P23

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

岡本 悠希[1]，井本 桂右[2]，高道 慎之介[3]，永瀬 亮太郎[1]，福森 隆寛[1]，山下 洋一[1]

[1] 立命館大学，[2] 同志社大学，[3] 東京大学

① 研究概要

環境音合成とは

■ 音声や楽音に限らないあらゆる音を人工的に合成する技術

・雨の音，人の足音，動物の鳴き声 . . .

■ 環境音合成の応用例：

・映画やゲームなどのメディアコンテンツの制作補助

環境音を模倣した音声を用いた環境音合成を提案

音声を入力に用いることで合成音の音高とリズムを制御可能に！！

② 従来研究の課題 / 本研究の方針

環境音合成の従来研究

■ 様々な入力情報からの環境音合成が提案されている

入力情報	リズム	音の高さ	時間的な変化	音源の種類
環境音ラベル [1]				○
画像 [2]				○
オノマトペ [3]			○	
音の説明文 [4]			○	○
音声（本研究）	○	○	○	

課題：従来の入力情報では音のリズムや高さを表現できていない

本研究の方針

■ 環境音を模倣した音声（模倣音声）のデータセット構築

■ 模倣音声を入力情報として利用した環境音合成の提案

・**模倣音声を利用することで合成音のリズムと音の高さを制御**

③ 環境音の模倣音声データセットの構築

使用した環境音データセット (ESC-50 [5]) の概要

■ 50 種類の環境音が収録

本研究では雑音などの少ない 31 種類の環境音に対して模倣音声を収録

・**動物の鳴き声：**犬，おんどり，豚，牛，猫，カエル，めんどり，羊

・**自然音 & 水の音：**水滴，火がパチパチ，水を注ぐ

・**人間の動作に関する音：**拍手，足音，歯を磨く

・**室内：**木のドアを叩く，マウスのクリック，キーボードのタイピング，木のドアが軋む
缶を開ける，掃除機，目覚まし時計，時計の秒針，ガラスが割れる

・**屋外：**チェーンソー，サイレン，車の警笛，教会の鐘，飛行機，花火，ノコギリ，エンジン

環境音に対する模倣音声の収録

■ 収録場所：立命館大学内のスタジオ（残響時間：0.2 秒以下）

■ 模倣音声の収録方法：

・手順 1：発話者に環境音をヘッドホンで聴取してもらう

・環境音は何度でも聴取可能

・手順 2：聴取した環境音を声で模倣してもらう

・音声の撮り直しは発話者が納得するまで何度でも可能

データセット

構築した環境音の模倣音声データセットの概要

■ 収録発話数：9,920 発話 (1,240 音 × 8 名)

