CE-008

未知環境音からの日本語擬音語自動生成手法の提案
Proposal of an Automatic Method for Generating Japanese Onomatopoeia from Unseen Environmental Sounds

藤吉　宙† 柊 和佑‡ 坂本　薫† 柳谷　啓子‡
Fujiyoshi Sora Hiiragi Wasuke Sakamoto Sumire Yanagiya Keiko

1. はじめに

近年、発話以外の「環境音」に対する認識や記述の重要性が高まりを見せている。ロボティクスや AR/VR、聴覚支援技術、スマートデバイスにおいて、音環境を理解し応答することが求められるようになっており、それに伴って環境音の自動認識や分類の研究も進展している。

とりわけ、近年では市民によるスマートフォンやアクションカメラなどを用いた無構成記録(Unstructured recordings)や、日常生活の中で偶発的に音を含む映像・音声を収録した非意図的環境記録(Incidental environmental capture）が急増している。これらのデータは、記録者の意図を超えた形で多様な環境音を含んでおり、従来の整備された録音・録画データとは異なる情報的価値を有する。にもかかわらず、それらに含まれる環境音は体系的に記述・検索される手段が乏しく、アーカイブ的利活用や意味抽出が困難である。

こうした環境音の記述手段として注目されているのが、日本語の擬音語（オノマトペ）である。オノマトペは単なる音の模倣にとどまらず、音の印象・質感・身体感覚までを象徴的に言語化する力を持っており、直感的な理解を促す表現手段として、創作・教育・医療・工学など多様な分野で活用されている。

従来の音響処理では、環境音を事前に定義されたラベル(例：dog_bark, thunder）に分類する「ラベリング」が主流であり、未知の音や印象的表現に対する柔軟性が乏しい。また、擬音語生成に関する研究の多くは、特定語の模倣や主観ラベルの収集に依存しており、音から新たな言語的記述を導出する手法はほとんど存在していない。

本研究では、未知環境音を入力とし、日本語モーラ列を出力するという新たな記号化枠組みを提案する。システムは、WebRTC VAD による発話検出・除去、FFT による雑音低減、YAMNet を用いた音響特徴抽出、そしてファインチューニング済みのモーラ分類器による Top-5 出力から構成されており、従来手法では難しかった未知音への対応と言語的・印象的表現の創出を可能にする。

†中部大学大学院 国際人間学研究科 言語文化専攻
‡中部大学人文学部

このような「音→モーラ列」という生成的アプローチは、音環境理解に新たな表現力をもたらすとともに、無構成記録や非意図的環境記録に内包される多様な音の検索性や意味付与を可能にする記述手段として、今後の音響記号化や感性インタフェースの基盤となると考えている。

2. 関連研究

本研究が対象とする「環境音に対する擬音語的表現の生成」は、音響分類・音響合成・記号化の複数の研究領域にまたがるテーマである。本章では、主に 3 つの観点から関連研究を整理し、本研究の新規性と位置づけを明らかにする。

2.1 環境音と擬音語対応の分析研究

環境音に対する擬音語的記号化の研究は、音響分類・記号化・合成といった複数の領域にまたがって展開されてきた。たとえば、RWCP-SSD-Onomatopoeia は、環境音に対する日本語擬音語の主観的対応関係を記述したデータセットであり、聴覚印象と語彙的記号との関係性を分析する上で有用である[1]。一方、SoundBeam などのターゲット音抽出技術は、自然言語で指定された語句（例：「猫の鳴き声」）に対応する音を混合音源から抽出するもので、音とテキストのクロスモーダルな対応を実現している[2] [3]。また、擬音語を条件とする音響合成技術では、U-Net ベースのモデルにより擬音語から対応する効果音を生成する手法が提案されている [4] [5]。

2.2 本研究の位置づけ

以上のように、既存研究はいずれも「音→クラス」「記号→音」「記号→抽出」といった一方向的な枠組みに留まっており、「未知環境音から新たなモーラ列を記号的に生成する」という本研究の枠組みは未踏である。特に、本研究では音韻的単位（モーラ）を出力単位とすることで、既存語彙に依存せず、柔軟かつ創造的な擬音語表現を実現できる点が特徴である。これは、環境音の分類・記号化・創作応用といった複数の文脈において高い表現力と適応力を備えたアプローチであり、音響処理研究における新たな可能性を提示するものである。

