペに対する環境音の表現性**
  提示されたオノマトペを表す音として提示された音がどの程度適しているかについて 1（非常に表現できていない）〜5（非常に表現できている）の 5 段階で評価。

実験 I-1，I-2 のオノマトペに対する環境音の許容度，表現性に関する評価の平均スコアと標準偏差をそれぞれ Fig. A.2 および Fig. A.3 に示す。結果より，従来手法，提案手法による合成音は，許容度と表現性の 2 つの評価指標においてデータセットに含まれる自然音 (natural sounds) と近いスコアを獲得した。各手法による合成音の特徴を確認するため，Fig. A.4 に各手法において「ズザー (/ z u z a: /) というオノマトペを入力した場合の合成音のスペクトログラムを示す。図より，Transformer を用いた提案手法による合成音は，seq2seq を用いた従来手法よりも長期的な時間構造をうまく捉えていることが確認できる。このように，Transformer により大局的な特徴を捉えることで，長期的な構造を特徴とする環境音に対しても頑健な音の合成が可能となることを示した。

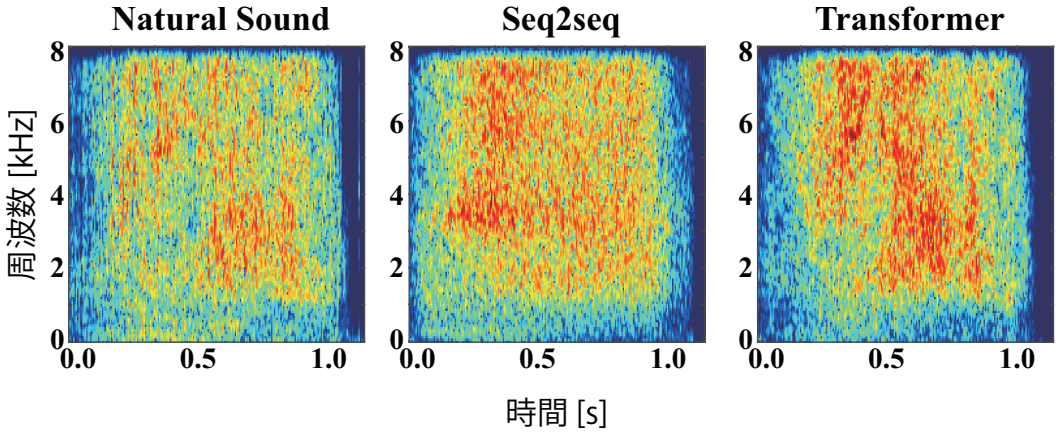


図 A.4: データセットに含まれる自然と各手法によるオノマトペからの合成音のスペクトログラム

A.3.3 環境音の品質に関する評価

環境音の品質を評価するため，データセットに含まれる自然 (natural sound) と合成音をランダムに被験者に提示し，以下の 2 つの評価を行った。

- **実験 II-1：環境音の全体的な印象**
  環境音の全体的な印象について，1（非常に悪い）〜5（非常に良い）の 5 段階で評価。

- **実験 II-2：環境音の自然性**
  環境音の自然性（実際に存在しそうな音であるか，音としての違和感がないか）について 1（非常に不自然である）〜5（非常に自然である）の 5 段階で評価。

実験 II-1，II-2 の環境音の全体的な印象，環境音の自然性に関する評価の平均スコアと標準偏差をそれぞれ Fig. A.5 および Fig. A.6 に示す。Fig. A.5 の結果より，提案手法は seq2seq を用いた従来手法と比較し，全体的な音の品質を損ねることなく合成可能であることがわかる。一方，natural sound や WaveNet による合成音と比較すると，音としての全体的な品質は劣るという結果になった。Fig. A.6 に示す自然性に関する評価においては，提案手法は WaveNet による従来手法と同程度の自然性を獲得する結果となった。これらの結果より，WaveNet を用いた従来手法には全体的な品質では劣るが，提案手法における合成音も従来手法と比較し環境音としては違和感のない品質になっていることが確認できる。

実験 II の環境音の全体的な印象の評価において，ひげ剃りの動作音や紙を引き裂く音の合成音はスコアが低くなる傾向が見られた。音のみを被験者に提示して

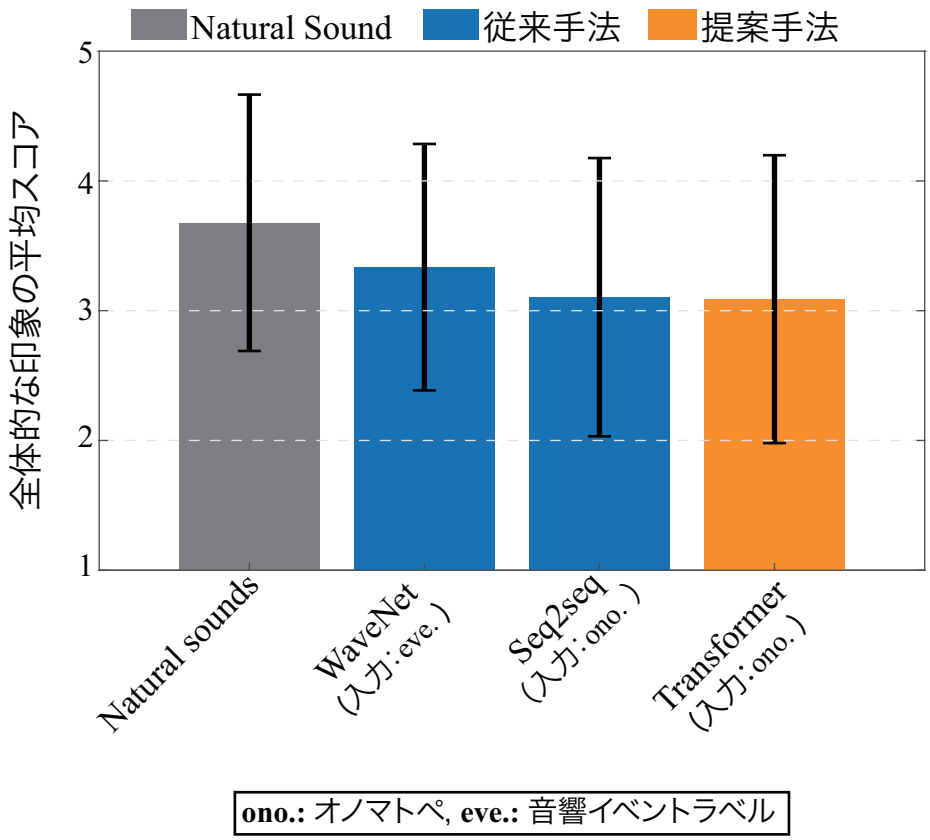


図 A.5: 環境音の全体的な印象に関する評価結果

評価したため，ひげ剃りの動作音のように音の継続長が長く，広い周波数帯域に特徴が現れている音はノイズのように被験者に捉えられてしまい，スコアが低くなったのではないかと考えられる。今後，音と同時にどの音響イベントを表現した音であるかを提示するなど，評価方法の改善が必要である。

A.4　付録 A のまとめ

本章では，Transformer を用いたオノマトペからの環境音合成手法を提案した。オノマトペに対する環境音の評価において，提案手法による合成音はデータセットに含まれる音と近いスコアを獲得した。seq2seq による合成手法と比較すると，提案手法は同程度のスコアを獲得する結果となった。一方，長期的な時間構造をもつような環境音に対して提案手法は，seq2seq と比べ頑健な音の合成を可能とした。環境音の品質に関する評価においては，自然性の項目で seq2seq を用いた合成手法，WaveNet による合成手法と同程度の品質を獲得した。今後，入力となったオノマトペを十分に表現できなかった合成音の特徴を詳細に分析する。

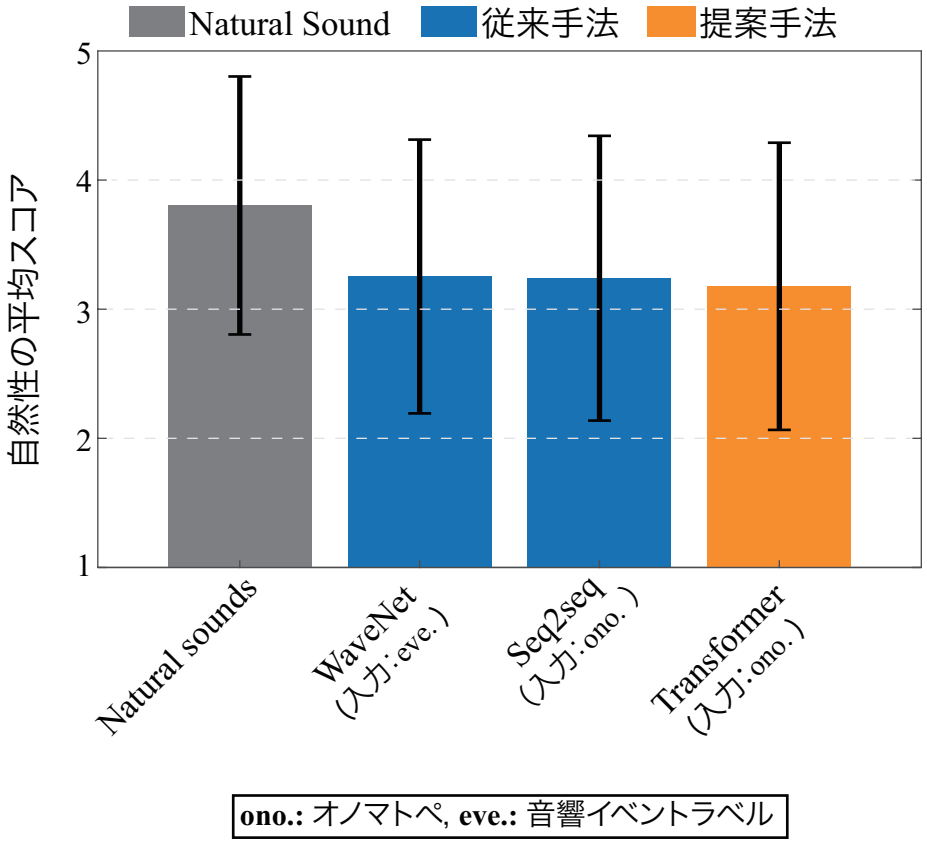


図 A.6: 環境音の自然性に関する評価結果

付 録B　オノマトペを用いた環境音抽出

B.1　はじめに

映画やゲームなどのメディアコンテンツにおいて，没入感や臨場感を高めるために環境音は欠かせない存在である。目的とする音を得るための方法として，環境音データベースから音を探して，利用する方法が考えられる。しかし，目的とする音がデータベース内にあるとは限らない。一方，インターネット上には音の種類（音響イベント）などのラベルが付与されていない環境音が大量に存在するが，音響イベントラベルを付与するには豊富なドメイン知識が必要なため，データベースを拡張することは容易ではない。また，データベースの規模が大きくなった場合でも，利用者がドメイン知識を持つ必要があるため，誰でも容易に使用できるとは限らない。それらを解消するため，音を直感的に検索できる手法も提案されている。例えば，音声 [25, 75, 76] やオノマトペ [77] を検索クエリとして利用した音の検索システムが存在する。これらの方法は，直感的に音を検索でき，利用者の満足度が高いことも報告されている [75]。そのため，直感的に目的の音を抽出できれば，コンテンツ制作の手助けになると考えられる。