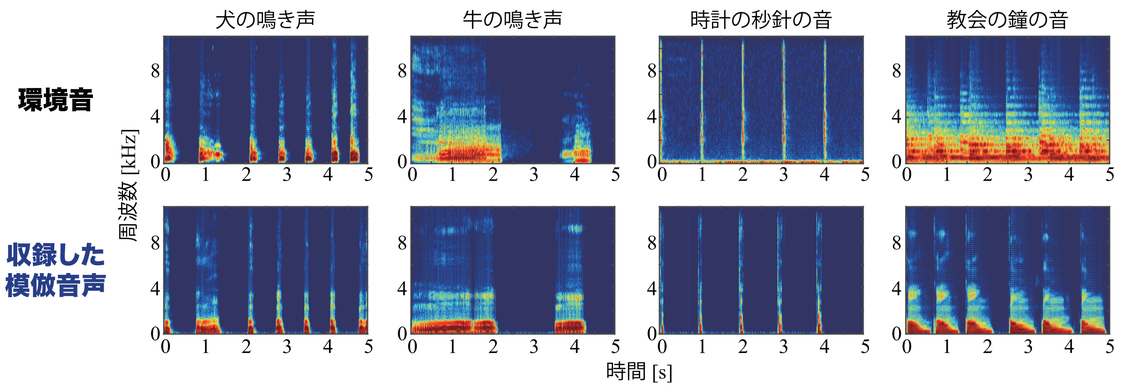
・1,240 音の環境音に対して収録（31 種類 × 40 音）

■ 話者数：8 名（男女各 4 名）

・立命館大学在学の 20 代を対象に収集

■ データ形式：48kHz, 16bit-linear PCM

収録した模倣音声の可視化



④ 環境音の模倣音声を用いた環境音合成

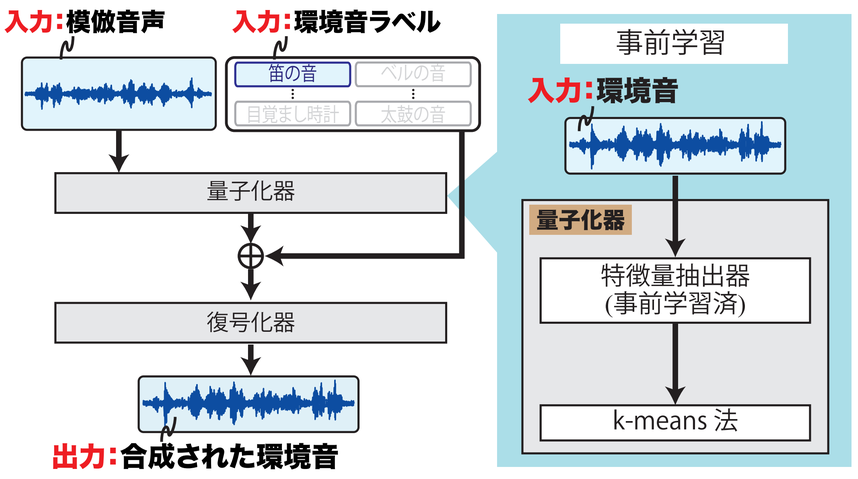
模倣音声と環境音ラベルを入力として環境音を合成

■ 量子化器と復号化器を利用して模倣音声と環境音の関係性をモデル化

・量子化器：BYOL-A（環境音によって学習された特徴量抽出器）+ k-menas 法

→ 入力となる模倣音声を量子化することで雑音に頑健な特徴量抽出を期待

・復号化器：Tacotron2 のデコーダ (Attention + 2層の LSTM ベース)



模倣音声と環境音ラベルを入力とする環境音合成のモデルの概要

デモ

実験条件

■ 利用したデータセット

・環境音：ESC-50 に含まれる 31 種類の環境音（各音源 35 サンプル）

・模倣音声：本研究で収録した模倣音声（各環境音に対して 4 名分）

■ 環境音合成手法

・従来法：VA-VAE + PixelSNAIL（環境音ラベルを入力）

・提案法：BYOL-A + Tacotron2 のデコーダ（環境音ラベル + 模倣音声を入力）

実験結果

■ 入力した模倣音声に対する合成音の再現度合いを主観評価

・模倣音声に対して合成音が音高・リズムの観点で妥当であるか 5 段階で評価

・1（非常に妥当でない）~ 5（非常に妥当である。）　※各音に対して 10 名で評価

音高に対する評価結果

合成手法	ドアを叩く音	全体の平均
自然音	3.90±1.05	3.81±1.03
従来法	2.88±1.19	2.54±1.15
提案法	3.44±0.91	2.65±1.04

リズムに対する評価結果

合成手法	ドアを叩く音	全体の平均
自然音	4.08±0.99	3.87±1.01
従来法	2.78±1.33	2.54±1.15
提案法	3.32±1.19	2.62±1.09

参考文献

[1] Y. Okamoto et al., “Overview of Tasks and Investigation of Subjective Evaluation Methods in Environmental Sound Synthesis and Conversion,” arXiv preprint arXiv: 1908.10055, 2019.
[2] Y. Zhou et al., “Visual to Sound: Generating Natural Sound for Videos in the Wild,” Proc. CVPR, pp. 3550-3558, 2018.
[3] Y. Okamoto et al., “Onoma-to-wave: Environmental Sound Synthesis from Onomatopoeic words,” APSIPA Transctions on Signal and Information Processing, Vol. 11, No. 1, e13, 2022.
[4] F. Kreuk et al., “AudioGen: Textually Guided Audio Generation,” Proc. ICLR, 2023.
[5] K. J. Piczak, “ESC: Dataset for Environmental Sound Classification,” Proc. ACM International Conference on multimedia, pp. 1015-1018, 2015.

今後の展望

■ 模倣音声のデータ拡張による合成音の品質向上

■ 量子化器，復号化器の部分に使用するモデル構造の改良

・使用する特徴量抽出器やクラスタ数の変更