3. 提案手法

本章では、本研究で構築した「未知環境音から日本語モーラ列を自動生成する」システムの構成と各処理モジュールについて詳細に述べる。本手法は、入力音声に対して段階的に前処理・特徴抽出・分類・生成の各工程を行い、最終的に人間が直感的に理解可能な短モーラ列（オノマトペ的文字列）を出力する。以下では、それぞれの処理ブロックについて順を追って解説する。(図 1)

3.1 音声・環境音の自動分離

まず、入力音声データ（WAV ファイル）に対し、WebRTC に基づく VAD（Voice Activity Detection）を適用し、人間の発話とそれ以外の環境音をフレーム単位で分離する。WebRTC VAD は、エネルギー分布、スペクトル特徴に基づき、発話か否かを確率的に判定する軽量なモデルであり、リアルタイム処理との親和性が高い。本研究では 30ms フレーム単位を選択し、発話セグメントと非発話セグメントを二分する。

このステップの目的は、以降のモーラ分類において「人間の発話に含まれる実語音」が擬音語として誤生成されることを防ぐためである。また、音源が混在したままでは分類精度が著しく低下するため、事前に「音の意味」にかかわらず発話というカテゴリを除去することは、本システムの汎用性を確保するうえで重要である。

3.2 FFT 帯域マスキングによる雑音除去

分離された環境音セグメントに対しては、さらに帯域特性に基づくノイズ除去処理を施す。具体的には、時間領域信号をフーリエ変換し、スペクトルのエネルギー分布に対し閾値処理を行うことで、環境ノイズや機器由来の背景雑音を抑圧する。再度逆変換（IFFT）を行うことで、クリアな周波数特性を持つ環境音信号が得られる。このマスキング処理により、後段の分類器がピュアな環境音特徴に基づいて推論できるようになる。特に、雨音や風音、こすれる音のように高周波雑音が重なる音源に対しては、この処理が有効に働くことが確認されている。

3.3 音響埋め込み抽出（YAMNet）

前処理済み音声に対しては、Google が公開している音響分類モデル「YAMNet」を用いて、音響特徴ベクトルを抽出する。YAMNet は AudioSet（20 万件以上のラベル付き音響イベント）で事前学習された MobileNet ベースの軽量分類器であり、入力波形から 1024 次元のフレーム埋め込みを抽出可能である[6]。また、これに類する構成で、時–周波数注意機構付き CNN を用いた高精度環境音分類も報告されている[7]。本研究では、一定時間（約 5 秒）の音声を対象とし、各フレームごとの埋め込みを以下の統計量で集約することで、3072 次元の固定長ベクトルを得る：

・平均値(mean)：音全体の基礎的傾向を捉える
・標準偏差(std)：変動性・リズムの有無を示唆
・最大値(max)：瞬間的な強音や鋭い音の検出

この「平均・分散・最大値」の統計的三層埋め込みは、単純な平均プーリングと比較して情報密度が高く、分類精度の向上が確認されている。

3.4 モーラ分類器によるモーラ系列生成

得られた 3072 次元ベクトルを入力として、事前にファインチューニングした日本語モーラ分類器によりモーラ列を生成する。本分類器は、Keras/TensorFlow ベースで実装した全結合層＋ソフトマックス出力を持つシンプルなマルチクラス分類モデルである。分類対象は、日本語の五十音に加え、濁音・半濁音・拗音を含む約 106 クラス（例：a, i, u, e, o, ka, ga, kya, pyo 等）で構成されている。

ファインチューニングには、筆者が独自収集・録音した明瞭な 1 音単位の日本語単語音声（1 秒単位）を用いた。各クラスには 100 の音声サンプルを用意し、YAMNet ベースの音響埋め込みと対応付けて教師あり分類を行っている。こうすることで、YAMNet が持つ「音響的印象」に基づいたベクトル空間から、直感的に分かりやすい日本語モーラ列へと変換する分類器を実現している。推論時には、ソフトマックス出力の上位 5 件を取り出し、それぞれのスコアとともに提示する。この

