擬音的発話のニュアンスを反映するインタラクティブ効果音合成

滝沢 力[2,a)]　平井 重行[1,b)]

概要：アニメや，映画，ゲームなどでは，場面に応じた効果音が様々に使用される．それらの音響制作には専門的な知識・ノウハウや試行錯誤，既存の膨大な効果音ライブラリから選定する手間などが発生する．本研究では，人が効果音を口真似することが比較的容易であることに着目し，音の細かなニュアンスまでも反映した，効果音の口真似音声（擬音的発音）を基にしたインタラクティブな効果音合成技術を提案する．ここでは，発話表現のし易さや，多種多様なニュアンスが発音可能な爆発音を合成題材とし，データセットおよびモデルの構築を行った．具体的には，効果音および口真似音声それぞれのメルスペクトログラム画像を Transformer で学習し，メルスペクトログラムを推定する前段処理と，メルスペクトログラムから波形を合成する後段処理のニューラルボコーダとして iSTFTNet を用い、2 種類の深層学習モデルを組み合わせて合成する．本稿では，提案モデルと学習データセットについて述べ，口真似の音声入力からの高音質な効果音合成結果を報告する．

1. はじめに

アニメや，映画，ゲームなどでは，場面に応じて様々な効果音が使用され，それらを制作する必要がある．音響制作時には，適切な素材を膨大なデータベースから選定し，それらを加工編集していく作業や，フォーリーなどの手法を用いて音を一から録音して作り上げていくような労力を費やす作業が必要になる [1][2][3][4]．経験の浅い制作者や素人からすると，このような作業は大変困難である．生成系深層学習技術の発展により，イラスト，音楽などのメディア作品を誰でも容易に制作できるようになってきている [5][6][7] が，このような物音や効果音を対象とした合成技術の例は僅かである．そこで，本研究では，人が効果音を口真似することが比較的容易であることに着目し，音の細かいニュアンスが表現可能である擬音を用いた効果音合成技術を提案する．研究では，発話表現のし易さや，多種多様なニュアンスが考えられる爆発音を対象とし，データセット，モデルの構築を行った．提案モデルは，ニュアンスをある程度考慮できた爆発音合成が可能であることが確認できた．本稿では，データセットと提案モデルの詳細および，現状得られている結果について述べる．

1 京都産業大学情報理工学部
2 京都産業大学大学院先端情報学研究科
a) takiriki1216@gmail.com
b) hirai@cc.kyoto-su.ac.jp

2. 関連研究

2.1 音声から非音声のサウンド合成

2.1.1 オノマトペのテキストからの環境音合成

オノマトペは音の特徴を表現する手法として有効であり，音の多様性の制御が期待される．岡本らは，そのようなオノマトペ（擬音語）のテキストからの環境音合成手法として，エンコーダに Bi-LSTM[10]，デコーダに LSTM[9] を使用したモデル [8] と，エンコーダ・デコーダとして Transformer[11] を使用したモデル [12] を提案している．前者では，オノマトペに加え，環境音の種類を示す音響イベントラベルも学習に用いることで，同じオノマトペでも様々な音の種類の環境音合成を実現していた．後者では，LSTM などの再帰構造を有さないニューラルネットワークとして Transformer を利用することで，長いオノマトペから長期的な特徴を捉えた合成を実現していた．

これらの手法では，多種多様な環境音の合成を可能にしている一方，音の細かなニュアンス変化を考慮した合成には対応していない．

2.1.2 KROTOS REFORMER PRO

KROTOS 社は，リアルタイムで入力された音声や，その他のオーディオ信号を解析し，それに応じて新しい音を作り出すことが可能な，REFORMER PRO[13] をリリースした．REFORMER PRO では，最大 4 つのサウンドライブラリが選択でき，それらのブレンドおよびそれぞれのパ

ラメータの制御が可能である．また，入力音の強弱や，長さなどの特徴を基に効果音を作り出していくことができ，映像を見ながら音を演じることで，フォーリー [4] のような制作を可能にしている．

2.2 オノマトペ関連のオーディオデータセット

2.2.1 VOCAL IMITATION SET

Bongjun Kim らは，非言語的な音と，それを人が真似たボーカルイミテーションが含まれるデータセット（VOCAL IMITATION SET）を作成した [16].

