一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report

環境音と説明文の意味的関連性の自動評価に向けたデータセット構築と基本性能評価

岡本　悠希[†]　金森　勇介[†]　高野　大成[†]　高道慎之介[†,††]
齋藤　佑樹[†]　永瀬亮太郎[†††]　猿渡　洋[†]

† 東京大学大学院情報理工学系研究科　〒113–8656 東京都文京区本郷 7–3–1
†† 慶應義塾大学理工学部　〒223–8522 神奈川県横浜市港北区日吉 3–14–1
††† 立命館大学情報理工学部　〒567–8570 大阪府茨木市岩倉町 2–150
E-mail: †y-okamoto@ieee.org

あらまし　機械学習技術や計算機の発展に伴い，音声や楽音に限らないあらゆる音である環境音を合成する取り組みが盛んに行われている．環境音合成技術の 1 つとして，「犬が吠える後ろで男性が話している」のような音を説明した文から環境音を合成する text-to-audio (TTA) が注目を集めている．TTA の評価として，入力となった説明文と合成された環境音の意味的な関連性が主観/客観の両側面で評価されている。しかしながら，人間による主観評価は金銭・時間の観点からコストが高く，既存の客観評価では人間による主観評価との相関が低いことが報告されている。そこで本研究では，音の説明文と環境音の意味的関連性の自動評価に向けたデータセットを構築する。また，構築したデータセットを用いて，説明文と環境音の関連性自動評価モデルを作成して，現状の到達点に関して議論する．
キーワード　環境音合成，text-to-audio，CLAPScore，主観評価

Construction and performance evaluation of dataset for automatic evaluation of the semantic relevance of environmental sounds and texts

Yuki OKAMOTO[†], Yusuke KANAMORI[†], Taisei TAKANO[†], Shinnosuke TAKAMICHI[†,††],
Yuki SAITO[†], Ryotaro NAGASE[†††], and Hiroshi SARUWATARI[†]

† The University of Tokyo
†† Faculty of Science and Technology, Keio University
††† College of Information Science and Engineering, Ritsumeikan University
E-mail: †y-okamoto@ieee.org

Abstract　With the development of machine learning, environmental sound synthesis, which is not limited to speech and music, has been proposed. One of the environmental sound synthesis that has been attracting attention is text-to-audio (TTA), which generates environmental sounds from text, such as “a dog barks while a man talks in the background.” In TTA, the relevance between text and output audio is an important evaluation aspect. Typically, it has been evaluated from both subjective and objective perspectives. However, subjective evaluation is costly in terms of money and time, and objective evaluation has low correlation to subjective evaluation scores. In this work, we construct an open-sourced dataset that subjectively evaluates the relevance. Also, we propose a model for automatically evaluating the subjective relevance between text and audio. Our model outperforms a conventional method, and that trend extends to many sound categories.
Key words　Environmental sound synthesis, text-to-audio, CLAPScore, subjective evaluation