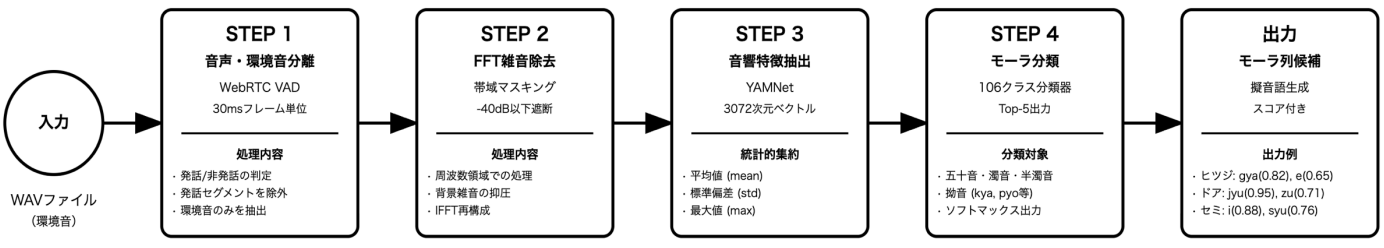

図 1. 未知環境音からの日本語擬音語自動生成システム

「Top-5 表示」は、擬音語という性質上、唯一解ではなく多義的・印象的な選択が望まれる場面において、表現の多様性と柔軟性を確保する効果がある。

4. 実験と考察

本章では、提案手法に基づいて構築したシステムを用いた実験とその結果について報告する。対象としたのは多様な実環境下で収録された音声クリップであり、それらに対して音声・環境音の分離、前処理、音響埋め込み抽出、モーラ分類という一連のプロセスを適用し、出力された日本語モーラ列の内容を分析した。

4.1 実験設定

評価対象とした音源は、筆者が独自に収録・編集した室内外の環境音クリップ 100 件である。音源は、動物の鳴き声（猫、ヤギなど）、生活音（ドアの開閉、電動工具等）、自然音（風、雨、雷）など幅広くカバーし、モノラル WAV ファイル（16kHz）に統一した。各クリップに対して、以下の処理を順に実行する：

1.WebRTC VAD による人声／環境音の自動分離
2.FFT 帯域マスキングによる雑音低減
3.YAMNet による音響埋め込み(3072 次元)の抽出
4.ファインチューニング済み分類器によるモーラ列生成(Top-5 出力)

分類器の出力結果は、モーラごとのソフトマックス確率とともに記録し、音源ごとに出力された上位 5 件のモーラ列を「自動生成擬音語候補」として評価した。

4.2 出力例と観察結果

出力されたモーラ列は以下のような傾向を示した：

・ヒツジの鳴き声に対して：「gya」「e」「ha」「ne」「te」など
・ミンミンゼミの鳴き声に対して：「i」「syu」「ti」「ki」「jyo」など(図 2, 図 3)
・ドアが閉まる音に対して：「jyu」「zu」「byu」「syu」など (図 4, 図 5)

出力されたモーラ列には、音響印象に対して一定の整合性が見られた。たとえば、ヒツジの鳴き声に対して、「gya」「ne」は実際の「メェー」に近い響きを含み、柔らかく高音域の鳴き声として妥当な印象を持つ。ミンミンゼミの鳴き声では、「i」が最上位に出力されおり、これは連続的かつ高周波成分を持つセミの音響特性を反映したものである。「syu」「ti」などの摩擦音を伴うモーラも含まれており、「ミーンミーン」という聴感的構造の一部を捉えていると考えられる。また、ドアの閉まる音では「jyu」が非常に高いスコアで出力され、これは単なる衝撃音ではなく、閉まる過程で発生する軋みや摩擦音の要素を含んだ音響特