VOCAL IMITATION SET[16] には，8 個のサウンドカテゴリ，それらをさらに細分化したサウンドクラスが 302 個設けられており，フリー素材から各クラス平均 10 個の音源を収集している．収集された各クラスの音源からボーカルイミテーション用の参照音源として高品質なものが一つ選択され，クラウドソーシングを用いて対応するボーカルイミテーションを収集している．音声信号処理の研究者が，収集された音声を評価していき，最終的に 5601 個のボーカルイミテーションを利用できるデータセットが公開されている．

2.2.2 RWCP-SSD-Onomatopoeia

岡本らは，2.1.1 小小節の手法に使用するデータセットとして，RWCP 実環境音声音響データベース（RWCP-SSD）[14] に含まれる 105 種類の環境音に対して，計 155,568 個のオノマトペの音素情報を付与したデータセット（RWCP-SSD-Onomatopoeia）を作成し，公開している [15].

3. 擬音的発話のニュアンスを反映する効果音合成手法

3.1 提案手法

音のニュアンスを含めて発話した音声（以下，擬音と省略する）は，オノマトペテキストなどのシンボル化した情報では表すことができない細かなニュアンスを含めることができ，このような発話は多くの人にとって比較的容易である．実際の音響制作現場でも，制作者間で音のイメージを共有する際にしばしば擬音が用いらることがあり，これらのことから，擬音を用いることで直感的に利用可能且つ音の細かいニュアンス制御が可能であると考えた．

図 1 に提案する擬音を用いた効果音合成手法の概略を示す．提案手法では，岡本らの提案モデル [12] と同様に系列変換モデルとして Transformer を採用する．

3.2 学習用データセット

系列変換に使用する Transformer を本タスク用に学習させるために，合成の対象となる効果音，それに対応する擬音のペアが必要になる．2.2.1 小小節で述べた VOCAL IMITATION SET では，様々なサウンドクラスとそれらに対応する擬音が用意されているが，擬音に対応する正解データがクラス毎に一つのみであり，音の品質も十分でない．また，2.2.2 小小節で述べた RWCP-SSD-Onomatopoeia では，環境音とオノマトペの音素（発音のシンボル列）のデータセットになっている．我々が合成したい音は，環境音のような現実に存在する音に限らず，非現実な音も含む様々な擬音表現を対象としており，かつ既存の言語的な発音に限らない細かなニュアンスをも扱いたいため，これらのデータセットを活用することはできなかった．そこで，本研究では効果音とそれに対して人が発話した擬音の音声とを対にしたデータセットを独自構築することとした．

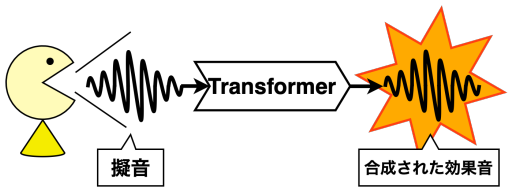

図 1 擬音音声から効果音の合成を行う手法の概略図

ここで，本研究においては，効果音のニュアンス表現を制御することに焦点をあてるべく，合成対象とする効果音の例を「爆発音」に絞ってまずはデータセットの構築と合成を試みることとした．爆発音は，数ある効果音の中でも，多種多様なバリエーションがあり，フォーリーとして演じてサウンド制作することが困難な音でもある．その一方で，人が口まねとして擬音的発話を行うことは比較的容易であり，様々なニュアンスとしても表現することができる．これらの理由から，爆発音は提案手法の検証に適していると考えた．

構築したデータセットの内容については，人が擬音的な発話をするための参照音（ターゲット音）となる爆発音をまずは収集した．ここでは，インターネット上で公開されているフリー素材（A.1）や，有料のデータセット（A.2）から約 1300 個収集した．この収集した各爆発音を参照音として聞きながら細かなニュアンス表現を含めて発話し，その音声を複数回ずつ録音した．これは同じ爆発音でも発音揺れが起こることを考慮し，参照音（爆発音）と擬音的発話音声のペアを増やしてデータ拡張を行うことを意図した．今回の学習用データセットに用いたデータ数は計 3,575 ペアとなる．録音は，全て同じマイク（A.3）を使用し，周囲のノイズが少ない防音室で収録を行った．全サウンドファイルは，サンプリング周波数 44100Hz で統一した．

