6T-03

深層学習を用いた背景音楽の低減

小澤　俊　　小河　誠巳
東京電機大学理工学部情報システムデザイン学系

1　はじめに

近年，動画投稿サイトやSNSの流行により動画をインターネット上に投稿することも珍しくなくなったが，その際に注意しなければならないのは著作権の問題である．例えば，街中や店内で撮影した動画の中で著作権のある音楽が流れており，権利者の許可なくその動画を投稿した場合，それは著作権侵害に当たる．そのため，音楽の使用許可が得られない場合にはその動画に何らかの処理を施してから投稿する必要がある．しかし，背景に流れる音楽を除去し，その他の音を残したい場合，現在の動画編集ソフト等の機能でそれを行うことは困難なのが現状である．本稿では以降，人が話している声を「発話」，発話や音楽，車の走行音や動物の鳴き声などと言った音全般を「音声」と呼ぶこととする．

本稿の目的は，車の走行音や人々の会話，音楽などの様々な音が混ざった音声から音楽のみを除去することである．例えば，音声強調を用いることで背景雑音を除去することが可能であり，55種類の歪みを考慮した深層学習による音声強調の研究[1]では高い精度を出している．しかし，本稿では音声強調に伴った音楽の除去ではなく，音楽のみの除去を目的としている．そのために，モノラル信号の発話と音楽の混合音から深層学習を用いて音楽のみを低減する予備実験を行なった．

2　関連研究

モノラル信号の発話と音楽の混合音に対する音楽低減に関連する研究として，複素NMFDの原理に基づいた発話と音楽の分離の研究[2]がある．発話と音楽のスペクトログラムのパターンの違いを学習させることで発話を抽出するマスクを作成し，混合音から発話の抽出を行う研究である．また，この研究では音楽の抽出についても同様に行なっており，複素NMFDの原理に基づく抽出は発話・音楽の両者に有効であることが確認されたことから，背景音楽抑圧への応用が期待されている．

3　使用データ

発話データには公開データセットLibriSpeech[3]のtrain-clean-100を使用した．このデータセットは男女251人による無言の時間を含む英文の朗読100時間分をモノラルチャンネル，サンプリング周波数16kHzで収録したデータセットである．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，本稿ではサンプリング周波数16kHz，チャンネル数1に変換したものを使用した．

4　手法

本稿では深層学習を用いることで音楽の除去を行なった．発話と音楽を任意のSN比で混合した後に学習器に入力し，音楽成分が消えるように学習器のパラメータを更新することによって学習・音楽除去を行なった．

学習データの複雑度を下げるため，音楽データにはジャンルをHip-Hop, Popの2種類に絞り込んだ計2700曲を使用した．また，音楽に混合する発話の組み合わせはランダムとしている．学習する際は音楽を混合させない発話のみのデータも学習データに含ませており，混合する際は音楽を雑音と見立ててSN比5デシベルで発話と混合して学習器に入力した．ネットワーク構造にはConv-TasNet[5]を使用した．Conv-TasNetは時間領域による音声分離の手法であり，混合音をエンコーダによって写像し，写像した信号からマスクを推定しマスキング，デコーダで時間波形に変換することで混合音を分離する手法である．パラメータの更新は学習器の出力と混合する前の発話とのRMSEが小さくなるように学習率0.0001で行い，オプティマイザーにはAdamを使用した．本稿ではサンプル長の違いによる精度を比較するため，学習器へ入力する混合音のサンプル長を50ms，100ms，250ms，500ms，750ms，1000msに変え，それぞれ20エポック学習させた．

5　結果

学習を終えた後，混合する際のSN比を変化させて推論したときのRMSEを，それぞれのサンプル長について比較したものを以下の図1に示す．また，テストデータに用いた音楽データにはボーカルが含まれており，発話データには学習データに用いた人物と同じ人物による発話を含んでいる．

Copyright ©2023 Information Processing Society of Japan.
All Rights Reserved.