本章では，音の様子を模倣した文字列であるオノマトペを用いた環境音抽出手法を提案する。オノマトペは音の長さ，高さ，音色など音の特徴を表現するのに有効であるとされている [23]。また，音に対するオノマトペの付与にはドメイン知識などを必要としないため，ラベル付けにかかるコストが低いという利点もある。そこで本章では，Fig. B.1 に示すように，抽出する音をオノマトペを用いて指定する。そして，音源分離や音抽出の研究 [78, 79, 80, 81] で用いられる U-Net [82] とオノマトペを用いて抽出対象の時間周波数マスクを推定し，オノマトペに対応する音のみを抽出する。

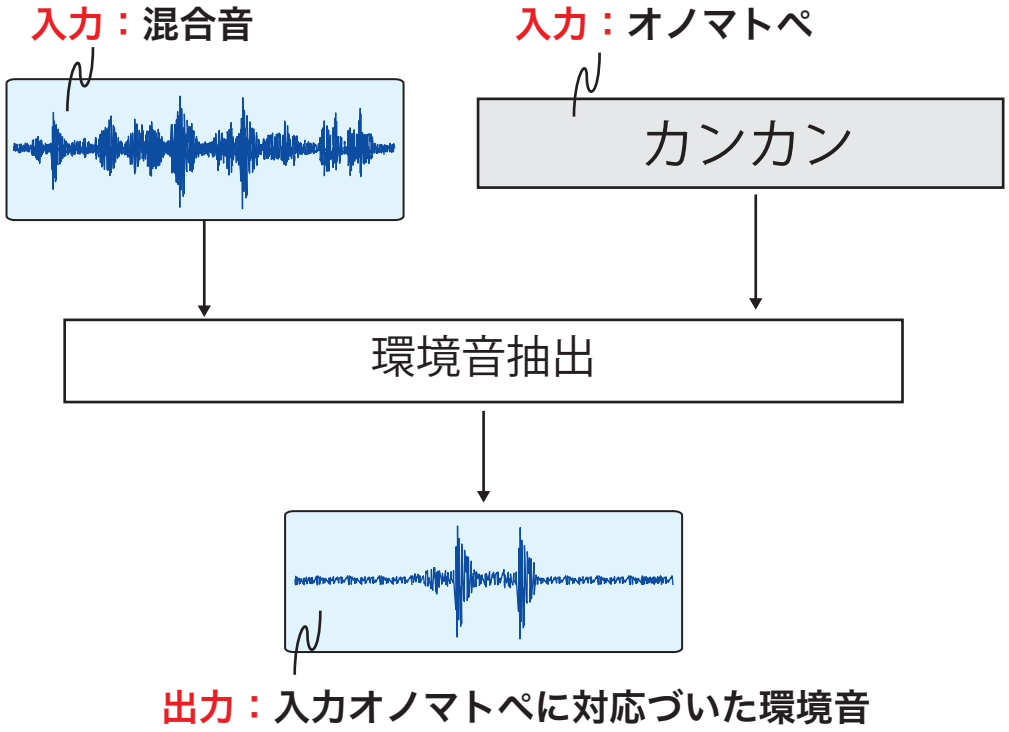


図 B.1: オノマトペを用いた環境音抽出の概要

B.2 関連研究

これまで，深層学習を用いた環境音の抽出・分離手法が提案されている [78, 79, 83, 84]。Sudo らは，U-Net に基づく環境音分離手法 [79] を提案している。また，楽器音の分離においても U-Net を用いた手法が提案されている。Ochiai らは，音声分離のために提案された Conv-TasNet を用いて，特定の音響イベントの音のみを抽出する手法 [83] を提案している。これらの手法 [78, 79, 83] は，抽出・分離する音を指定するために音響イベントのクラスを使用している。しかし，環境音には，音の長さ，音高，音色のように音の種類だけでは表現できない様々な特性がある。例えば，音高に関係なく「笛の音」という 1 つのクラスを定義すると，従来の手法 [78, 79, 83] では所望の音高の音のみを抽出することが困難である。これらの解決策として「甲高い笛の音」,「低い笛の音」のようにより細かな音響イベントクラスを定義することが挙げられる。しかし，ラベル付けにコストを要する。このように，より詳細な音響イベントクラスを定義できたとしても，クラス内の音の多様性は常に存在し，それらの違いを区別する方法は存在しない。そのため，音響イベントクラスによって条件付けする従来手法は，特定の音を抽出することに適していない。

目的とする音の特徴をハミングによって表現した歌声抽出の手法も提案されて

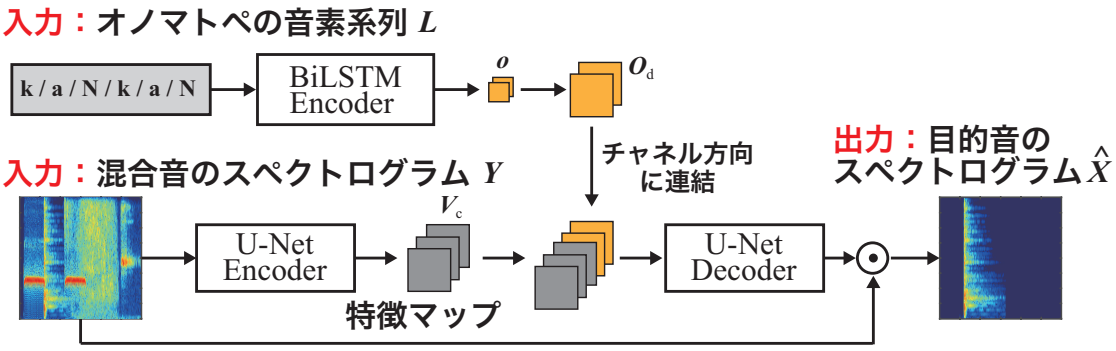


図 B.2: オノマトペを用いた環境音抽出のモデル構造

いる [85]。歌声の場合，対象となる音が常に人の声であるため，ハミングで抽出したい音の特徴を十分表現できる。しかし，環境音抽出の場合，ハミングでは音色を表現することが困難であるのに加えて，鼻濁音などハミングでは表現できない種類の音も存在するため，目的音の特徴を決定付けるには不十分である。

B.3　オノマトペを用いた環境音抽出手法

B.3.1　オノマトペを用いた環境音抽出の概要

提案手法は，オノマトペ w で指定された目的音 $\boldsymbol{x}$ を混合音 $\boldsymbol{y}$ から再構成する。具体的には，混合音 $\boldsymbol{y}$ とオノマトペ w に対し，非線形変換 $\mathsf{Extractor}(\cdot,\cdot)$ を用いて $\hat{\boldsymbol{x}}$ を推定する。

$$\hat{\boldsymbol{x}} = \mathsf{Extractor}(\boldsymbol{y}, w) \tag{B.1}$$

$\mathsf{Extractor}(\cdot,\cdot)$ の詳細は，B.3.2 項で述べる。

B.3.2　オノマトペを用いた環境音抽出手法の提案

Fig. B.2 に提案手法の詳細な構造を示す。提案手法では，U-Net を用いた時間周波数マスク推定並びに，オノマトペからの特徴ベクトル抽出を行う。抽出対象の環境音を指定するため，U-Net の encoder 部の出力に対してオノマトペによる条件付けをする。従来研究 [78, 79, 83] では，抽出する目的音を音響イベントクラスで条件付けしたり，音響イベント検出の結果を組み合わせることで条件付けを行っていた。これらの研究より，混合音のスペクトログラムが CNN を通過したあとの中間的な特徴に対して，音響イベントクラスなどを条件付けすることが有効であるとされている。そこで，提案手法においても U-Net encoder によって得られる

中間的な特徴に対する条件付けを利用する。

提案手法では，Fig. B.2 に示すように，2 種類の特徴量を入力とする。まず 1 つ目の特徴量は，入力された混合音 $\boldsymbol{y}$ から抽出されたスペクトログラム $\boldsymbol{Y} \in \mathbb{R}^{F \times T}$ である。ここで，F 及び T は音響特徴量の次元数と時間フレーム数を表す。2 つ目の特徴量は，オノマトペ w から抽出された埋め込みベクトルである。$\boldsymbol{Y}$ は畳み込み層から構成される U-Net encoder に与えられる。U-Net encoder の各層では，時間方向，周波数方向の次元がそれぞれ半分になり，チャネル数 C が 2 倍に増えていく。そのため，$C\left(=2^K\right)$ の特徴マップが以下のように計算される。

$$[\boldsymbol{V}_1, \ldots, \boldsymbol{V}_C] = \mathsf{UNetEncoder}\left(\boldsymbol{Y}\right) \in \mathbb{R}^{F' \times T' \times C} \tag{B.2}$$

ここで，$\boldsymbol{V}_c \in \mathbb{R}^{F' \times T'} (c = 1, \ldots, C)$ は c 番目のチャンネルの特徴マップを表す。

入力されたオノマトペ w は音素列 L に変換された後，Bi-directional Long Short Term Memory (BiLSTM) による encoder に入力される。そして，オノマトペ全体を捉えた特徴ベクトル $\boldsymbol{o}$ が以下のように抽出される。

$$\boldsymbol{o} = \mathsf{BiLSTMEncoder}\left(\boldsymbol{L}\right) \in \mathbb{R}^{D} \tag{B.3}$$