3.3 提案モデル

図 2 に本研究で提案するモデルと処理フローを示す．処理の流れとしては，波形編集等を行う前処理，系列変換を行う Transformer，ニューラルボコーダを用いた波形合成処理の三つで構成される．

前処理では，中段の Transformer に入力する形式として，

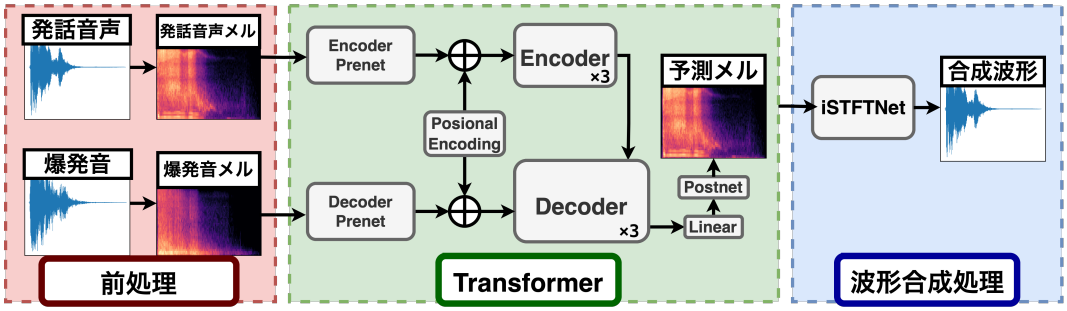

図 2 Transformer を用いた擬音からの爆発音合成モデル

表 1 メルスペクトログラムの仕様

FFT 窓幅	2048
移動幅	512
メル周波数軸の次元数	256

表 2 用意した未知の爆発音の擬音

入力名	ニュアンス
爆発音の擬音 A	爆風がフェードアウトしていくような擬音
爆発音の擬音 B	時間を空けて二回爆発しているような擬音
爆発音の擬音 C	炎がボワっと立ち上がるような擬音

二次行列（画像）である音響特徴量に変換する．本手法では，音声データを入力として用いるため，人間の聴覚特性に基づいた音響尺度を用いたメルスペクトログラムを音響特徴量とする．表 1 に，ここで扱う音響特徴量のパラメータを示す．

Transformer では，初めに Encoder/Decoder Prenet にて，音響特徴量の縦軸（周波数軸）を 256 次元から 512 次元へ拡張する．擬音と爆発音それぞれの音響特徴量を 3 層の Encoder/Decoder に入力する．Encoder にて，擬音の音響特徴量から特徴抽出を行い，Decoder にて，抽出された特徴と爆発音の音響特徴量を入力とし，新たな音響特徴量を生成していく．Decoder が生成する音響特徴量の縦軸（周波数軸）を元の 256 次元に射影し，Postnet にて音響特徴量の残差を予測し最終的な出力に加算する．これらにより，擬音の音響特徴量から爆発音の音響特徴量（画像から画像）への系列変換を学習することができる．

波形合成処理では，ニューラルボコーダとして iSTFTNet[17] を使用する．iSTFTNet は，メルスペクトログラムをアップサンプリングしていき，位相・振幅スペクトログラムを予測するニューラルボコーダである．予測された位相・振幅スペクトログラムを用いて逆短時間フーリエ変換を行うことで波形合成が可能になる．

4. 提案モデルの検証

4.1 学習

作成した学習用データセットを用いて提案モデルで使用する Transformer をエポック数 20000，バッチサイズ 64 で学習した．また，波形合成処理で用いる iSTFTNet には，作成したデータセットの内，爆発音のみを使用し，エポック数 10000，バッチサイズ 64 で学習した．それら学習には，NVIDIA DGX A100 環境（A.4.1）を用いた．

