音韻と韻律を含めたオノマトペ音声からのTransformerによる爆発音合成

滝沢 力 * 平井 重行 †

概要. メディア作品における音響制作では，クリエイターが知見・技術・経験を基に，場面や演出に適した効果音を制作していく．経験の浅いクリエイターや素人が音響制作を行うことは非常に困難であるが，求めている音を想像し，それらをオノマトペ（擬音語）として発話表現することは，比較的容易である．そこで，本研究では音のニュアンスを含めた発話音声から効果音を合成することで，容易に所望の効果音を作り出す手法について研究を行っている．ここでは，数ある効果音の中でも，多種多様なニュアンス表現が可能である爆発音に焦点を当て，深層学習モデルによる発話音声から爆発音への変換を学習し，爆発音合成を試みた．本稿では，Transformer を用いた合成手法と学習に用いたデータセット，および現状で得られている合成音について述べる．

1 はじめに

アニメや映画，ゲーム等のメディア作品では，場面に応じた様々な効果音が使用される．メディア制作現場では，プロのサウンドエンジニアがサウンドデータセットから素材を選定したり，素材そのものの録音も行う．そして，それらの素材に対して適した編集・加工などの作業を行い，最終的に使用される効果音へと仕上げていく [1][2][3]．これらの工程は，音響制作の知識や経験のある人が行っていくものであり経験の浅い人には困難である．しかし，求める音を想像・オノマトペとして発話し，音韻や韻律による微妙なニュアンスを含めて表現を行うことは，多くの人にとって容易である．そこで，本研究では，それらニュアンス表現されたオノマトペ発話音声から効果音合成を自在に行う手法の確立を目指す．そして，発話と合成をインタラクティブに繰り返し，所望の効果音素材を入手する環境の提供を目的とする．ここでは具体的な効果音として，爆発音を題材として取り組む．

2 関連研究

岡本らによるOnoma-to-Wave[4]では，テキスト音声合成技術による効果音合成を試みた．LSTMで構成された系列変換モデルにより，音の音素を表すオノマトペテキスト（音素列）から，対応する効果音への変換を行い，オノマトペテキストに加え，効果音の種類を表す音響イベントラベルを付与して学習させることで，多種多様な効果音の合成を可能にした．しかし，使用している系列変換モデルは，再帰構造を利用しており，長い系列の特徴を捉えた学習が困難であった．そこで，使用している系列変換モデルを，RNNやLSTMなどを使用せず系列の長期的な特徴を捉えることが可能であるTransformer[5]に置き換え，上記の改善を試みている [6]．

上記のモデル [4][6] では，音のバリエーションを考慮した合成を実現している一方で，音の韻律などのニュアンスを考慮した合成には特化していない．

3 提案手法

3.1 手法

本研究では，系列の長期的な依存関係を捉えることができ，音声以外の音の合成が可能であるTransformerを用いて，音韻・韻律等のニュアンス情報を含んだオノマトペ音声から効果音への変換を学習させる．また，変換の対象として，多種多様なニュアンス表現が可能である爆発音に焦点を当てており，細かいニュアンスの違いを表現してデータセットを作成する．提案モデルを学習させた学習済みモデルを使用することで，発話音声からの爆発音合成をその場でインタラクティブに行うことができる．

3.2 データセット

モデル学習用のデータセットは，爆発音とそれを口頭で擬音語として発話表現した音声とで対を成す形で構成する．まず，爆発音素材として，インターネット上で公開されている効果音データ [7]〜[18] や，販売されている効果音集[1] [2] から音源を収集した．これら全てを聞きながら発話表現して，録音した．ま

* 京都産業大学大学院
† 京都産業大学

[1] 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)