ここで，D は LSTM のユニット数を表す。そして，$\boldsymbol{o}$ は，時間周波数方向に引き伸ばされ，特徴マップ $\boldsymbol{O}_d \in \mathbb{R}^{F' \times T'}$ が生成される。そして，U-Net encoder と BiLSTM encoder から得られた特徴マップは，チャネル方向に連結し，転置畳み込み層から構成される U-Net decoder に与えられる。最後に，時間周波数ソフトマスク $\boldsymbol{M} \in (0, 1)^{F \times T}$ が，U-Net decoder により推定される。

$$\boldsymbol{Z} = \mathsf{UNetDecoder}\left([\boldsymbol{V}_1, \ldots, \boldsymbol{V}_C, \boldsymbol{O}_1, \ldots, \boldsymbol{O}_D]\right) \boldsymbol{M} = \sigma\left(\boldsymbol{Z}\right) \in (0, 1)^{F \times T} \tag{B.4}$$

推定されたソフトマスク $\boldsymbol{M}$ と混合音から抽出されたスペクトログラム $\boldsymbol{Y}$ の要素積を求めることで目的音のスペクトログラム $\hat{\boldsymbol{X}}$ を得る。

$$\hat{\boldsymbol{X}} = \boldsymbol{M} \odot \boldsymbol{Y} \in \mathbb{R}^{F \times T} \tag{B.5}$$

モデル学習時は，目的音のスペクトログラム $\boldsymbol{X} \in \mathbb{R}^{F \times T}$ と $\hat{\boldsymbol{X}}$ の二乗平均誤差を損失関数として使用する。

$$L\left(\boldsymbol{X}, \hat{\boldsymbol{X}}\right) = \sqrt{\frac{1}{TF}\left\|\boldsymbol{X} - \hat{\boldsymbol{X}}\right\|_F^2} \tag{B.6}$$

推論時は，推定された $\hat{\boldsymbol{X}}$ を環境音波形に復元するため，Griffin–Lim アルゴリズム [51] を使用する。

表 B.1: オノマトペを用いた環境音抽出で使用した各音響イベントクラスの superclass と subclass

Superclass	Subclass	Superclass	Subclass
metal	metal05, metal10, metal15	**bells**	bells1, bells2, bells3, bells4, bells5
dice	dice1, dice2, dice3	**coin**	coin1, coin2, coin3
bottle	bottle1, bottle2	**coins**	coins1, coins2, coins3, coins4, coins5
cup	cup1, cup2		
particl	particl1, particl2	**whistle**	whistle1, whistle2, whistle3
cap	cap1, cap2		
clap	clap1, clap2	**phone**	phone1, phone2, phone3, phone4
claps	claps1, claps2		
clip	clip1, clip2	**toy**	toy1, toy2
bell	bell1, bell2		

B.4 評価実験

B.4.1 学習・評価用データの作成

混合音の作成には，RWCP-SSD を使用した。RWCP-SSD に含まれる音響イベントは，“whistle1”，“whistle2” のように “音響イベント名 + ID” の形式でラベル付けされているものが多く存在する。そこで，同じ音響イベント名を持つラベルをグループ化することで，階層的な音響イベントのクラスを作成した。まず，RWCP-SSD から 44 種類の音響イベント（以降，subclass とする）を選択し，16 種類の音響イベントクラス（以降，superclass とする）にグループ化した。本章で使用した subclass ならびに superclass は Table B.1 に示す。各サブクラスの音を概ね 7 : 2 : 1 の割合で分割し，学習用，検証用，評価用データとして使用した。各音に対応するオノマトペは，3 章にて構築した RWCP-SSD-Onomatopoeia を使用した。RWCP-SSD-Onomatopoeia は各音に対するオノマトペが 15 個以上含まれている。本実験では，1 音あたり 3 個のオノマトペをランダムに選択して，使用した。

本実験では以下の 3 種類の評価用データを作成して提案手法を評価した。

- **Inter-superclass dataset**: データセット内の混合音は，目的音と干渉音から構成されており，干渉音は目的音とは異なる superclass のみを持つ環境音である。

- **Intra-superclass dataset**: データセット内の混合音は，目的音と干渉音か

表 B.2: オノマトペを用いた環境音抽出の実験条件

混合音の長さ	5 s
サンプリング周波数	16 kHz
音波形の圧縮形式	16-bit linear PCM
U-Net encoder のブロック数	4
U-Net decoder のブロック数	4
BiLSTM encoder 層の数	1
BiLSTM encoder 層のユニット数	512
バッチサイズ	8
最適化手法	RAdam
音響特徴量	振幅スペクトログラム
フレーム長	128 ms（2,048 サンプル）
フレームシフト	32 ms（512 サンプル）

ら構成されており，目的音と干渉音の superclass は同じであるが，目的音と干渉音の subclass が異なる。

- **Intra-subclass dataset**: データセット内の混合音は，目的音と干渉音から構成されており，目的音と干渉音の subclass が同じであるが，付与されているオノマトペが異なる。

各データセットの混合音は，信号対雑音比 (SNR) を $\{-10, -5, 0, 5, 10\}$ dB で変化させて作成した。目的音 $\boldsymbol{s}_{\text{target}}$ と干渉音 $\boldsymbol{s}_{\text{interference}}$ の SNR を次式のように定義する。

$$\text{SNR} = 10 \log_{10} \left(\frac{\|\boldsymbol{s}_{\text{target}}\|^2}{\|\boldsymbol{s}_{\text{interference}}\|^2} \right) \tag{B.7}$$

学習データは 7,663 音，検証データは 2,160 音から構成されている。また，3 種類の評価用データはそれぞれ 1,107 音の混合音で構成されている。

B.4.2 実験条件

Table B.2 に提案手法のモデルパラメータ並びに実験条件を示す。また，比較手法として superclass や subclass の音響イベントクラスに応じて条件付けを行う手法も評価をした。superclass や subclass の音響イベントクラスで条件付けを行う場合，それぞれの音響イベントクラスを one-hot で表現して使用した。

各手法を評価するため，評価指標として Signal-to-Distortion Ration (SDR) [86] の向上率である SDR Improvement (SDRi) を用いた。SDRi は，混合音に対する

表 B.3: オノマトペを用いた環境音抽出の SDRi [dB] による評価結果

Dataset	Method	SNR −10 dB	−5 dB	0 dB	5 dB	10 dB
Inter-superclass dataset	Superclass-conditioned method	5.11 ± 3.02	4.72 ± 2.75	4.06 ± 2.55	2.70 ± 2.13	1.33 ± 2.12
	Subclass-conditioned method	5.06 ± 2.97	4.75 ± 2.85	4.04 ± 2.52	2.81 ± 2.31	1.25 ± 2.09
	Onomatopoeia-conditioned method	4.63 ± 2.58	4.57 ± 2.69	4.02 ± 2.53	2.77 ± 2.22	1.41 ± 2.12
Intra-superclass dataset	Superclass-conditioned method	2.05 ± 2.37	1.97 ± 2.40	1.86 ± 2.38	1.50 ± 2.19	0.82 ± 1.89
	Subclass-conditioned method	5.03 ± 2.56	4.77 ± 2.59	4.19 ± 2.45	2.74 ± 2.12	1.26 ± 2.06
	Onomatopoeia-conditioned method	5.61 ± 2.78	5.36 ± 2.75	4.73 ± 2.52	3.10 ± 2.27	1.42 ± 2.06
Intra-subclass dataset	Superclass-conditioned method	2.03 ± 2.40	2.06 ± 2.54	1.87 ± 2.37	1.49 ± 2.09	0.79 ± 1.98
	Subclass-conditioned method	3.14 ± 2.78	3.09 ± 2.77	2.84 ± 2.63	2.21 ± 2.29	1.01 ± 2.12
	Onomatopoeia-conditioned method	5.83 ± 2.43	5.68 ± 2.53	5.11 ± 2.58	3.34 ± 2.24	1.64 ± 2.02

目的音の SDR と，抽出された音に対する SDR の差として以下で定義される。

$$\mathrm{SDRi} = 10\log_{10}\left(\frac{\|\boldsymbol{x}\|^2}{\|\boldsymbol{x}-\hat{\boldsymbol{x}}\|^2}\right) - 10\log_{10}\left(\frac{\|\boldsymbol{x}\|^2}{\|\boldsymbol{x}-\boldsymbol{y}\|^2}\right) \tag{B.8}$$

SDRi による評価は，B.4.1 項で作成した評価用データセットそれぞれに対して実施した。

B.4.3 実験結果

Table B.3 に各評価用データにおける SDRi を示す。結果より，inter-superclass dataset を用いた評価において，superclass によって条件付けする手法は，目的音のみを抽出できている。一方，intra-superclass 並びに intra-subclass dataset では SDRi が低下することが確認された。また，subclass で条件付けする手法は，inter-superclass 並びに intra-superclass dataset では目的音のみを抽出できるが，intra-subclass dataset では SDRi が低下することが確認された。提案手法（Table C.6 の onomatopoeia-conditioned method）は，3 種類のデータセット全てにおいて同程度の SDRi であった。この結果より，オノマトペは subclass よりも細かなクラスとして振る舞えることを示唆した。

Fig. B.3 に subclass によって条件付けした手法と提案手法によって抽出された音のスペクトログラムを示す。なお，混合音には intra-subclass dataset の音を使用した。図より，subclass で条件付けする手法は目的音以外も抽出されているのに対し，提案手法では目的音のみ抽出されていることが確認できる。提案手法は，superclass または subclass で条件付けする手法よりも高い抽出性能を示したが，目的音と干渉音が重なり合っている場合は抽出が難しいという結果となった。重なりあった音の抽出に関しては，今後さらなる検討が必要である。