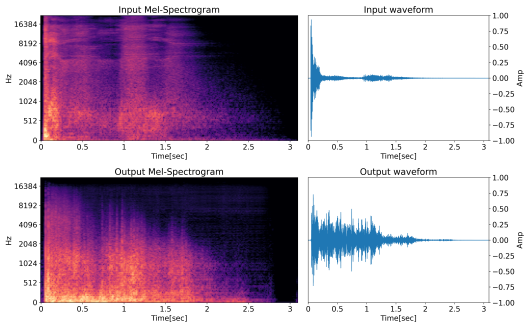

図 3 爆発音の擬音 A を入力とした合成結果
（上段：入力の擬音音声 / 下段：合成結果）

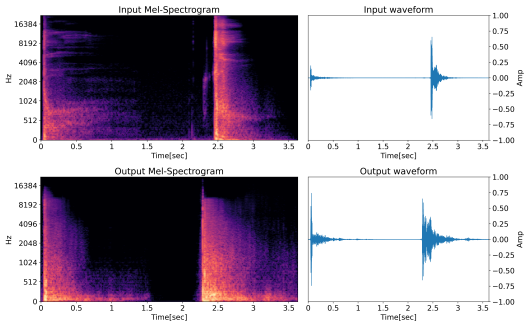

図 4 爆発音の擬音 B を入力とした合成結果
（上段：入力の擬音音声 / 下段：合成結果）

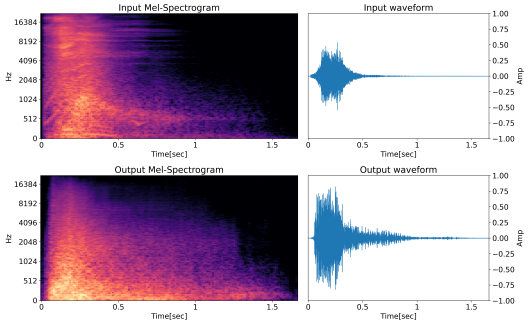

図 5 爆発音の擬音 C を入力とした合成結果
（上段：入力の擬音音声 / 下段：合成結果）

4.2 合成

表 2 に示す，未知の爆発音の擬音を入力とし，学習済みの Transformer，iSTFTNet を用いて爆発音合成を行った．モデルの生成（推論時の）処理では，GPU として GeForce RTX 3090（A.4.2）を一つ使用しており，擬音の発話長 + 数ミリ秒程度の処理時間で合成が可能である．

図 3〜5 に結果を示す．それぞれの合成結果で，入力のニュアンスを反映した特徴（時系列）の変化が見られ，音声と比較して低周波帯域の成分が強くなっていることから，爆発音らしい迫力のある音として合成されていることが確認できる．

5. 効果音制作としての使い方の狙い

5.1 提案モデルによる Human-In-The-Loop

4.2 小節で述べたように，現状の学習済みモデルでの生成処理に性能の高い GPU を用いることで，素早い効果音合成ができ，何度も擬音を入力し合成を行いながら試行錯誤が可能である．このことを踏まえて，図 6 に，想定とする，提案手法による効果音制作の概略図を示す．

制作手法は，望んでいる音を想像，それを擬音として発話，モデルに入力し合成という流れで進めていく．合成された音が所望の効果音と異なる場合は，再度先の手順を繰り返し行っていくことで，合成結果を望んでいる音に近づけていくことが可能になる．このように，擬音からインタラクティブに効果音合成が行える本手法において，ユーザが制作に介入して，繰り返し合成を試みる Human-In-The-Loop としての活用が有効であると考える．

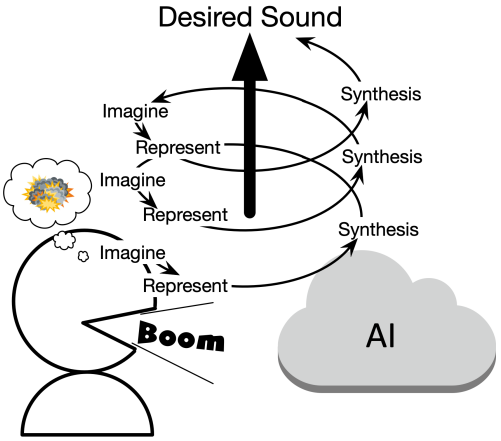

図 6 本手法における効果音制作（Human-In-The-Loop）

5.2 繰り返し合成の例

ここでは，図 6 の制作サイクルを実行し，合成音を目的の爆発音に近づけていくタスクを行う．

図 7 に，合成音の目標となる，データセットから選定した爆発音のメルスペクトログラムと波形を示す（なお，目標の爆発音は，ライフルで発砲したような音になっている）．

