進捗報告

1　先週決まったこと

とりあえず, 音声の生成タスクについて触れて動かしてみようということになった. 実装までこぎつけなかったので先行研究や理論をまとめる.

2　Onoma-to-wave

2.1　説明

Onoma-to-Wave [1] とは, 井本らが考案した系列変換モデルを用いたオノマトペからの環境音合成手法である. 背景として, 従来の環境音合成では, 入力として音響イベントラベル (風の音, 笛の音 ...), 出力として合成された音声という, 音響イベントラベルと出力として求められる音が一対一対応であった (KanaWave). この問題点としては, 生成される音に多様性 (ピッチ, 音色など) が欠けてしまうことである. 例として, 同じ笛の音でも出力としては「ピー」だけでなく「ビー」や「ピッピ」のように多義的である. そこで, 井本らは音響イベントラベルではなく, オノマトペを音の特徴をより良い表現で表せる手段として採用し, 音の多様性の制御が可能であると考えた.

図 1 に Onoma-to-wave のモデル概略図を示す. 左側のモデルは入力としてオノマトペのみを用いて, Encoder ではオノマトペの分散表現を抽出し, Decoder 側で音響特徴量を推定し, スペクトログラムを生成している. この研究ではスペクトログラムから波形への復元方法としては Griffin-Lim アルゴリズムを使用している. (その他の方法としては WaveNet Vocoder 等が挙げられる.)

右側のモデルはオノマトペと音響イベントラベル (One-hot 表現) 両方を入力としたモデルである. 音響イベントラベルも入力に加えた理由としては, 例えば同じ「パン」というオノマトペでも風船が割れる音なのか, ピストルの音なのか分からず, 音響イベントの種類が制御できない可能性があるからである.

また, 井本らの最新の研究 [2] では図 2 に示すように, Encoder 及び Decoder 部を Transformer に置き換えた手法が提案されている. なお, 筆者らによる学習済みモデルの公開はされていない.

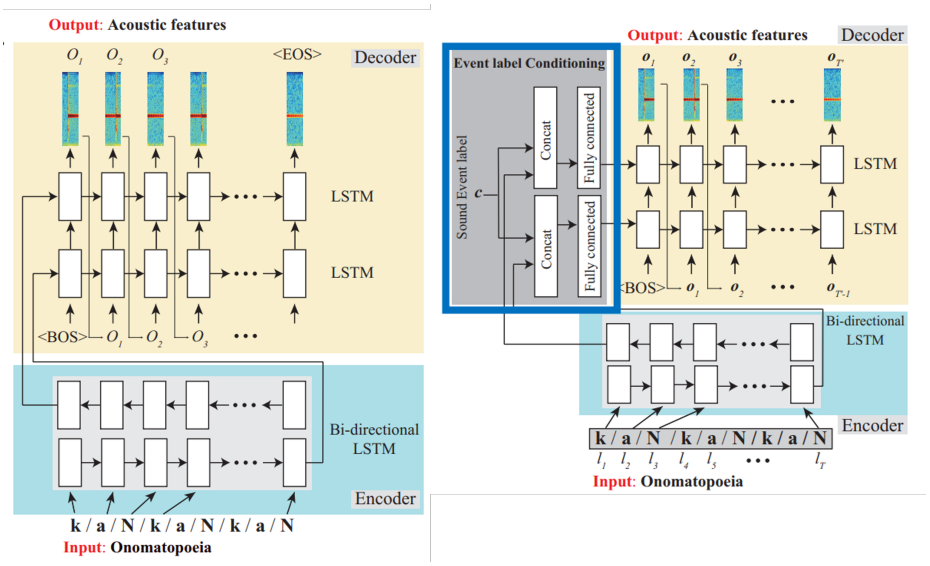

図 1: Onoma-to-Wave モデル概略図
(左) オノマトペのみを入力
(右) オノマトペ + 音響イベントラベルを入力

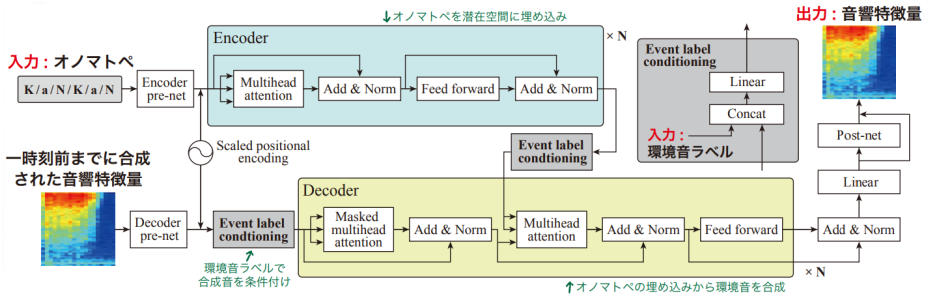

図 2: Onoma-to-Wave：Transformer による環境音・効果音の合成モデル

2.2 データセット

学習には, RWCP 実環境音声・音響データベース https://research.nii.ac.jp/src/RWCP-SSD.html に含まれる 10 種類の環境音（計 1000 サンプル）及び, 井本らによって作成された RWCP-SSD-Onomatopoeia https://github.com/KeisukeImoto/RWCPSSD_Onomatopoeia [3] に含まれる 14250 個のオノマトペを用いている. 後者は前者のデータセットに含まれる環境音一つ一つに対して多くのオノマトペをラベリングしたもので, オノマトペの単語に対する自己報告の信頼度スコアと, 他人から報告された受け入れスコアも含まれているため, オノマトペの単語の適切性を評価するために使用できる.

2.3 実験結果

音のみ, または音とオノマトペを被験者に提示し, 4 つの指標 (環境音の全体的な印象, 環境音の自然性, オノマトペに対する環境音の許容度, オノマトペに対する環境音の表現性) を 5 段階で評価した結果, 従来手法 (KanaWave, WaveNet) と比較して同程度の品質を獲得でき, 許容度および表現性 h はともに高いスコアを獲得した. また, 音響イベントラベルも使用することで同一オノマトペから様々な音響イベントを表現可能であることが分かった. デモページはこちら https://y-okamoto1221.github.io/Onoma_to_wave_Demonstration_jp/

3 やること

- RWCP 実環境音声・音響データベースの中身の確認
- 学習できそうなら回して生成してみる
- また相談します

参考文献

[1] Yuki Okamoto, Keisuke Imoto, Shinnosuke Takamichi, Ryosuke Yamanishi, Takahiro Fukumori, and Yoichi Yamashita. Onoma-to-wave: Environmental sound synthesis from onomatopoeic words. *APSIPA Transactions on Signal and Information Processing*, 11(1):–, 2022.

[2] Yuki Okamoto, Keisuke Imoto, Shinnosuke Takamichi, Takahiro Fukumori, and Yoichi Yamashita. How should we evaluate synthesized environmental sounds, 2022.

[3] 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)*, pages 125–129, 2020.