[2] 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）

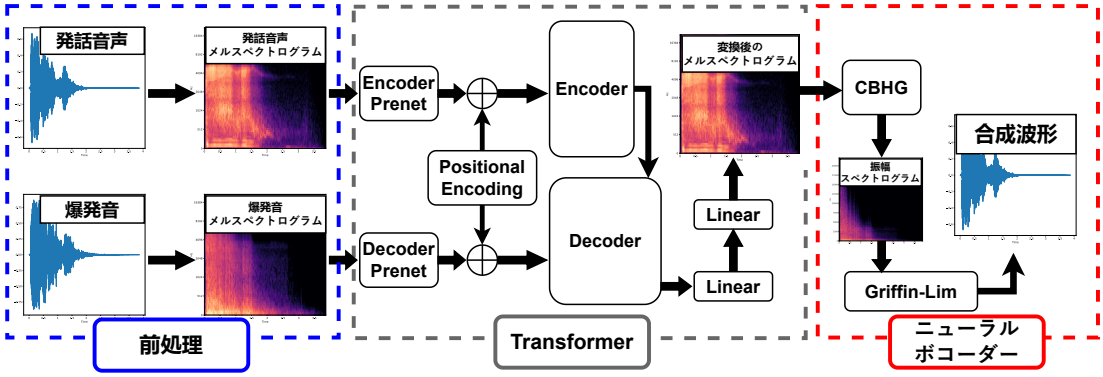

図 1. 提案モデルの処理フロー

た，一部の音源に対し複数回録音を行った．これにより，3575 対のデータを用意した．これらはすべて 44.1kHz・16bit サンプリングのオーディオデータとして学習に用いた．

3.3 提案モデル

提案モデルは，図 1 に示すように，前処理・Transformer・ニューラルボコーダーの 3 つの処理に分けられる．前処理では，波形データの後ろ十分の一の箇所を振幅のフェードアウト処理を施し，メルスペクトログラムの画像データに変換する．

Transformer では，発話音声のメルスペクトログラム，爆発音のメルスペクトログラムそれぞれをエンコーダ・デコーダに入力し，デコーダが出力するメルスペクトログラムの残差を Postnet で算出していくことで，変換を学習する．

ニューラルボコーダーでは，CBHG[19] を用いて，メルスペクトログラムから振幅スペクトログラムへの変換を学習する．合成時の Transformer の出力はメルスペクトログラム画像であるため，それを CBHG にて振幅スペクトログラムに変換し，それを Griffin-Lim アルゴリズム [20] によって合成波形を得る．

4 学習・合成

3 章のモデル・データセットを用いて，Transformer，ニューラルボコーダーをそれぞれエポック数 4000，2000，バッチサイズを 16 の設定で学習させた．学習済みモデルに，未知データとしてロケット花火のような音，破裂音のオノマトペ音声を入力した結果を図 2，3 に示す．これら図のそれぞれは，上段に入力に用いたオノマトペ音声，下段が合成結果の爆発音として示している．これらはニュアンスを含めて上手く合成できている．一方で、連続的に複数回の爆発が起こるような音については、適切な合成音が得られないことも確認できている．これは，そのような学習データが少ないことに起因するため，該当するデータを追加することで対応可能と考えている．

5 おわりに

本研究では，その場ですぐに合成を試すことができ，様々なニュアンスの爆発音を合成できるようなモデルの確立を目指し，その実装に取り組んだ．そして，音韻と韻律によるニュアンスを反映した合成が可能であることを確認した．うまく合成できない場合もあるため，引き続きデータセットの拡充や学習手法の改良などに努めていく．

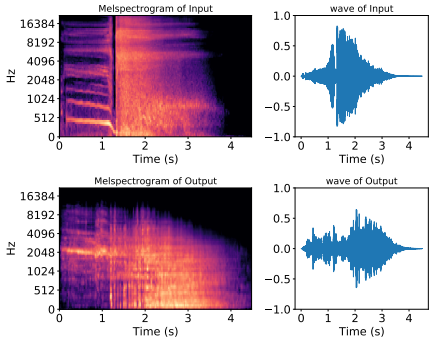

図 2. ロケット花火のような音の合成例
（上段：入力音声，下段：合成された効果音）

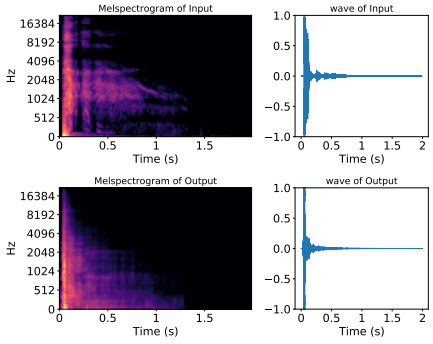

図 3. 破裂音の合成例
（上段：入力音声，下段：合成された効果音）

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] Y. Okamoto, K. Imoto, S. Takamichi, R. Yamanishi, T. Fukumori, Y. Yamashita, ”ONOMA-TO-WAVE: ENVIRONMENTAL SOUND SYNTHESIS FROM ONOMATOPOEIC WORDS,” arXiv: 2102.05872v3, 20 Oct 2021

[5] A. Vaswani, et al., ”Attention is all You need,” Proc. NIPS, pp. 6000-6010, 2017.