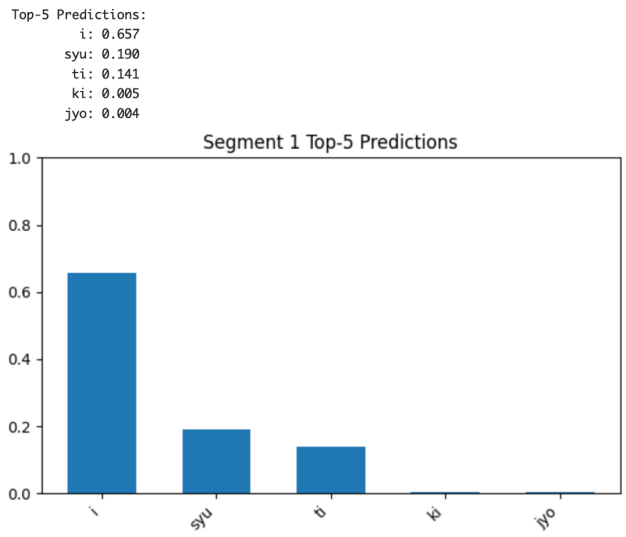

図 2. 「ミンミンゼミ」のモーラ列

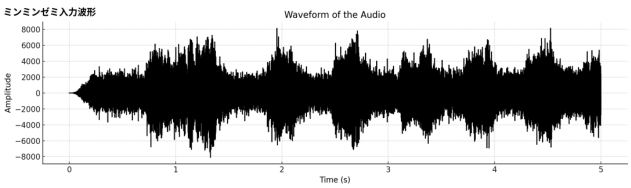

図 3. 「ミンミンゼミ」の波形

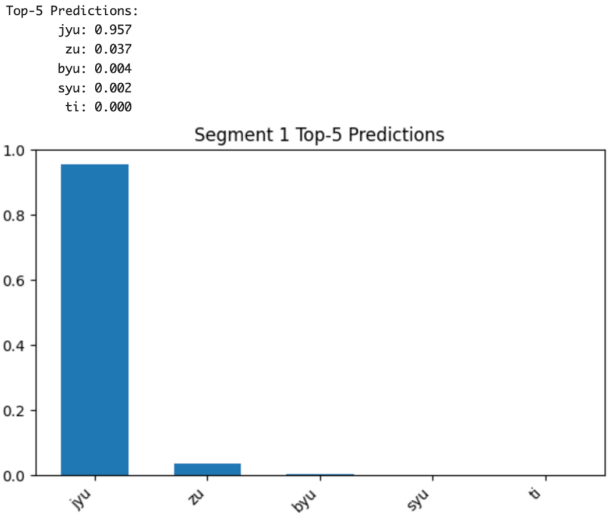

図 4. 「ドアが閉まる音」のモーラ列

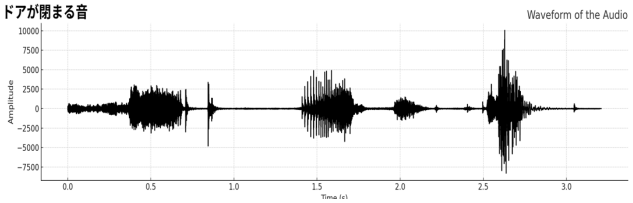

図 5. 「ドアが閉まる音」の波形

徴と音韻構造との対応を学習的に抽出していることを示すものである。生成されたモーラ列は、いずれも音響印象に基づく新たな記号化として機能している。本手法が実現する「音響印象→モーラ列」という抽象的変換は、従来の「音→ラベル」あるいは「擬音語→音響信号」といった固定的な処理とは異なり、未知音や非定型音に対しても柔軟な記述生成を可能とする点で、新たな音響記号化アプローチとして有効である。

生成されたモーラ列が正解モーラ列とどれだけ一致するかを評価指標として用いた。具体的には、1 モーラ単位での一致数を全体のモーラ数で割った平均トークン精度を算出している。各音声には 1 つの正解擬音語が割り当てられ、比較が行われた。現時点ではこの精度を暫定的な評価指標としており、今後は主観評価や多様性指標なども含めた評価基準の整備を検討中である。今回の実験では平均トークン精度 0.88 を達成した。

本研究における「平均トークン精度 0.88」は、評価用のテストセットに対して出力されたモーラ列のうち、実際の正解モーラと一致したモーラが全体の 88%にのぼることを示す。この指標は、分類結果の単純な一致率ではなく、モーラ単位での逐次的な正確性を評価するものである。