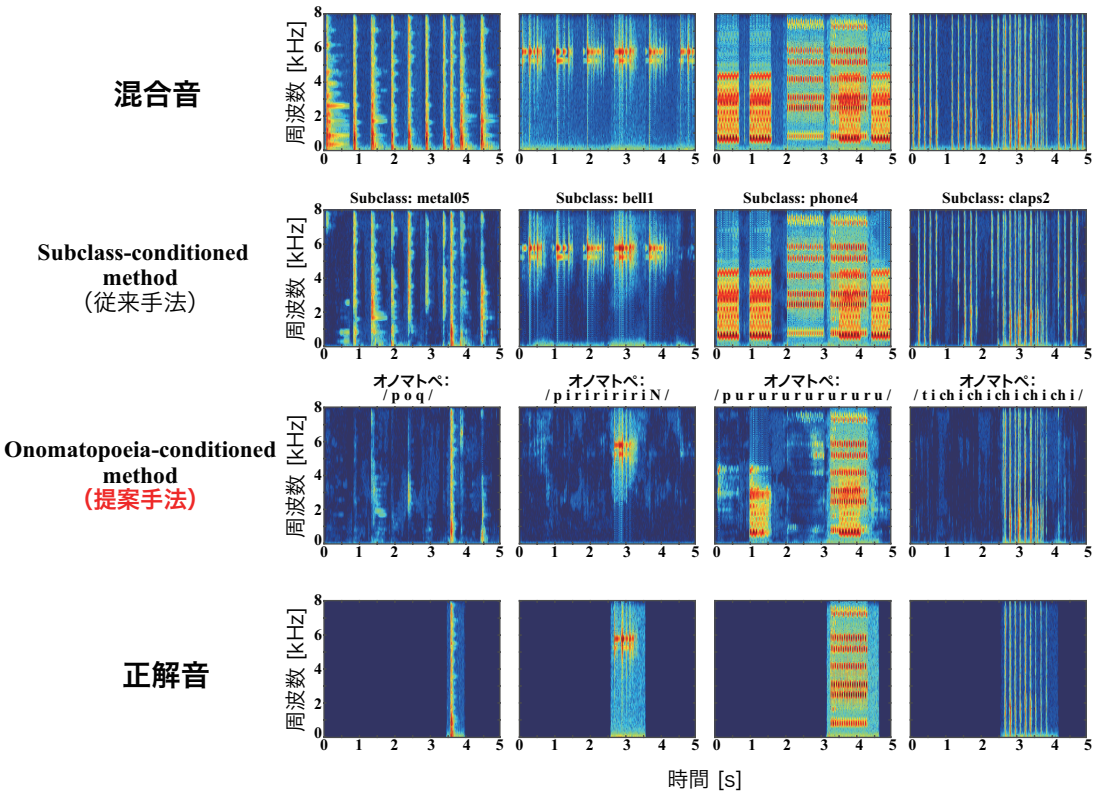


図 B.3: オノマトペを用いた環境音抽出によって抽出された環境音のスペクトログラム

B.5　付録Bのまとめ

本章では，オノマトペを用いた環境音抽出手法を提案した。提案手法では，オノマトペで指定された目的音の時間周波数マスクを U-Net を用いて推定し，混合音との要素積を求めることで目的音のみを抽出した。実験結果より，提案手法はオノマトペを入力として，混合音から特定の音のみを抽出可能であることを確認した。また提案手法は，音響イベントクラスで条件付けする従来手法よりも優れた性能を示した。この結果より，オノマトペは，音のラベリングに特別なドメイン知識を必要としないにもかかわらず，音響イベントクラスよりも細かなクラスのように振る舞えることが明らかになった。音に対し付与されるオノマトペは，言語圏によって異なるため，今後，異なる言語の話者が付与したオノマトペに対しても提案手法が有効であるかどうか検証する。

付 録C　単一音源に説明文を付与した環境音データセットの構築

C.1　はじめに

映画やゲームといったメディアコンテンツ作品において，臨場感や没入感を高めるために環境音は欠かせない存在である。目的とする環境音を得るための方法として，大規模な環境音データベースから音を探して，利用する方法が考えられる。しかし，必要とする環境音がデータベース内に存在するとは限らない。一方，インターネット上には大量の環境音データが存在するが，それらの多くは複数の音響イベントの音が混ざり合った混合音であることが多く，目的とする音のみを得ることは容易ではない。

単一の音響イベントのみを含む環境音（以下，単一音源）を得るため，複数の音が混ざり合った混合音から目的とする音のみを抽出する環境音抽出手法が提案されている [79, 83, 84, 87, 88, 89]。例えば，音響イベントクラスを指定して，指定した音響イベントクラスに対応する音のみを抽出する手法が提案されている [78, 79, 83]。また，付録 B で提案したように，抽出したい音を音の特徴を模倣したオノマトペによって表現し，音響イベントクラスよりもさらに細かい粒度で音を抽出する手法もある。しかしながら，オノマトペは言語や文化に依存するため，環境音抽出モデルの学習に使用するデータセットをシステムを使用するユーザに合わせて変更する必要がある。言語や文化に依存しない方法として，音の様子を説明した文（以下，音の説明文）によって抽出したい音の特徴を表現し，目的とする音のみを抽出する手法も提案されている [90, 91, 92]。従来の音の説明文を用いた環境音抽出では，複数音源を含む環境音に対して説明文が付与されたデータセット [93, 94, 95] を使用している。表 C.1 に示すように，複数音源を含む環境音に対して付与された音の説明文は，単一音源に対して付与された説明文と比べて，1 個の音響イベントに対する説明の詳細度合いが低い。そのため，混合音から単一音源の環境音のみを抽出するためには，単一音源に対して付与された音の説明文が必要である。

表 C.1: 構築したデータセットに含まれる音の説明文の例

音源の数	音の説明文
単一	• One or two keyboards continue to be pressed slowly with a light, high-touch sound.
	• The sound of the buttons on the keyboard being pressed one by one quite slowly.
	• The electronic tone of the digital alarm clock is high-pitched and continues to sound slowly to gradually faster.
	• The bright, high-pitched alarm clock bells are ringing.
複数	• Keyboard typing and mouse clicking noises are continually heard.
	• The sound of a toaster rang out as if to counteract the sound of typing on a small keyboard.
	• The alarm clock is drowned out by the loud noise of the hair dryer.
	• The alarm clock beeps at equal intervals, with one final sound to close the lock.

目的とする音を得る手段として，上述した環境音抽出手法だけでなく，5 から 7 章までで述べた環境音を人工的に作り出す環境音合成技術も考えられる。これまでの章にて，環境音を合成する手法として，音響イベントクラスやオノマトペを入力とした環境音合成手法を提案した。また，音の説明文を入力とした環境音合成 [19, 20, 21, 96] も提案されている。しかしながら，音の説明文を利用した環境音抽出と同様に，単一の音のみを表現した音の説明文が存在しないため，音の説明文を利用した環境音合成においても，生成する音をより細かに調整することが困難であると考えられる。

そこで本章では，環境音を利用する様々なタスクにおいて利用可能な，単一音源に対して音の説明文を付与したデータセットを構築する。なお，作成したデータセットを CAPTDURE (CAPTioned sound Dataset of single soURcEs) と呼ぶ。また，構築したデータセットの利用例として，環境音抽出の実験を行い，単一音源に対して付与された音の説明文の利用が，混合音中の特定の音のみを抽出するために有効であることを示す。

C.2 データセットの構築

C.2.1 構築したデータセットの概要

以下の内容から構成されるデータセットを構築した。

- **単一音源の環境音**
  表 C.2 に示す 14 種類の日常生活で発生する音響イベントの音を収録した。

表 C.2: 収録した音響イベントクラス

音響イベントクラス	音の説明	Subclass の数	音ファイル数	音の長さ [s]
Keyboard	キーボードのタイピング音	5	60	465.0
Door	扉の開閉音	4	48	279.0
Mouse	マウスのクリック音	7	84	600.0
Water tap running	水の流れる音	5	60	431.0
Dryer	ドライヤーの稼働音	6	252	1,746.0
Ventilation fan	換気扇の稼働音	3	36	284.0
Door lock	鍵を閉める音	6	48	245.0
Intercom	インターホンを鳴らす音	5	56	350.0
Door knock	ドアをノックする音	6	72	430.0
Microwave	電子レンジの稼働音	4	48	329.0
Toaster	トースターの稼働音	5	60	465.0
Cutlery	食器がぶつかる音	7	52	362.0
Clock	目覚まし時計が鳴る音	5	60	420.0
Fan	扇風機の稼働音	3	108	779.0
合計		71	1,044	7,185.0

各音響イベントに対し，音の鳴らし方などの収録条件を変えながら，約 40〜100 音程度の音を収録した。各音の長さは 5〜9 秒に設定した。また，各音響イベントクラス内は，音の鳴り方や音源の種類などの違いによってサブクラスに分割した。

- **複数音源を含む環境音**
  収録した単一音源の環境音を用いて，合計 1,044 音の複数音源を含む環境音を作成した。収録した環境音の中から異なる音響イベントクラスの環境音を 2 音ランダムに選び，SNR が 0 dB になるように混合した。なお，混合する音の系列長が一致しない場合は，系列長の短い音の末尾を 0 埋めすることによって 2 音の音の系列長を揃えた。

- **単一音源の環境音に対する音の説明文**
  各単一音源の環境音に対して 3 個以上，計 4,902 個の説明文を収集した。音に対する説明文の収集方法については C.2.3 項にて述べる

- **複数音源を含む環境音に対する音の説明文**
  各複数音源を含む環境音に対して 3 個，計 3,132 個の説明文を収集した。

- **各音の説明文の対する妥当性のスコア**
  音の説明文を付与した作業者とは異なる作業者から，音に対して収集した説明文の妥当性のスコアを収集した。このスコアによって，他者による音の説明文の評価が可能である。妥当性のスコアの収集方法に関しては C.2.4 項で述べる。

表 C.3: 環境音の収録に使用した機材

収録機材	ソフトウェア	マイクロフォン	オーディオ インターフェース
Apple/Macbook Pro	Audacity[1]	SHURE/MX150B/O-XLR	Roland/Rubix24
TASCAM/DR-44WL	内蔵	内蔵	内蔵
Apple/iPhone SE	PCM recording[2]	内蔵	内蔵
Huawei/dtab d01-G	AudioRec[3]	内蔵	内蔵