[6] 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,” 日本音響学会講演論文集 P.943-946, 2021 年 9 月

[7] E エフェクツ <https://esffects.net> (最終アクセス日:2021 年 12 月 13 日)

[8] On-Jin 音人 ”” <https://on-jin.com/sound> (最終アクセス日:2021 年 12 月 27 日)

[9] クラゲ工匠 <http://www.kurage-kosho.info> (最終アクセス日:2021 年 12 月 27 日)

[10] 効果音ラボ <https://soundeffect-lab.info/sound> (最終アクセス日:2021 年 12 月 27 日)

[11] OtoLogic <https://otologic.jp/free/se> (最終アクセス日:2021 年 12 月 27 日)

[12] 効果音工房 <https://umipla.com/> (最終アクセス日:2021 年 12 月 27 日)

[13] 魔王魂 <https://maou.audio/> (最終アクセス日:2021 年 12 月 27 日)

[14] VSQ plus+ <https://vsq.co.jp/plus/> (最終アクセス日:2021 年 12 月 27 日)

[15] HURT RECORD<https://www.hurtrecord.com/> (最終アクセス日:2021 年 12 月 27 日)

[16] Sounds-mp3.com <https://sounds-mp3.com> (最終アクセス日:2021 年 12 月 27 日)

[17] Free Sound Dataset <https://annotator.freesound.org/fsd/explore/>（最終アクセス日:2022 年 8 月 18 日）

[18] Royalty Free Sound Effects Archive <https://sonniss.com/gameaudiogdc>（最終アクセス日:2022 年 8 月 16 日）

[19] Y. Wang, R. Skerry-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”, in Proc. Interspeech, 2017, pp. 4006-4010.

[20] D. W. Griffin and J. S. Lim, ”Signal estimation from modified short-time Fourier transform,” IEEE Transactions on Acoustices, Speech and Signal Processing, pp. 236-243, 1984.

未来ビジョン

ゲームやアニメ、映画等のメディア作品の素材を容易く制作できるツールや環境が整ってきている．画像やイラスト作成では，MidJourney や Stable Diffusion 等の画像生成系 AI の登場により，あるレベルのコンテンツは容易かつ大量に生成を試せる時代になった．音楽・音響系制作でも，人の音声や歌声，音楽の AI 生成技術の発展で，ある程度合成ができるようになってきたが，効果音についてはほとんど扱われていない．効果音は映像作品やゲーム等でシーンの演出や臨場感表現に欠かせないが，シンセサイザを駆使して試行錯誤したり，大量の効果音データベースから選定したりと，望む音の入手には時間がかかる．代替音で効果音作成する手法もあるが，制作のアイデアや経験の要求、フォーリースタジオ環境など容易でない事情もある．こういった中で，様々なニュアンスを表現し，その場でオノマトペとしての発音を試行錯誤し，様々な効果音素材を合成して試せると，作業の大幅な効率化が見込まれる．これは，制作経験の浅い人でも様々な工程を飛ばして所望の効果音が得られ，またプロのクリエイターでもまずは自分で発話してイメージに近い音素材を得ることができることになる．高品質で本当にイメージ通りの音素材が得られるようになるにはデータセットの拡充は不可欠で，それは既存データの活用ができるものではないため，容易ではない．しかし，技術が確立してデータさえ整えば，人間のオノマトペ表現の能力が最大限に活用できるようになり，ひいてはそれが口頭での音表現に関する人間拡張技術になるとも言える．そして，それが新たなコンテンツ制作に繋がったり，従来にないアイデアの効果音演出やメディア表現にも繋がり，コンテンツ制作の新たなムーブメントが起こることすら期待できる．