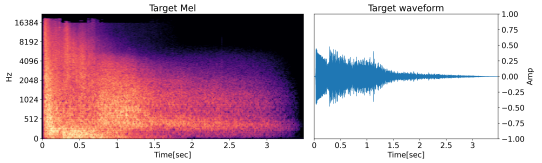

図 7 目標の爆発音

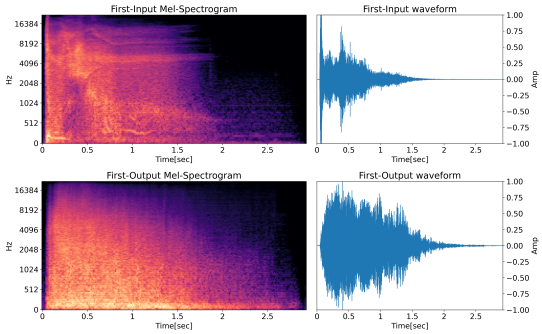

図 8 爆発音再現の 1 回目
（上段：入力の擬音音声 / 下段：合成結果）

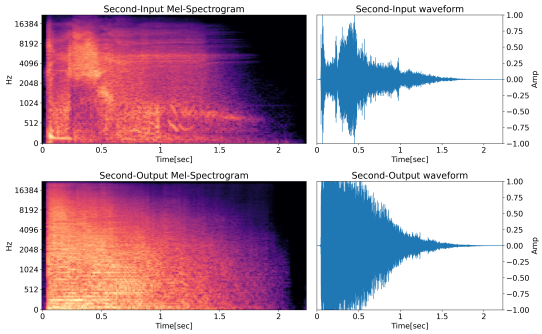

図 9 爆発音再現の 2 回目
（上段：入力の擬音音声 / 下段：合成結果）

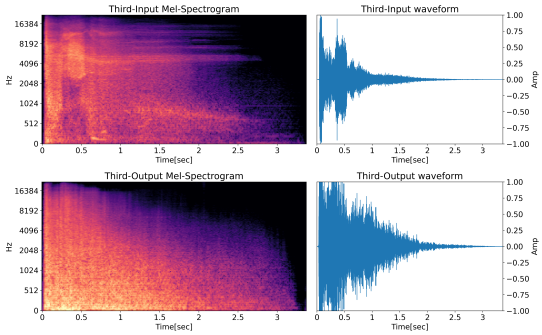

図 10 爆発音再現の 3 回目
（上段：入力の擬音音声 / 下段：合成結果）

一回目の試行では，目標の爆発音を聞こえたまま発話し，合成を行った．結果を図 8 に示し，合成音の波形から確認できるように，発砲音のようなアタックが無く，フェードインするような爆発音として合成されていた．図 9 に示す二回目の試行結果でも，一回目と同じようなニュアンスの擬音を入力した．一回目の結果と比較してアタックのある

音になっているが，波形が大きくクリップしており，品質の低い音が合成されていた．一回目と二回目では，アタックの音として「D」のような濁った発音で表現していたが，図 10 に示す三回目の試行では，それを「T」の発音にすることで，比較的高い周波数成分を持つ音として表現でき，より発砲音に近い音の合成ができると考えた．結果より，二回目の合成音と比較して，アタック部分が大きくクリップしておらず，アタック直後は減衰していくような音が合成できており，目標の爆発音に非常に近い音が得られた．

6. 考察

4.2 節で示したように，提案モデルによる爆発音合成では，ある程度ニュアンスを制御でき，比較的品質の高い合成結果が得られた．その一方で，発音とは明らかに異なるニュアンスの合成音となる場合もあった．5.2 節では，少ない試行回数で発話者が頭でイメージする爆発音に近い合成音が得られたものの，発音と異なるニュアンスの音が合成されると，イメージに近付けるための発話の微調整がうまくできなくなる可能性がある。これらが現時点での課題と言える．