- **作業者 ID**
  音の説明文並びに説明文の妥当性スコアを付与した作業者の匿名化された ID を含む。

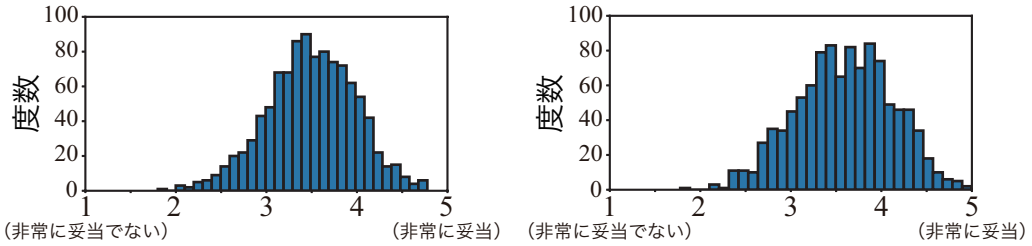
C.2.2 環境音の収録環境と条件

データセットに含まれる環境音は，2 箇所の防音室（防音室 A，防音室 B とする）にて収録した。防音室 A は，広さ 3.1 m × 5.4 m × 2.7 m，残響時間は 0.2 秒である。防音室 B は，広さ 4.9 m × 4 m × 2.5 m，残響時間は 0.2 秒である。一部，防音室にて収録が困難である環境音に関しては，鉄筋コンクリート造マンションの一室，並びに戸建木造住宅の一室にて収録した。収録に使用した機材を Table C.3 に示す。各機材にて，サンプリング周波数は 48kHz，量子化ビット数は 16bit で収録を行った。収録時は，収録対象物とマイクロフォンの距離を約 0.3 m〜0.5 m に設定した。収録対象音の音量が極めて大きい場合は，対象物とマイクロフォンの距離を十分に確保して収録した。

C.2.3 収録した音に対する説明文の収集

本章では、クラウドソーシングサービスである Lancers を利用して日本語話者から音に対する日本語の説明文を収集した。クラウドソーシングサービスは，効率的に人手を集めることが可能であるのに加え，多くの作業者から音の説明文を収集することが可能であるため，1 音に対しても多様な説明文の収集が期待できる。事前実験として，Amazon の提供するクラウドソーシングサービスである Amazon mechanical Turk (MTurk) の利用も試みた。しかしながら，Lancers と MTurk からそれぞれ収集した音の説明文を比較したところ，Lancers を利用するほうが質の高い音の説明文を収集できることが確認された。そのため，本章においては Lancers にて音の説明文を収集した。

音の説明文の収集では，1 人の作業者につき 5 音を提示した。予備実験において，5 音が全て異なる音響イベントとなるように音を提示した場合，「電子レンジ



(a) Single-source sounds　　(b) Multiple-source sounds

図 C.1: 音の説明文に対して収集した妥当性スコアのヒストグラム

の音」のように各音に付与された説明文の詳細度合いが下がることが確認された。そのため本章では，提示する 5 音は同一の音響イベントからなる音を提示し，各音に付与する説明文は必ず異なる文を記述するように指示した。それに加え作業者に対して,「電子レンジの音」のような音の種類だけの回答や,「カンカン」のようなオノマトペのみの回答は避けるように指示した。

単一音源の環境音に対しては各音 3 個以上，計 4,902 個の説明文を収集した。また，複数音源を含む環境音に対しては，各音につき 3 個，計 3,132 個の説明文を収集した。また，日本語の音の説明文に加えて，DeepL API を用いて英語に翻訳した音の説明文を収録している。

C.2.4　収集した音の説明文の評価

C.2.3 項にて収集した音の説明文が環境音に対してどの程度妥当であるかをクラウドソーシングサービスを用いて評価した。環境音と付与された音の説明文を作業者に提示する。作業者は，それぞれの環境音に対して付与された説明文がどの程度妥当であるかについて 1（非常に妥当でない）〜5（非常に妥当である）の 5 段階で回答した。なお，各音の説明文に対して 3 名以上作業者から妥当性のスコアを収集した。

図 C.1 に収集した妥当性スコアのヒストグラムを示す。なお，図は，各音ごとに収集されたスコアの平均をヒストグラムにしたものである。図より，ほとんどの音に対して，妥当性のスコアが 3 以上である説明文が付与されていることがわかる。よって，各音を表現するために妥当な多くの説明文を収集できたことが確認できる。

表 C.4: 構築したデータセットにおける学習，検証，評価セットの統計的情報

データセット	音ファイル数	音の説明文の数	単語数 (en) / 文字数 (ja)
単一音源の環境音			
学習	795	3,774	10.53 (en) / 22.80 (ja)
検証	82	374	10.62 (en) / 22.91 (ja)
評価	167	501	10.43 (en) / 23.05 (ja)
複数音源の環境音			
学習	795	2,385	16.83 (en) / 36.47 (ja)
検証	82	246	16.60 (en) / 35.53 (ja)
評価	167	754	17.79 (en) / 38.06 (ja)

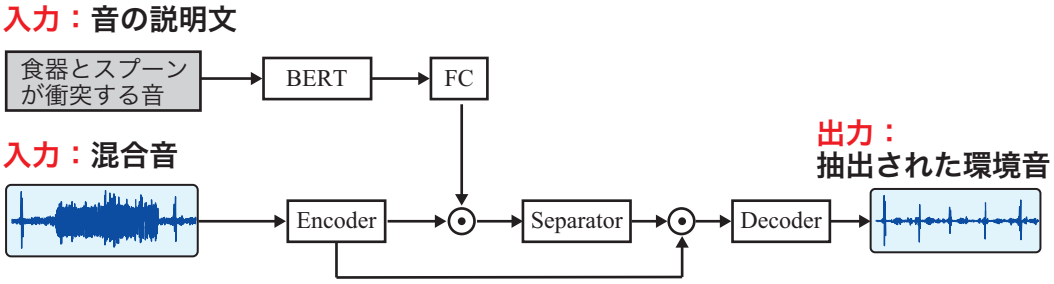


図 C.2: 音の説明文に基づく環境音抽出のモデル概要

C.2.5　収集したデータの分割

単一音源の環境音と複数音源を含む環境音を音響イベントの各サブクラスを学習，検証，評価でそれぞれ約 7:1:2 の割合になるように分割した。各音に付与された音の説明文も同様に分割を行った。各セットの統計的な情報を表 C.4 に示す。

C.3　評価実験

構築したデータセットの性能評価のため，音の説明文をクエリとして混合音から目的の環境音のみを抽出する環境音抽出抽出タスクを実施する。

C.3.1　音の説明文に基づく環境音抽出のモデル構造

本実験では，Fig. C.2 に示す Conv-TasNet [97] を基とした環境音抽出モデルを構築した。まず，混合音 $\boldsymbol{x} \in \mathbb{R}^{1\times T}$ を 1 次元畳み込みニューラルネットワーク (1-D convolution) から構成される encoder に入力する。

$$\boldsymbol{W} = \mathsf{Encoder}\left(\boldsymbol{x}\right) \in \mathbb{R}^{C\times T'} \tag{C.1}$$

次に，音の説明文 $\boldsymbol{l}$ を Bidirectional Encoder Representations from Transformers (BERT) [98] に入力する。なお，本実験においては，学習データに使用する言語に応じて，日本語，英語でそれぞれ事前学習された BERT のモデルを使用する。そして、BERT から得られた特徴ベクトルを線形層 $Linear(\cdot)$ に入力して，C 次元の特徴ベクトルに変換する。

$$\boldsymbol{o} = \mathsf{Linear}\left(\mathsf{BERT}\left(\boldsymbol{l}\right)\right) \in \mathbb{R}^{C} \tag{C.2}$$

次に，encoder の出力から得られた行列 $\boldsymbol{W}$ と特徴ベクトル $\boldsymbol{o}$ を用いて，ソフトマスク $\boldsymbol{M}$ を得る。

$$\boldsymbol{M} = \mathsf{Separator}\left(\ \boldsymbol{W} \odot [\underbrace{\boldsymbol{o}, \boldsymbol{o}, \ldots, \boldsymbol{o}}_{T'}]\ \right) \in (0,1)^{C \times T'} \tag{C.3}$$

$\odot$， $\mathsf{Separator}(\cdot)$ はそれぞれ行列の要素積，1-D convolution を表す。最後に，ソフトマスク $\boldsymbol{M}$ と Encoder の出力 $\boldsymbol{W}$ の要素積を計算したものを 1-D convolution から構成された Decoder に入力することで，入力となった音の説明文に対応する環境音のみを得る。

$$\hat{\boldsymbol{y}} = \mathsf{Decoder}\left(\boldsymbol{M} \odot \boldsymbol{W}\right) \in \mathbb{R}^{1 \times T} \tag{C.4}$$

モデル学習の際は，モデルから推定された波形 $\hat{\boldsymbol{y}}$ と正解波形 $\boldsymbol{y}$ の L1 ノルムを損失関数として利用する。

C.3.2 学習・評価用データの作成

収録した単一音源の環境音からランダムにそれぞれ異なる音響イベントクラスを持つ 3 音を選択して，混合音を作成した。学習セットと検証セットはそれぞれ，2,385 音，246 音の混合音から構成される。音に対する説明文は，1 音あたりのデータ数を揃えるために 1 音につき 3 通りの文をランダムに選択し，それぞれを目的音に対する説明文として使用した。

収録した単一音源の環境音を利用して，inter-event-class dataset と intra-event-class dataset の 2 種類の評価用データセットを作成した。混合音は目的音とそれ以外の干渉音から構成されており，inter-event class dataset は目的音と干渉音が異なる音響イベントクラスによって構成された混合音からなるデータセットとする。また，intra-event class dataset は目的音と干渉音が同一の音響イベントの混合音から構成されたデータセットとする。各混合音の SNR は $\{-10, -5, 0, 5, 10\}$ dB で変化させて作成した。なお，各評価用データセットは各 SNR ごとに 501 音の混合

表 C.5: 音の説明文に基づく環境音抽出の実験条件

音波形	
混合音の長さ	10 s
サンプリング周波数	16 kHz
音波形の圧縮形式	16-bit linear PCM
Conv-TasNet パラメータ	
オートエンコーダのフィルタ数	256
フィルタの長さ	20
ボトルネック層のチャネル数	256
CNN 層のチャネル数	512
CNN 層のカーネルサイズ	3
CNN 層の数	8
モデル学習のパラメータ	
バッチサイズ	1
エポック数	120
学習率	0.0001
最適化手法	RAdam

音で構成される。

C.3.3 実験条件

本実験で使用するモデルのパラメータ並びに実験条件は Table C.5 に示す。本実験では，以下の 2 種類のモデル学習方法を比較して，単一音源に対して付与された説明文の利用が，目的とする音のみの抽出に有効であることを示す。

- **Training using multiple-source caption**
  複数音源を含む環境音に付与された音の説明文を利用したモデル学習方法。本モデル学習手法は，従来のデータセットに含まれる音の説明文を用いる場合と同等であると言える。

