4Q-08

音楽除去に特化した深層学習モデル

小澤 俊　　小河 誠巳　　神戸 英利
東京電機大学大学院 理工学研究科 情報学専攻

1　はじめに

近年，動画投稿サイトやSNSの流行により個人で動画を編集することも珍しくなくなった．より質の良い動画を作成する上で音声は非常に重要な構成要素であり，スマートフォンの性能やAIの精度の向上に伴って音声を編集する技術も一般ユーザー向けに普及し始めている．そんな中，インターネット上に動画を投稿する際には著作権への注意が必要であり，動画投稿サイトによっては音楽の検出が自動で行われるため，使用許可を得ていない音楽が動画内に偶然含まれてしまったような場合においては，不必要なトラブルを回避するためにも何らかの処理を施すことが望ましい．また，動画の質という観点で見ると，処理を行なった後の音声は可能な限り元の音声を維持していることが理想的である．しかし，このような元音源の用意が難しい状況において，現在の動画編集ソフト等の機能で対象の音楽のみを除去することは困難なのが現状である．また，本稿では以降，音楽として一般に公開されている学習に使用可能なデータを「音楽」と定義し，音楽以外の音全てを「環境音」と呼ぶこととする．

本稿は，音楽と様々な種類の環境音が混ざった音声から音楽のみを除去することを目的とし，その予備実験としてモノラル信号を対象としたシミュレーション環境による音楽除去の実験を行なった．

2　関連研究

音楽の除去に関連する研究として，convolutional denoising autoencoder (CDAE) を用いて人の声と音楽の混合音から音楽を除去する研究 [1] が行われている．この研究はCDAEは様々なジャンルの音楽のパターンを学習可能であり，CDAEに基づいた音楽除去は音声認識の性能を大幅に向上させるという結果を示している．しかしながら，この研究は音楽を除去することで音声認識の精度を向上させることが目的であり，音楽が除去されているかの評価は行われていない．また，様々な環境音と音楽の混合音からテキストで指定した音だけを分離する研究 [2] も行われており，環境音と楽器音の分離に対する評価が行われている．しかし，環境音と音楽の枠組みでの分離は行われておらず，音楽の除去に特化した手法は確立されていない．

3　使用データ

3.1　環境音データセット

環境音には公開データセットESC-50[3]を使用した．このデータセットは50種類の環境音を長さ5秒，モノラルチャンネル，サンプリング周波数41kHzで収録したデータセットであり，本稿ではサンプリング周波数16kHzにダウンサンプリングしたものを使用した．

3.2　音楽データセット

音楽データには公開データセットFree Music Archive[4]のMediumサイズを使用した．このデータセットは16ジャンルの音楽を1曲あたり30秒，2チャンネル，サンプリング周波数41kHzで収録したボーカルを含む音楽が含まれているデータセットであり，本稿ではサンプリング周波数16kHz，チャンネル数1にダウンサンプリングしたものを使用した．

4　手法

本稿では，環境音 e と音楽 s の混合音 x から，混合音 x をもとに推定した音楽 $\hat{s}$ を引くことで環境音 $\hat{e}$ の推定を行う．音楽の推定には，時間領域による音源分離の手法であるTasNet[5]と同様に非負値行列因子分解（NMF）の計算方法を真似た特徴量を用いる．一般的にNMFでは時間周波数領域に変換した非負値の行列を非負値の基底行列と係数行列の積として表すが，TasNetでは時間領域の信号を直接，非負値の基底行列と係数行列の積として表し，そのような基底行列と係数行列はニューラルネットワークのパラメータとして学習される．すなわち，本稿では環境音に混ざった音楽を「マスクした非負の重み行列 $\hat{\mathbf{W}}$」と「非負の基底行列 $\mathbf{B}$」の積として表すことができるような行列 $\hat{\mathbf{W}}$ と行列 $\mathbf{B}$ を学習によって求めていく．

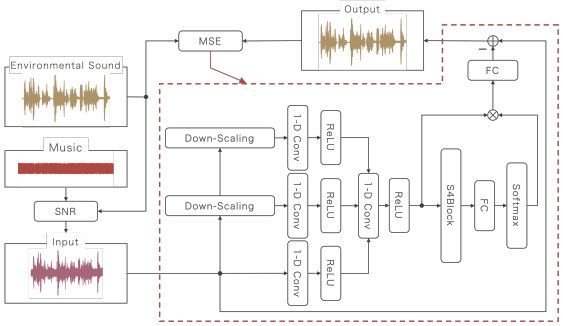

図 1: ネットワークモデル

まず，入力された混合音 x を長さ L，K 個のセグメント $\boldsymbol{x}_k$ に分割し，1次畳み込みを行わない信号，1回行った信号，2回行った信号の3つのスケールの信号に変換する．それぞれのスケールの信号から重みベクトル $\boldsymbol{w}_k$ を以下のようにして求める．

$$\boldsymbol{w}_k = ReLU\left(Conv1D(\boldsymbol{x}_k)\right)$$

$Conv1D$ は1次畳み込み層を表す．その後，求めた3つの重みベクトルを1次畳み込みにより1つにまとめる．

次に，入力の混合音に含まれる音楽を再構成できるよう重み行列 $\mathbf{W}$ に対するマスク $\mathbf{M}$ の推定を行う．本稿では TasNet における Separation 内の LSTM を S4[6] に変更したものを使用した．ここでは重みベクトル $\boldsymbol{w}_k$ を用いて K 個間の時間的な依存関係を見ながらマスク $\mathbf{M}$ の推定を行う．