これらの課題に対しては，データセットのバリエーションを増やすことで改善できると考えられる．現時点では，爆発音と発話音声のペアのデータセットをそのまま Transformer で学習したのみであり、自己教師有り学習による事前学習等は行っていない．発話音声を追加収録してデータセット数を増やすよりも，まずは事前学習とファインチューニングを行うことで，これらの課題の改善を試みる予定である．一方で，今回は，爆発音のみを対象として合成を試みたが，別の効果音でも合成を行うには，別途対象の効果音の収集，録音などのデータセット構築を行う必要がある．

7. おわりに

本研究では，誰でも容易に所望の効果音制作可能な手法として，擬音から効果音合成を行う技術を提案した．本稿では，様々なニュアンス，バリエーションが考えられる爆発音を合成の対象とし，データセット，モデルの構築を行った．提案モデルによる合成結果では，擬音のニュアンスがある程度反映され，比較的品質の高い成果が得られた．提案モデルを活用した Human-In-The-Loop としての効果音制作手法の検証では，少ない試行回数で目標の音に近い音が合成できた．

今後は，自己教師有り学習による事前学習とファインチューニングによる精度向上の試みに加え，複数話者でのデータセット構築や爆発音以外（ゲームなどで使用される魔法の音など）の効果音検証も行っていく予定である．

参考文献

[1] 木村哲人，＜キムラ式＞音の作り方，筑摩書房， 1999.
[2] 小川哲弘，サウンドエフェクトの作り方 [改訂版]， 工学社，2021.
[3] デイヴィッド・ゾンネンシャイン，Sound Design 映画を響かせる「音」のつくり方，フィルムアー ト社，2015.
[4] Vanessa Theme Ament. 2021. The Foley Grail. Routledge.
[5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser and Bj ¨ orn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10674-10685, doi: 10.1109/CVPR52688.2022.01042.
[6] David Holz. Midjourney. https://midjourney.com/home/?callbackUrl=%2Fapp%2F
[7] Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen and Zhou Zhao. 2022. DiffSinger：Singing Voice Synthesis via Shallow Diffusion Mechanism. arXiv. (Mar. 2022). https://arxiv.org/abs/2105.02446
[8] Okamoto. Y, Imoto. K, Takamichi. S, Yamanishi. R, Fukumori. T and Yamashita. Y. 2021. ONOMATO-WAVE：ENVIRONMENTAL SOUND SYNTHESIS FROM ONOMATOPOEIC WORDS. arXiv,（Feb. 2021）. DOI：10.48550/arxiv.2102.05872
[9] Sepp; Hochreiter and Jurgen Schmidhuber. Long Short-Term Memory. Neural coputation, 9(8):1735-1780, 1997.
[10] Graves, A., Fernández, S., Schmidhuber, J. (2005). Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition. In: Duch, W., Kacprzyk, J., Oja, E., Zadrożny, S. (eds) Artificial Neural Networks: Formal Models and Their Applications – ICANN 2005. ICANN 2005. Lecture Notes in Computer Science, vol 3697. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11550907_126
[11] A. Vaswani, et al.,“Attention is all You need,” Proc. NIPS, pp. 6000–6010, 2017.
[12] Okamoto. Y, Imoto. K, Takamichi. S, Fukumori. T and Yamashita. Y. 2021. Environmental sound synthesis from onomatopoeic words using Transformer model. Acoustical Society of Japan (ASJ). (Sep. 2021). ROMBUNNO. 3-3-1. https://jglobal.jst.go.jp/detail?JGLOBAL_ID=202102274740950003
[13] https://www.minet.jp/brand/krotos/reformer-pro/
[14] S. Nakamura, K. Hiyane, F. Asano, and T. Endo, ”Acousticalsound database in real environments for sound scene under-standing and hands-free speech recognition,” Proc. Language Resources and Evaluation Conference (LREC), pp. 965–968,2000.
[15] Yuki Okamoto, Keisuke Imoto, Shinnosuke Takamichi, Ryosuke Yamanishi, Takahiro Fukumori, and Yoichi 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.
[16] Bongjun Kim and Bryan Pardo, “Vocal Imitation Set v1.1.3 : Thousands of vocal imitations of hundreds of sounds from the AudioSet ontology”. Zenodo, Aug. 06, 2018. doi: 10.5281/zenodo.1340763.
[17] T. Kaneko, K. Tanaka, H. Kameoka and S. Seki,” ISTFTNET：Fast and Lightweight Mel-Spectrogram Vocoder Incorporating Inverse Short-Time Fourier Transform,” ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing（ICASSP）, Singapore, Singapore, 2022, pp. 6207-6211, doi：10.1109/ICASSP43922.2022.9746713.