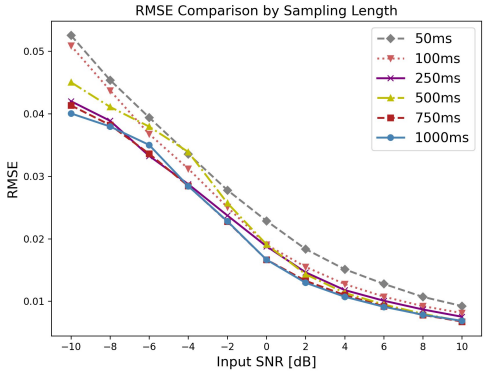

図 1: サンプル長ごとの RMSE の比較

サンプル長 1000ms で学習させた学習器で推論したときの入力の混合音，出力の推論結果，目的信号である混合前の発話のスペクトログラムをそれぞれ図 2，3，4 に示す．

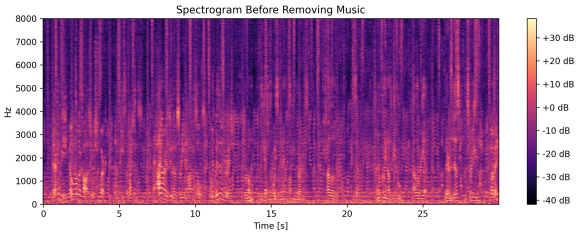

図 2: 音楽除去前のスペクトログラム

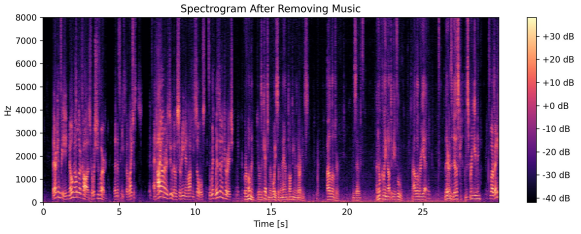

図 3: 音楽除去後のスペクトログラム

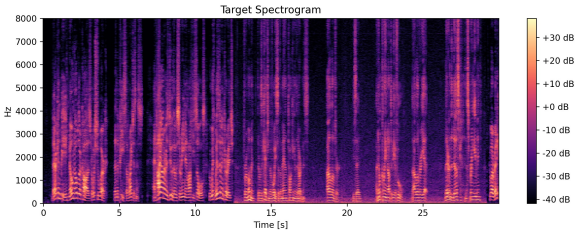

図 4: 混合前の発話のスペクトログラム

6 考察

図 1 より，サンプル長が長い方が精度が良い傾向にあることがわかる．このことから，学習する際にはサンプル長が長い方が音楽の特徴を捉えることができるのではないかと考えられる．また，図 2 と図 3，図 4 を比べると，音楽除去後のスペクトログラムは音楽除去前に比べて発話のスペクトログラムに似ていることから，音楽が低減されていることがわかる．また，著者が図 3 の音声の試聴を行なったところ，ボーカルを含む音楽全体が低減されているが，ボーカルの残る箇所がしばしば確認された．また，全体的に音楽は低減されているものの，音楽ではなく発話の特徴を学習している可能性があるため，音楽の特徴を学習しているかどうかを確かめることが今後の課題である．本稿では使用する音楽データのジャンルを絞っているため，ジャンルやデータ数を増やして学習を行うことでさらに精度が向上するのではないかと考えられる．

7 まとめ

本研究では，様々な音が混ざった音声から音楽のみを除去することを目標とし，発話と音楽の混合音から深層学習を用いることで音楽を除去するシミュレーション実験を行なった．その結果，学習に用いるデータのサンプル長は，50ms から 1000ms の範囲では長い方が精度が良くなる傾向にあることが確認された．また，試聴を行ったところ，音楽の低減は確認されたものの完全な除去までには至らなかった．今後の課題として，発話ではなく音楽を学習しているかの検証が挙げられる．また，使用する音楽データのジャンル，音楽・発話のデータ数を増やしての学習を行なっていきたい．

参考文献

[1] J. Serr à , S. Pascual, J. Pons, et al.: Universal Speech Enhancement with Score-based Diffusion, arXiv:2206.03065, 2022.

[2] 北野佑，亀岡弘和，柏野邦夫ほか：複素 NMFD による音声抽出マスクの設計と背景音楽抑圧への応用，日本音響学会 2009 年春季研究発表会講演論文集，pp.719-720，2009.

[3] V. Panayotov, G. Chen, D. Povey, et al.: Librispeech: an asr corpus based on public domain audio books, 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 2015, pp.5206-5210, doi: 10.1109/ICASSP.2015.7178964.

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

[5] Y. Luo, N. Mesgarani: Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation, arXiv:1809.07454, 2018.

Copyright ©2023 Information Processing Society of Japan.
All Rights Reserved.