- **Training using single-source caption**
  単一音源に対して付与された説明文を利用したモデル学習方法。

評価では，C.3.2 項にて作成した評価用データセットを用いて，各モデル学習手法ごとに環境音抽出の性能を評価した。

[1] https://www.audacityteam.org/
[2] https://ko-yasui.com/
[3] https://audiorec.jp.aptoide.com/app

表 C.6: 音の説明文に基づく環境音抽出の SDRi [dB] による評価結果

データセット	モデル学習に使用した説明文	SNR −10 dB	−5 dB	0 dB	5 dB	10 dB
Inter-event-class dataset	Multiple-source caption (ja)	5.12 ± 4.22	4.39 ± 4.04	3.00 ± 4.21	0.93 ± 4.92	-1.85 ± 5.95
	Multiple-source caption (en)	4.67 ± 4.03	3.88 ± 3.88	2.43 ± 4.04	0.17 ± 4.71	-2.78 ± 5.67
	Single-source caption (ja)	7.28 ± 5.19	6.26 ± 5.02	4.63 ± 5.18	2.09 ± 5.71	-1.08 ± 6.54
	Single-source caption (en)	6.42 ± 4.84	5.14 ± 4.66	3.13 ± 4.89	0.40 ± 5.54	-2.91 ± 6.40
Intra-event-class dataset	Multiple-source caption (ja)	3.22 ± 2.78	2.49 ± 2.37	1.11 ± 2.31	-1.01 ± 3.08	-3.93 ± 4.34
	Multiple-source caption (en)	3.09 ± 2.50	2.38 ± 2.11	1.15 ± 2.09	-0.86 ± 2.90	-3.61 ± 4.22
	Single-source caption (ja)	4.36 ± 3.60	3.24 ± 3.18	1.36 ± 3.01	-1.33 ± 3.66	-4.69 ± 4.73
	Single-source caption (en)	4.47 ± 3.09	3.23 ± 2.41	1.29 ± 2.49	-1.48 ± 3.27	-4.87 ± 4.42

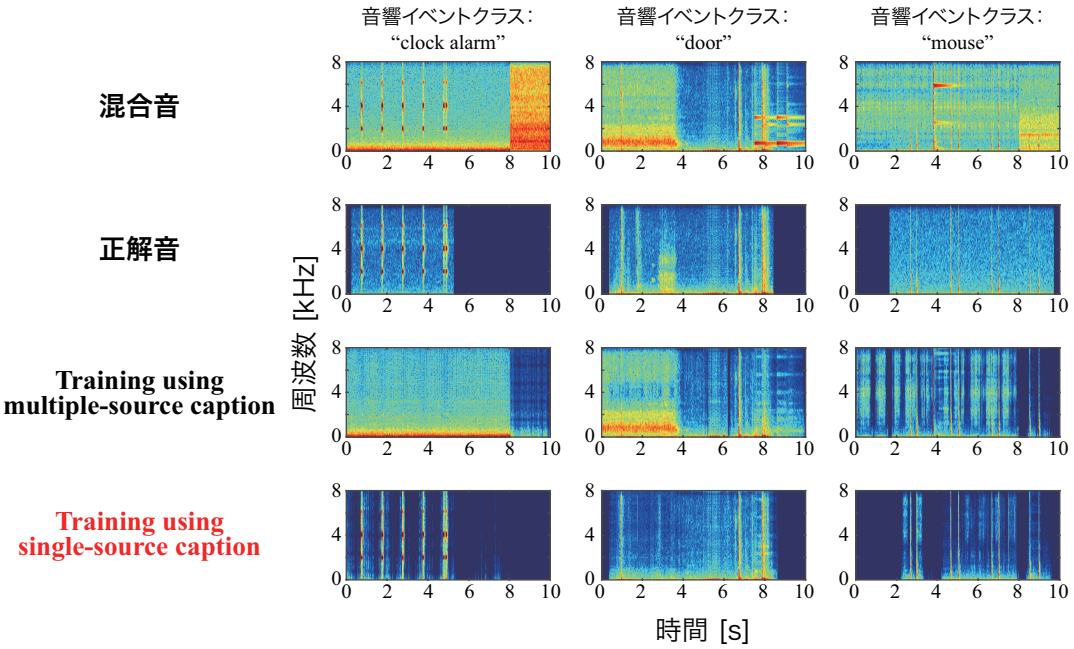


図 C.3: 音の説明文に基づく環境音抽出によって抽出された環境音のスペクトログラム

各モデル学習方法の評価には，SDRi を用いた。SDRi は，混合音に対する目的音の SDR と，抽出された音に対する SDR の差として以下で定義される。

$$\mathrm{SDRi} = 10\log_{10}\left(\frac{\|\boldsymbol{y}\|^2}{\|\boldsymbol{y}-\hat{\boldsymbol{y}}\|^2}\right) - 10\log_{10}\left(\frac{\|\boldsymbol{y}\|^2}{\|\boldsymbol{y}-\boldsymbol{x}\|^2}\right) \tag{C.5}$$

$\boldsymbol{x}$，$\boldsymbol{y}$，$\hat{\boldsymbol{y}}$ はそれぞれ，混合音，正解音，抽出音を表す。

C.3.4　実験結果

Table C.6 に各手法における評価データに対する SDRi を示す。結果より，単一音源の環境音に付与した説明文によってモデル学習させる手法 (training using single-source caption) は，複数音源を含む環境音に対して付与された説明文によってモデル学習する手法 (training using multiple-source caption) よりも，説明文にて指定した目的音を精度高く抽出可能であることが確認された。また，intra-event-class dataset における評価結果より，単一音源の環境音に付与した説明文を利用することで，混合音中に同一の音響イベントが含まれている場合においても，目的音のみを抽出可能であることがわかった。

Fig. B.3 に各手法によって抽出された音のスペクトログラムを示す。図は，inter-event-class dataset の SNR が 0 dB の混合音に対して抽出を行った結果である。図より，複数音源を含む環境音に付与した説明文を利用した場合は，目的音以外の環境音も抽出されてしまうことが確認できる。一方，単一音源の環境音に付与した説明文を利用した場合は，説明文で指定した目的音のみを抽出できていることがわかる。これらの結果より，単一音源の環境音に付与した説明文を環境音抽出のモデル学習に使用することで，目的とする環境音をより精度高く抽出可能であることが明らかになった。

C.4　付録 C のまとめ

本章では，環境音を利用する様々なタスクで利用可能な単一音源の環境音に対する音の説明文を付与したデータセットを構築した。計 1,044 個の単一音源の環境音を収録して，収録した音に対してクラウドソーシングサービスを用いて計 4,902 個の音の説明文を収集した。混合音から目的音を抽出する環境音抽出タスクにおいて，単一音源の環境音に対する説明文を利用することで，混合音から精度高く目的音のみを抽出可能であることが明らかになった。今後，環境音合成など，環境音を利用するその他のタスクにおいても本データセットの有効性を検証する必要がある。

参考文献

[1] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *arXiv preprint, arXiv:1609.03499*, 2016.

[2] D. B. Lloyd, N. Raghuvanshi, and N. K. Govindaraju, "Sound synthesis for impact sounds in video games," *Proc. Symposium on Interactive 3D Graphics and Games. ACM*, pp. 55–61, 2011.

[3] K. Wang, H. Cheng, and S. Liu, "Efficient sound synthesis for natural scenes," *Proc. IEEE Virtual Reality* (*VR*), pp. 303–304, 2017.

[4] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), pp. 344–348, 2017.

[5] Q. Kong, Y. Xu, T. Iqbal, Y. Cao, W. Wang, and M. D. Plumbley, "Acoustic scene generation with conditional sampleRNN," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 925–929, 2019.

[6] D. Svozil, V. Kvasnicka, and J. Pospichal, "Introduction to multi-layer feed-forward neural networks," *Chemometrics and Intelligent Laboratory Systems*, vol. 39, no. 1, pp. 43–62, 1997. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169743997000610

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Proc. Advances in neural information processing systems*, vol. 25, 2012.

[8] J. L. Elman, "Finding structure in time," *Cognitive Science*, vol. 14, no. 2, pp. 179–211, 1990.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Proc. Advances in Neural Information Processing Systems*, vol. 30, 2017.

[11] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, "SampleRNN: An unconditional end-to-end neural audio generation model," *Proc. International Conference for Learning Representations* (*ICLR*), pp. 1–11, 2017.

[12] F. Gontier, M. Lagrange, C. Lavandier, and J. F. Petiot, "Privacy aware acoustic scene synthesis using deep spectral feature inversion," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 886–890, 2020.

[13] E. Grinstein, N. Q. K. Duong, A. Ozerov, and P. Pérez, "Audio style transfer," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 586–590, 2018.

[14] P. K. Mital, "Time domain neural audio style transfer," *arXiv preprint arXiv:1711.11160*, 2017.

[15] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), pp. 2414–2423, 2016.

[16] D. Schwarz, "State of the art in sound texture synthesis," *Proc. Digital Audio Effects* (*DAFx*), pp. 221–232, 2011.

[17] H. Caracalla and A. Roebel, "Sound texture synthesis using RI spectrograms," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 416–420, 2020.

[18] Y. Zhou, Z. Wang, C. Fang, T. Bui, and T. L. Berg, "Visual to sound: Generating natural sound for videos in the wild," *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), pp. 3550–3558, 2018.

[19] F. Kreuk and G. Synnaeve and A. Polyak and U. Singer and A. Défossez and J. Copet and D. Parikh and Y. Taigman and Y. Adi, "AudioGen: Textually guided audio generation," in *Proc. International Conference on Learning Representation* (*ICLR*), 2023.

[20] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "AudioLDM: Text-to-audio generation with latent diffusion models," *arXiv preprint arXiv:2301.12503*, 2023.

[21] D. Yang, J. Yu, H. Wang, W. Wang, C. Weng, Y. Zou, and D. Yu, "Diffsound: Discrete diffusion model for text-to-sound generation," *arXiv preprint arXiv:2207.09983*, 2022.

[22] S. Sundaram and S. Narayanan, "Vector-based representation and clustering of audio using onomatopoeia words," *Proc. American Association for Artificial Intelligence* (*AAAI*) *Symposium Series*, pp. 55–58, 2006.

[23] G. Lemaitre and D. Rocchesso, "On the effectiveness of vocal imitations and verbal descriptions of sounds," *The Journal of the Acoustical Society of America*, vol. 135, no. 2, pp. 862–873, Feb. 2014.