推定したマスク $\mathbf{M}$ によってマスクされた重み行列 $\hat{\mathbf{W}}$ に基底行列 $\mathbf{B}$ を掛けることで混合音に含まれる音楽 $\boldsymbol{s}_k$ が再構成される．基底行列 $\mathbf{B}$ は重み行列 $\mathbf{W}$，マスク $\mathbf{M}$ の推定と同時にネットワーク全体で学習される．また，推定した音楽 $\hat{s}$ を混合音 x から引くことによって音楽の除去が行われる．

5　実験

学習に用いる音楽はジャンルが ”Country” の計 142 曲とし，ランダムに選んだ環境音 12 個を SN 比が平均 5，標準偏差 3 の正規分布からランダムに生成した値になるように混合した．長さ 30 秒の混合音 142 個をそれぞれ 320 サンプル（20ms），オーバーラッピングなしで 15,000 個に分割し，基底ベクトルを 1,000 本に設定した．学習率 0.001，オプティマイザーに Adam，損失関数を MSE として 2,000 エポック学習した．

6　結果

ランダムに選んだ環境音 12 個をランダムな時間に配置した環境音（目的信号），ジャンルが”Country”であるテストデータ，目的信号である環境音に SN 比 5dB で音楽を混ぜた混合音（入力信号），その混合音を学習したモデルを用いて音楽除去した信号（出力信号）の４つをスペクトログラムに変換したものを以下の図 2 に示す．

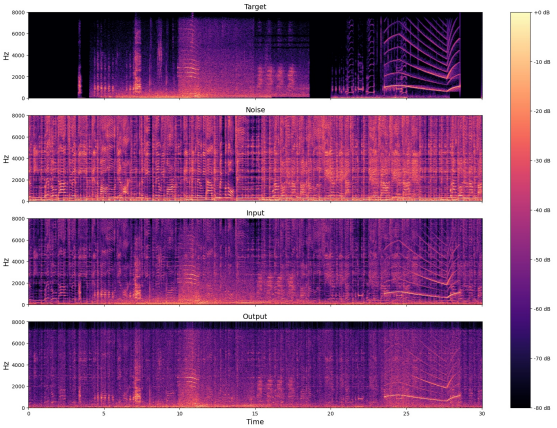

図 2:（上から）目的信号，音楽，入力信号，出力信号

音楽を除去した出力信号について試聴を行なったところ，音楽はわずかに低減されているものの，まだ音楽であると認識できる程度の除去の精度であった．

7　考察

図 2 より，目的信号である環境音と音楽除去後のスペクトログラムを見比べると，差は歴然であり音楽除去の精度は悪いことがわかる．しかし，データ数の少なさや，ハイパーパラメータの調整不足，ネットワーク構造の細部への試行錯誤が足りていない点を鑑みると，提案したネットワーク構造が音楽除去に不向きであるとは必ずしも言えないと考えられる．

8　まとめ

本稿では，TasNet のネットワーク構造に基づいた時間領域による音楽除去の手法を提案した．環境音と音楽から混合音を作成し，混合音中の音楽が再構成可能となるように基底行列と係数行列の学習を行なった．その結果，ある程度音楽の低減が確認された．

9　今後の課題

本稿では，損失関数に MSE を用いているが，MSE は振幅の大きさのずれを考慮できないため，SI-SNR を用いた学習を行なっていきたい．また，データ数や音楽のジャンルを増やしての学習や，音楽を事前学習し，基底行列をある程度学習させた転移学習についても行なっていきたい．

参考文献

[1] M. Zhao, D. Wang, Z. Zhang, et al., ”Music removal by convolutional denoising autoencoder in speech recognition,” APSIPA ASC 2015, Hong Kong, China, 2015, pp. 338-341.

[2] X. Liu, Q. Kong, Y. Zhao, et al., ”Separate Anything You Describe,” arXiv:2308.05037, 2023.

[3] Karol J. Piczak, ”ESC: Dataset for Environmental Sound Classification,” Proceedings of the 23rd Annual ACM Conference on Multimedia, 2015, pp.1015-1018, isbn: 978-1-4503-3459-4.

[4] M. Defferrard, K. Benzi, P. Vandergheynst, et al.: FMA: A Dataset For Music Analysis, arXiv:1612.01840, 2016.

[5] Y. Luo and N. Mesgarani, ”TaSNet: Time-Domain Audio Separation Network for Real-Time, Single-Channel Speech Separation,” 2018 IEEE ICASSP, Calgary, AB, Canada, 2018, pp. 696-700.

[6] A. Gu, K. Goel, C. Ré, ”Efficiently Modeling Long Sequences with Structured State Spaces,” arXiv:2111.00396, 2022.