5. 結論と今後の課題

本研究では、未知環境音に対して日本語のモーラ列を自動生成する音響記号化手法を提案した。従来の音響分類や音響生成と異なり、本手法は「擬音語的な記号列」を出力することで、人間の直感的理解を支援する新たな枠組みを示した。語彙依存を排し、モーラ単位で音を記述することで、未知音や曖昧な音に対しても柔軟な表現が可能である。

本システムは、(1) WebRTC VAD による発話除去、(2) FFT による雑音低減、(3) YAMNet による音響特徴抽出、(4) モーラ分類器による Top-5 出力という段階的処理を通じ、高い汎用性と応用性を実現した。

実験では、出力されたモーラ列と実際の擬音語の印象的対応も観察され、分類器が音響印象を反映した妥当な出力を行っていることが確認された。課題としてはまず、評価が定量指標に偏っており、「使いやすさ」「伝わりやすさ」などの主観的妥当性を十分に検証できていない点がある。今後は被験者による主観評価（再現性・会話使用可能性・違和感の有無等）の導入が必要である。これにより、単なる数値精度だけでなく、「ユーザーにとって使いやすいか」「響きが自然か」といった実用上の妥当性も評価可能となる。

また、本手法は単一音源に特化しており、連続音・複合音に対しては出力の一貫性が低下する傾向がある。これに対しては、RNN や Transformer などの時系列モデルの導入が有効と考えられる。

以上より、本研究は「未知音の音響印象を日本語的記号として抽出する」という未踏の課題に対して、理論・実装・応用可能性の各面から新たな解を提示するものであり、音響処理および人間中心インタフェース設計の両領域に貢献する可能性を持つと考えられる。

また、本研究で提案した「音→モーラ列」の記号化枠組みは、従来のタグベースのラベリング手法では捉えきれなかった、未知音や印象的・主観的な音の記述を可能にするものである。この手法をデジタルアーカイブ分野に応用することで、映像資料や音声記録に含まれる環境音を、日本語の擬音語という直感的かつ検索可能な形で記述・付与することが可能となり、特に非統制的に収集された無構成映像や市民収録資料における音の意味的アクセシビリティが大きく向上する。これにより、記録映像の再解釈、感性的検索インタフェースの構築、文化的・生活的記録の言語的補完といった新たな活用の可能性が開かれ、音を軸としたデジタルアーカイブの記述・発見・再利用の実践に貢献すると考えている。また、音効果ライブラリの自動索引化に取り組む研究[8]とも親和性が高く、記述性・検索性の両立という課題に対する議論を補強する。

参考文献

[1] Okamoto Y., Imoto K., Takamichi S., Saruwatari H., “RWCP-SSD-Onomatopoeia: Onomatopoeic Word Dataset for Environmental Sound Synthesis,” *Proc. DCASE Workshop 2020*.

[2] Zhang Y., Kong Q., Delcroix M., et al., “SoundBeam: Target Sound Extraction Conditioned on Sound-Class Labels and Enrollment Clues for Increased Performance and Continuous Learning,” *IEEE/ACM Trans. Audio, Speech, and Language Processing*, (2022).

[3] Delcroix M., Bennasar Vázquez J., Ochiai T., et al., “Few-shot learning of new sound classes for target sound extraction,” *arXiv*, Jun 2021.

[4] Park H., Lee K., et al., “Learning to Generate Onomatopoeic Words for Audio Events,” *Findings of ACL 2020*.

[5] Okamoto Y., Horiguchi S., Yamamoto M., et al., “Environmental Sound Extraction Using Onomatopoeic Words,” *arXiv*, Dec 2021.

[6] Mu W., Yin B., Huang X., Xu J., Du Z., “Environmental Sound Classification using Temporal-Frequency Attention Based CNN,” *Scientific Reports*, vol. 11, art. 21552 (2021).

[7] Fang Z., Yin B., Du Z., Huang X., et al., “Fast Environmental Sound Classification based on Resource Adaptive Convolutional Neural Network,” (2022).

[8] Piczak K. J., “ESC: Dataset for Environmental Sound Classification,” *Proc. ACM Multimedia 2015*.