付　　録

A.1 データセットの参照用音源に用いたフリーの効果音ソース

- E エフェクツ <https://esffects.net> (最終アクセス日:2021 年 12 月 13 日)
- On-Jin 音人 "" <https://on-jin.com/sound> (最終アクセス日:2021 年 12 月 27 日)
- クラゲ工匠 <http://www.kurage-kosho.info> (最終アクセス日:2021 年 12 月 27 日)
- 効果音ラボ <https://soundeffect-lab.info/sound> (最終アクセス日:2021 年 12 月 27 日)
- OtoLogic <https://otologic.jp/free/se> (最終アクセス日:2021 年 12 月 27 日)
- 効果音工房 <https://umipla.com/> (最終アクセス日:2021 年 12 月 27 日)
- 魔王魂 <https://maou.audio/> (最終アクセス日:2021 年 12 月 27 日)
- VSQ plus+ <https://vsq.co.jp/plus/> (最終アクセス日:2021 年 12 月 27 日)
- HURT RECORD<https://www.hurtrecord.com/> (最終アクセス日:2021 年 12 月 27 日)
- Sounds-mp3.com <https://sounds-mp3.com> (最終アクセス日:2021 年 12 月 27 日)
- Free Sound Dataset <https://annotator.freesound.org/fsd/explore/>（最終アクセス日:2022 年 8 月 18 日）
- Royalty Free Sound Effects Archive <https://sonniss.com/gameaudiogdc>（最終アクセス日:2022 年 8 月 16 日）
- Pixabay https://pixabay.com/sound-effects/search/explosion/ （最終アクセス日:2023 年 5 月 10 日）
- taira-komori https://taira-komori.jpn.org/arms01.html （最終アクセス日:2023 年 5 月 24 日）
- ザ・マッチメイカァズ http://osabisi.sakura.ne.jp/m2/#top （最終アクセス日:2023 年 5 月 24 日）
- TAM Music Factory https://www.tam-music.com （最終アクセス日:2023 年 5 月 24 日）
- ユーフルカ https://youfulca.com （最終アクセス日:2023 年 5 月 24 日）
- Pocket Sound https://pocket-se.info （最終アクセス日:2023 年 5 月 24 日）
- ヤミーズカフェ http://yamicafe.nekonikoban.org/index.html（最終アクセス日:2023 年 5 月 24 日）
- SoundJay https://www.soundjay.com（最終アクセス日:2023 年 5 月 24 日）
- Adobe https://www.adobe.com/products/audition/offers/AdobeAuditionDLCSFX.html （最終アクセス日:2023 年 5 月 24 日）
- 効果音辞典 https://sounddictionary.info/battle-1/ （最終アクセス日:2023 年 6 月 14 日）
- Mixkit https://mixkit.co/free-sound-effects/warfare/ （最終アクセス日:2023 年 7 月 3 日）
- Fesliyanstudios https://www.fesliyanstudios.com （最終アクセス日:2023 年 7 月 3 日）

A.2 データセットの参照用音源に用いた有料の効果音ソース

- SONICWIRE "爆発・災害に関するサウンドを中心に収めた効果音パック" <https://sonicwire.com/product/A9582>(最終アクセス日:2022 年 1 月 5 日)
- Pro Sound Effects "Anime Sound Effects Library"<https://shop.prosoundeffects.com/products/anime>（最終アクセス日:2022 年 9 月 6 日）

A.3 使用マイク

ZOOM Handy Recorder H4n
https://www.zoom.co.jp/ja/products/handy-recorder/h4nsp-handy-recorder

A.4 使用 GPU

A.4.1 学習時

NVIDIA DGX A100 640GB モデル
https://www.nvidia.com/ja-jp/data-center/dgx-a100/

A.4.2 合成時

NIVDIA GeForce RTX 3090
https://www.nvidia.com/ja-jp/geforce/graphics-cards/30-series/rtx-3090-3090ti/