[24] S. Sundaram and S. Narayanan, "Classification of sound clips by two schemes: Using onomatopoeia and semantic labels," *Proc. IEEE International Conference on Multimedia and Expo*, pp. 1341–1344, 2008.

[25] B. Kim and B. Pardo, "Improving content-based audio retrieval by vocal imitation feedback," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 4100–4104, 2019.

[26] S. Nakamura, K. Hiyane, F. Asano, and T. Endo, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," *Proc. Language Resources and Evaluation Conference* (*LREC*), pp. 965–968, 2000.

[27] K. Drossos, S. Lipping, and T. Virtanen, "CLOTHO: An audio captioning dataset," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 736–740, 2020.

[28] S. Lipping, K. Drossos, and T. Virtanen, "Crowdsourcing a dataset of audio captions," *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events* (*DCASE*), pp. 139–143, 2019.

[29] S. Takamichi and H. Saruwatari, "CPJD corpus: Crowdsourced parallel speech corpus of japanese dialects," *Proc. Language Resources and Evaluation Conference* (*LREC*), pp. 434–437, 2018.

[30] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JVS corpus: free Japanese multi-speaker voice corpus," *arXiv preprint arXiv:1908.06248*, 2019.

[31] "Speech Segmentation Toolkit using Julius," https://github.com/julius-speech/segmentation-kit.

[32] X. Liu, T. Iqbal, Z. Turab, J. Zhao, Q. Huang, M. D. Plumbley, and W. Wang, "Conditional sound generation using neural discrete time-frequency representation learning," *Proc. IEEE International Workshop on Machine Learning for Signal Processing* (*MLSP*), pp. 1–6, 2021.

[33] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet audio distance: A metric for evaluating music enhancement algorithms," *arXiv preprint arXiv:1812.08466*, 2018.

[34] X. Zhou, Z.-H. Ling, and S. King, "The blizzard challenge 2020," *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, pp. 1–18, 2020.

[35] Y. Okamoto, K. Imoto, T. Komatsu, S. Takamichi, T. Yagyu, R. Yamanishi, and Y. Yamashita, "Overview of tasks and investigation of subjective evaluation methods in environmental sound synthesis and conversion," *arXiv preprint arXiv:1908.10055*, 2019.

[36] Y. Okamoto, K. Imoto, S. Takamichi, R. Yamanishi, T. Fukumori, and Y. Yamashita, "Onoma-to-wave: Environmental sound synthesis from onomatopoeic words," *APSIPA Transactions on Signal and Information Processing*, vol. 11, e13, 2022.

[37] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to seqeuce learning with neural networks," *arXiv preprint arXiv:1409.3215*, 2014.

[38] J.-Y. Liu, Y.-H. Chen, Y.-C. Yeh, and Y.-H. Yang, "Unconditional audio generation with generative adversarial networks and cycle regularization," *arXiv preprint arXiv:2005.08526*, 2020.

[39] D. Yang, J. Yu, H. Wang, W. Wang, C. Weng, Y. Zou, and D. Yu, "Diffsound: Discrete diffusion model for text-to-sound generation," *arXiv preprint arXiv:2207.09983*, 2022.

[40] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," *Proc. International Conference on Learning Representation* (*ICLR*), pp. 1–13, 2020.

[41] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 776–780, 2017.

[42] "BBC Sound Effects," https://sound-effects.bbcrewind.co.uk/.

[43] A. Mesaros, T. Heittola, and T. Virtanen, "Tut database for acoustic scene classification and sound event detection," *Proc. European Signal Processing Conference (EUSIPCO)*, pp. 1128–1132, 2016.

[44] "Pulse code modulation (PCM) of voice frequencies," *ITU-T Recommendation P.711*, 1988.

[45] https://www.ksuke.net/demos.

[46] "KanaWave," https://www.vector.co.jp/soft/win95/art/se232653.html.

[47] S. Ikawa and K. Kashino, "Generating sound words from audio signals of acoustic events with sequence-to-sequence model," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 346–350, 2018.

[48] Y. Wang, R. S.-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[49] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence - video to text," *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 4534–4542, 2015.

[50] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, N. Jaitly, B. Li, J. Chorowski, and M. Bacchiani, "State-of-the-art speech recognition with sequence-to-sequence models," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4774–4778, 2018.

[51] D. Griffin and J. Lim, "Signal estimation form modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[52] S. Ikawa and K. Kashino, "Neural audio captioning based on conditional sequence-to-sequence model," *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events* (*DCASE*), pp. 99–103, 2019.

[53] J.-X. Zhang, Z.-H. Ling, and L.-R. Dai, "Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 540–552, 2020.

[54] M. Cartwright and B. Pardo, "Vocalsketch: Vocally imitating audio concepts," in *Proc. Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 43–46.

[55] B. Kim and M. Ghei and B. Pardo and Z. Duan, "Vocal imitation set: a dataset of vocally imitated sound events using the audioset ontology," in *Proc. Detection and Classification of Acoustic Scenes and Events* (*DCASE*), 2018, pp. 148–152.

[56] Y. Zhang, B. Pardo, and Z. Duan, "Siamese style convolutional neural networks for sound search by vocal imitation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 429–441, 2019.

[57] M. Cartwright and B. Pardo, "Synthassist: An audio synthesizer programmed with vocal imitation," in *Proc. the 22nd ACM International Conference on Multimedia*, 2014, pp. 741–742.

[58] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2018, pp. 4779–4783.

[59] R. Takizawa and S. Hirai, "Synthesis of explosion sounds from utterance voice of onomatopoeia using transformer," in *Proc. Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, 2023, pp. 87–90.

[60] J. Park, S. Takamichi, T. Nakamura, K. Seki, D. Xin, and H. Saruwatari, "How generative spoken language modeling encodes noisy speech: investigation from phonetics to syntactics," in *Proc.* (*INTERSPEECH*), 2023, pp. 1085–1089.

[61] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the royal statistical society. series c (applied statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[62] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: an open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2022.

[63] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Byol for audio: Exploring pre-trained general-purpose audio representations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 137–151, 2023.

[64] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proc. the 23rd ACM International Conference on Multimedia*, 2015, pp. 1015–1018.

[65] https://github.com/nttcslab/byol-a/tree/master/v2.

[66] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed, and E. Dupoux, "On Generative Spoken Language Modeling from Raw Audio," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1336–1354, 2021.

[67] J. Kong and J. B. J. Kim, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[68] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *ACM International Conference on Multimedia*, 2014, pp. 1041–1044.

[69] https://github.com/liuxubo717/sound_generation.

[70] https://github.com/DCASE2023-Task7-Foley-Sound-Synthesis/dcase2023_task7_baseline.

[71] https://voice-to-foley.github.io/.

[72] L. R. Medsker and L. Jain, "Recurrent neural networks," *Design and Applications*, vol. 5, no. 64-67, p. 2, 2001.

[73] Q. Wang, B. Li, T. Xiao, J. Zhu, C. Li, D. F. Wong, and L. S. Chao, "Learning deep transformer models for machine translation," in *Annual Metting of the Association for Computational Linguistics*, 2019, pp. 1810–1822.

[74] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 6706–6713.

[75] Y. Zhang, J. Hu, Y. Zhang, B. Pardo, and Z. Duan, "Vroom!: A search engine for sounds by vocal imitation queries," in *Proc. Conference on Human Information Interaction and Retrieval* (*CHIIR*), 2020, pp. 23–32.

[76] Y. Zhang and Z. Duan, "IMISOUND: An unsupervised system for sound query by vocal imitation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2016, pp. 2269–2273.

[77] S. Ikawa and K. Kashino, "Acoustic event search with an onomatopoeic query: measuring distance between onomatopoeic words and sounds," in *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events* (*DCASE*), 2018, pp. 59–63.

[78] G. M.-Brocal and G. Peeters, "Conditioned-U-Net: Introducing a control control mechanism in the U-Net for multiple source separations," in *Proc. International Society for Music Information Retrieval* (*ISMIR*), 2019, pp. 159–165.

[79] Y. Sudo, K. Itoyama, K. Nishida, and K. Nakadai, "Environmental sound segmentation utilizing Mask U-Net," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*), 2019, pp. 5340–5345.

[80] A. Jansson, E. J. Humphrey, N. Montecchio, R. M. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep U-Net convolutional networks," in *Proc. International Society for Music Information Retrieval* (*ISMIR*), 2017, pp. 745–751.

[81] Q. Kong, Y. Wang, X. Song, Y. Cao, W. Wang, and M. D. Plumbley, "Source separation with weakly labelled data: an approach to computational auditory scene analysis," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2020, pp. 101–105.

[82] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention* (*MICCAI*), 2015, pp. 234–241.

[83] T. Ochiai, M. Delcroix, Y. Koizumi, H. Ito, K. Kinoshita, and S. Araki, "Listen to what you want: Neural network-based universal sound selector," in *Proc. INTERSPEECH*, 2020, pp. 1441–1445.

[84] I. Kavalerov, S. Wisdom, H. Erdogan, B. Patton, K. Wilson, J. L. Roux, and J. R. Hershey, "Universal sound separation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), 2019, pp. 175–179.

[85] P. Smaragdis and G. J. Mysore, "Separation by "humming": User-guided sound extraction from monophonic mixtures," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), 2019, pp. 69–72.

[86] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[87] M. Delcroix, J. B. Vázquez, T. Ochiai, K. Kinoshita, Y. Ohishi, and S. Araki, "Soundbeam: Target sound extraction conditioned on sound-class labels and enrollment clues for increased performance and continuous learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 121–136, 2023.

[88] J. H. Lee, H.-S. Choi, and K. Lee, "Audio query-based music source separation," in *Proc. International Society for Music Information Retrieval* (*ISMIR*), 2019, pp. 878–885.

[89] E. Tzinis, S. Wisdom, J. R. Hershey, A. Jansen, and D. P. W. Ellis, "Improving universal sound separation using sound classification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2020, pp. 96–100.

[90] X. Liu, H. Liu, Q. Kong, X. Mei, J. Zhao, Q. Huang, M. D. Plumbley, and W. Wang, "Separate what you describe: Language-queried audio source separation," in *Proc. INTERSPEECH*, 2022, pp. 1801–1805.

[91] K. Kilgour, B. Gfeller, Q. Huang, A. Jansen, S. Wisdom, and M. Tagliasacchi, "Text-driven separation of arbitrary sounds," in *Proc. INTERSPEECH*, 2022, pp. 5403–5407.

