日本音響学会　2023 年春季研究発表会に参加して

岡　田　健　佑
Kensuke OKADA
情報メディア学専攻修士課程　2022 年度修了

1. はじめに

私は，2023 年 3 月 15 日から 17 日にオンラインで開催された日本音響学会 2023 年春季研究発表会に参加し，「残響除去を目的とした Conv-TasNet の学習法」という題目で発表を行った．

2. 研究背景と目的

残響除去モデルの一つに時間領域で直接信号を推定する Conv-TasNet（fully-convolutional time-domain audio separation network）[1]がある．このモデルは時間周波数領域で学習した従来の残響除去より，残響除去性能が高い結果が報告されている[2]．しかし残響除去の際，処理後の音声に歪み発生する問題点がある．本研究では Conv-TasNet による残響除去で発生する処理後の歪みを軽減する学習法について，学習時の損失関数や用いる学習データの観点から検討を行う．

3. 提案手法

本研究では，高域成分が考慮されにくい点と残響除去後音声に歪みが発生する問題点を解決のために，次のような 3 つの学習構成の検討を行い，実験結果より有効性を評価する．

3.1 プリエンファシスフィルタの適用

高周波数帯域を考慮して学習するようにする目的で学習時のターゲット音声に高域強調フィルタを適用する．今回は検討実験の結果より $p=0.05$ とした．

3.2 損失関数の変更

損失関数を時間領域の Si-SDR で行うのではなく，モデルからの出力に短時間フーリエ変換を行い，時間周波数領域で損失を計算する．損失関数は MSE（Mean Square Error）を用いる．

3.3 段階的な残響除去

残響成分を一括で除去するのではなく，残響成分を段階的に少しずつ減らしていく．構成を図 1 に示す．従来の手法では様々な残響時間の音声を入力した時の出力は全て残響がついていない音声が目標となる．段階的な残響除去では複数のそれぞれ役割が異なるモデルを用い，残響成分を段階的に除去する．例として図 1 にある 0.5－＞0.4 モデルという学習モデルでは，残響時間 0.5 秒の残響音声を残響時間 0.4 秒の残響音声に近づけるように学習を行う．そのため（b）では，使用するモデル数は 5 つでそれぞれ役割が異なるモデルとなる．

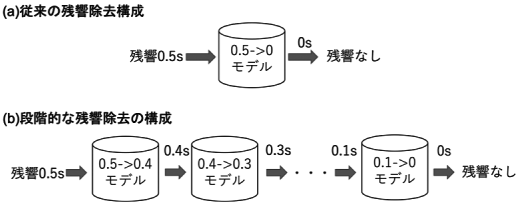

図 1　段階的な残響除去構成

4. 評価実験

提案手法の性能を評価するため，シミュレーションにおける評価実験を行った．

4.1 実験条件

実験に用いる Conv-TasNet のパラメータは[1]の論文と同じパラメータを採用した．学習時の損失関数は Si-SDR を用いた．学習と評価に用いるデータは Pyroom Acoustics ライブラリを用いて作成した．音声データは約 8 秒の男女各 48 名，計 96 名の音声で，RT60 は 0.3 秒と 0.5 秒，マイクから音源の距離は 1m の距離に配置してシミュレーションを行な

った．雑音は hoth 雑音を用いている．学習時の SNR は 20, 30dB，評価時の SNR は 20dB としている．音声データのサンプリング周波数は 16kHz である．残響除去の客観的な尺度として PESQ, STOI と Si-SDR を使用した．また主観評価として DMOS 評価を用いた．

4.2 客観評価実験結果

表 1 は RT0.5 の残響除去を行った結果を示す．処理前の場合と従来の時間領域（Si-SDR Loss）での残響除去処理の結果を示している．表の Si-SDR 以降は 3 章で示した検討内容順に示している．実験結果では，PESQ と STOI に関しては全てにおいて改善している．しかし段階的な残響除去においては MSE loss に比べて Si-SDR の評価値が及ばなかった．

表 1　客観評価実験結果

	PESQ	STOI	Si-SDR
処理前	1.898	0.399	-7.410
Si-SDR Loss	2.234	0.528	10.06
プリエンファシス	2.250	0.529	10.18
MSE Loss	2.317	0.537	10.85
段階的な残響除去	2.349	0.552	10.40

次に図 2 は原音と MSE loss，段階的な残響除去のスペクトログラムを示す．(b) では低周波数帯域の発話後の無音部にある残響成分が消えているのに対して，(c) では発話後の残響成分が残っていることがわかる．また発話部分に含まれている残響成分も除去しきれていないため，Si-SDR の評価値が及ばなかったと考えられる．

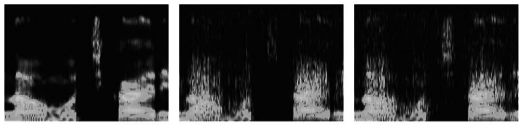

図 2　残響除去後のスペクトログラム

4.3 主観評価実験結果

客観評価指標のみでは除去後の歪みを評価できないため主観評価実験を行った．提案手法は段階的な残響除去のことを示す．評価は 20 代の男女計 12 名の DMOS である．評価結果を図 3 に示す．DMOS は原音声と残響除去後音声を比較して 1～5 の値で評価する．値が高いほど音声の劣化がない，つまり歪みがないことを示す．実験結果から段階的な残響除去の構成は従来の手法と比べて除去後の歪みが抑えられていることがわかる．

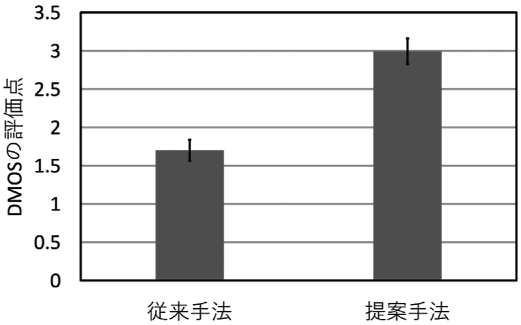

図 3　DMOS の平均点と 95% 信頼区間

5. おわりに

発表に参加し，多くの方々から意見を頂き，大変参考になった．研究や発表に対して多大なご指導を頂いた片岡章俊教授に深く感謝いたします．

参考文献

［1］ Y. Luo et al., IEEE/ACM TASLP, vol. 27, no. 8, pp.1256-1266, 2019.

［2］ Y. Luo and N. Mesgarani. Real-time single-channel dereverberation and separation with time-domain audio separation network. In Interspeech, pages 342-346, 2018.