[92] H.-W. Dong and N. Takahashi and Y. Mitsufuji and J. McAuley and T. Berg-Kirkpatrick, "Clipsep: Learning text-queried sound separation with noisy unlabeled videos," in *Proc. International Conference on Learning Representation* (*ICLR*), 2023.

[93] C. D. Kim, B. Kim, H. Lee, and G. Kim, "AudioCaps: Generating captions for audios in the wild," in *Proc. the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 2019, pp. 119–132.

[94] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: an audio captioning dataset," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2020, pp. 736–740.

[95] M. Wu, H. Dinkel, and K. Yu, "Audio caption: Listen and tell," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2019, pp. 830–834.

[96] R. Huang, J. Huang, D. Yang, Y. Ren, L. Liu, M. Li, Z. Ye, J. Liu, X. Yin, and Z. Zhao, "Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models," *arXiv preprint arXiv:2301.12661*, 2023.

[97] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[98] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1*, 2019, pp. 4171–4186.

本論文に関連する研究業績

学術論文

[J.1] **Y. Okamoto**, K. Imoto, S. Takamichi, R. Yamanishi, T. Fukumori, and Y. Yamashita, "Onoma-to-wave: Environmental Sound Synthesis from Onomatopoeic Words," APSIPA Transactions of Signal and Information Processing, Vol. 11, No. 1, e13, 2022.

国際会議

[C.1] **Y. Okamoto**, K. Imoto, S. Takamichi, R. Nagase, T. Fukumori, and Y. Yamashita, "Environmental Sound Synthesis from Vocal Imitations and Sound Event Labels," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024.（採録決定）

[C.2] **Y. Okamoto**, K. Imoto, S. Takamichi, T. Fukumori, and Y. Yamashita, "How Should We Evaluate Synthesized Environmental Sounds," Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 307-312, 2022.

[C.3] **Y. Okamoto**, K. Imoto, S. Takamichi, R. Yamanishi, T. Fukumori, and Y. Yamashita, "RWCP-SSD-Onomatopoeia: Onomatopoeic Word Dataset for Environmental Sound Synthesis," Proc. Detection and Classification of Acoustic Scenes and Events (DCASE), pp. 125-129, 2020.

[C.4] **Y. Okamoto**, K. Imoto, T. Komatsu, S. Takamich, T. Yagyu, R. Yamanishi and Y. Yamashita, "Overview of Tasks and Investigation of Subjective Evaluation Methods in Environmental Sound Synthesis and Conversion," Speech and Audio in the Northeast (SANE), 2019.

国内発表

[D.1] 岡本 悠希, 井本桂右, 高道 慎之介, 永瀬 亮太郎, 福森 隆寛, 山下 洋一, “環境音の模倣音声を用いた環境音合成の検討とデータセット構築, ” IDR ユーザフォーラム 2023，2023.

[D.2] 岡本 悠希，井本桂右，高道 慎之介，永瀬 亮太郎，福森 隆寛，山下 洋一，“Voice-to-foley: 環境音を模倣した音声を入力とする環境音合成, ” 日本音響学会 2023 年秋季研究発表会，pp. 1071–1074，2023.

[D.3] 岡本 悠希，井本桂右，高道 慎之介，福森 隆寛，山下 洋一，“環境音合成の入力情報に応じた主観評価手法の検討, ” 日本音響学会 2022 年秋季研究発表会，pp. 1257–1260，2022.

[D.4] 岡本 悠希，井本桂右，高道 慎之介，福森 隆寛，山下 洋一，“環境音合成における主観評価手法の検討, ” 日本音響学会 2022 年春季研究発表会，pp. 1071–1074，2022.

[D.5] 岡本 悠希，井本桂右，高道 慎之介，福森 隆寛，山下 洋一，“Transformer を用いたオノマトペからの環境音合成手法の提案, ” 日本音響学会 2021 年秋季研究発表会，pp. 943–946，2021.

[D.6] 岡本 悠希, 井本桂右, 高道 慎之介, 山西 良典, 福森 隆寛, 山下 洋一, “Onoma-to-Wave: オノマトペを利用した環境音合成手法の提案, ” 日本音響学会 2021 年春季研究発表会，pp. 843–846，2021.

[D.7] 岡本 悠希，井本 桂右，高道 慎之介，山西 良典，福森 隆寛，山下 洋一，“豊かな環境音の生成〜オノマトペを利用した環境音合成手法の提案〜, ” 日本音響学会 関西支部 第 23 回若手研究者交流研究発表会，2020.

[D.8] 岡本 悠希，井本 桂右，高道 慎之介，山西 良典，山下 洋一，“オノマトペを用いた環境音合成のためのデータセット構築とその分析, ” 日本音響学会 2020 年春季研究発表会，pp. 1099-1102，2020.

[D.9] 岡本 悠希，井本 桂右，小松 達也，高道 慎之介，柳生 拓巳，山西 良典，山下 洋一，“多様な環境音の合成をめざして〜環境音合成における評価方法の検討〜, ” 日本音響学会 関西支部 第 22 回若手研究者交流研究発表会，2019.

[D.10] 岡本 悠希，柳生 拓巳，井本桂右，小松 達也，高道 慎之介，山西 良典，山下 洋一，“多様な環境音の合成と変換のための基礎検討，” 日本音響学会 2019 年秋季研究発表会，pp. 1003-1006，2019.

Preprint

[A.1] **Y. Okamoto**, K. Imoto, T. Komatsu, S. Takamich, T. Yagyu, R. Yamanishi and Y. Yamashita, “Overview of Tasks and Investigation of Subjective Evaluation Methods in Environmental Sound Synthesis and Conversion”, arXiv preprint, arXiv: 1908.10055, 2019.

受賞

[H.1] 第 5 回 IEEE Signal Processing Society (SPS) Tokyo Joint Chapter Student Award, 2021.

[H.2] 日本音響学会 第 22 回学生優秀発表賞，2021.

その他の研究業績

学術論文

[J.1] 砺波 紀之，井本 桂右，岡本 悠希，福森 隆寛，山下 洋一，“誤検出の深刻さを考慮した音響イベント検出のための評価指標，” 日本音響学会誌，Vol. 78，No. 5，pp. 217–226，2022.

国際会議

[C.1] Keunwoo Choi, Jaekwon Im, Laurie Heller, Brian McFee, Keisuke Imoto, **Yuki Okamoto**, Mathieu Lagrange, and Shinosuke Takamichi, “Foley Sound Synthesis at the DCASE 2023 Challenge,” Proc. Detection and Classification of Acoustic Scenes and Events (DCASE), pp. 16–20, 2023.

[C.2] Shunsuke Tsubaki, Yohei Kawaguchi, Keisuke Imoto, Tomoya Nishida, Kota Dohi, Takashi Endo, and **Yuki Okamoto**, “Audio-Change Captioning to Explain Machine-Sound Anomalies,” Proc. Detection and Classification of Acoustic Scenes and Events (DCASE), pp. 201–205, 2023.

[C.3] **Yuki Okamoto**, Kanta Shimonishi, Keisuke Imoto, Kota Dohi, Shota Horiguchi, and Yohei Kawaguchi, “CAPTDURE: Captioned sound Dataset of Single Sources,” Proc. INTERSPEECH, pp. 1683-1687, 2023.

[C.4] Hien Ohnaka, Shinnosuke Takamichi, Keisuke Imoto, **Yuki Okamoto**, Kazuki Fujii and Hiroshi Saruwatari, “Visual Onoma-to-Wave: Environmental Sound Synthesis from Visual Onomatopoeias and Sound-Source Images,” Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5, 2023.

[C.5] **Yuki Okamoto**, Shota Horiguchi, Masaaki Yamamoto, Keisuke Imoto, and Yohei Kawaguchi, "Environmental Sound Extraction using Onomatopoeic Words," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 221–225, 2022.

[C.6] Noriyuki Tonami, Keisuke Imoto, Ryotaro Nagase, **Yuki Okamoto**, Takahiro Fukumori, and Yoichi Yamashita, "Sound Event Detection Guided by Semantic Contexts of Scenes," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 801-805, 2022.

[C.7] Noriyuki Tonami, Keisuke Imoto, **Yuki Okamoto**, Takahiro Fukumori, and Yoichi Yamashita, "Sound Event Detection Based on Curriculum Learning Considering Learning Difficulty of Events," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 875–879, 2021.

[C.8] **Yuki Okamoto**, Keisuke Imoto, Naoki Tsukahara, Kxen Nagata, and Koh Sueda, Ryosuke Yamanishi, and Yoichi Yamashita, "Crow Call Detection Using Gated Convolutional Recurrent Neural Network," Proc. RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP), pp. 171–174, 2020.

国内発表

[D.1] 岡本 悠希，井本 桂右，土肥 宏太，川口 洋平，"DCASECaps: 単一音源に説明文を付与した環境音データセット，" 日本音響学会 2023 年春季研究発表会，pp. 133-136，2023.

[D.2] 大中 緋慧，高道 慎之介，井本 桂右，岡本 悠希，藤井 一貴，猿渡 洋，Visual onoma-to-wave: 画像オノマトペと音源画像を利用した環境音合成の提案，電子情報通信学会 音声研究会，pp. 78-82，2023.

[D.3] 岡本 悠希，堀口 翔太，山本 正明，井本 桂右，川口 洋平，"擬音語を用いた環境音抽出，" 日本音響学会 2022 年春季研究発表会，pp. 247-250，2022.

[D.4] 砺波 紀之，井本 桂右，永瀬 亮太郎，岡本 悠希，福森 隆寛，山下 洋一，"事前定義されていないシーン情報を利用可能な音響イベント検出，" 日本音響学会 2022 年春季研究発表会，pp. 243-246，2022.

[D.5] 井本 桂右, 岡本 悠希, 高道 慎之介, 福森 隆寛, 山下 洋一, “RWCP 音声・音響データベースを用いた環境音・効果音合成の検討とオノマトペ拡張データセットの構築,” IDR ユーザフォーラム, 2021.

特許

[P.1] 井本桂右, 秋山大知, 岡本 悠希, 山西良典, 山下洋一, “音響モデル生成方法、音響分析方法、演算装置、及び、コンピュータプログラム”, 特願 2020-101291, (2020).

受賞

[H.1] 第 17 回 IEEE Signal Processing Society (SPS) Japan Student Conference Paper Award, 2024.