Transformer を用いた人工データでの学習による劣化音源からの音声復元手法の提案

小島 巧実 [1]　高野 敏明 [1]
[1] 静岡理工科大学大学院　理工学研究科
2421010.kt@sist.ac.jp takano.toshiaki@ieee.org

概要

フィールドワークやインタビュー調査などさまざまな場面で音声データが収集され，その記録や分析が行われている．しかし実際の収録環境では，環境雑音や録音機器の制約，録音媒体の劣化などの影響により音質が低下し，聞き取りや解析が困難となる場合がある．本研究では，劣化音声から聴感的に自然な音声を復元することを目的として，メルスペクトログラムを入力とする Transformer ベースの音声復元手法を提案する．提案手法は，時間周波数特徴量を Transformer に入力し，得られたメルスペクトログラムを HiFi-GAN によって音声波形へ変換する．VoiceBank-DEMAND を用いた評価の結果，特に低 SI-SNR 条件において，PESQ および STOI による知覚的評価指標が改善する傾向が確認された．

1　はじめに

屋内外のさまざまな環境で収録される音声データは，研究活動や記録保存において重要な情報源となっている．会議や講義などの室内収録に加え，フィールドワークやインタビュー調査などの屋外収録においても音声記録が広く用いられている．しかし音声信号は，環境雑音や録音媒体の劣化などの影響により品質が低下しやすいという問題を持つ．このため，劣化音声から有用な情報を復元する音声デノイズ技術は，音声情報処理分野における重要な研究課題となっている．特に，スマートフォンや携帯型ボイスレコーダなどに搭載される小型マイクは，指向性や防風性能に制約があるため，録音環境を十分に制御することが難しい．その結果，風音や周囲雑音が混入しやすく，音声品質が大きく劣化する場合が多い．そのため，収集した音声が聞き取りにくくなり，分析や記録作業に支障をきたすという問題が生じている．

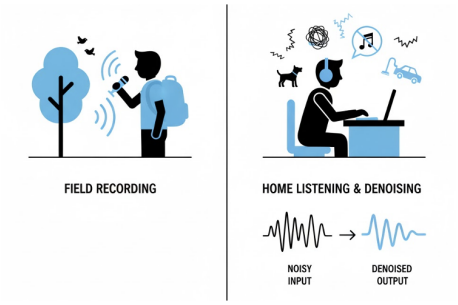

図 1　録音環境における雑音混入の例

従来の音声デノイズ手法としては，U-Net や RNN に代表される畳み込み型および再帰型ニューラルネットワークが広く用いられてきた [1]．これらの手法は局所的な時間・周波数特徴の抽出に優れる一方で，発話全体にわたる長期的な時間文脈を十分に活用することが難しいという課題を持つ．近年，Transformer は系列全体の依存関係を明示的に扱うことができるモデルとして注目されており，音声デノイズタスクへの応用が報告されている [2]．

本研究では，Transformer とニューラルボコーダを用いた音声復元手法を提案する．人間の聴覚特性を反映した特徴量を利用することにより，雑音環境下においても自然で理解しやすい音声を得ることを目的とする．

2　提案手法

本研究では，音声デノイズタスクに対して，メルスペクトログラム [3] を入力とする Transformer ベースの音声復元モデルを提案する（図 2）．提案手法は，

1. メルスペクトログラムによる特徴量抽出
2. Transformer による時間周波数特徴の雑音抑圧
3. HiFi-GAN による波形生成

の三段階で構成される．本手法は，波形を直接復元するのではなく，聴覚特性に基づく中間表現を介した生成的復元を行う点に特徴がある．音声波形の復元を最大化するのではなく聴感的品質の向上を目的

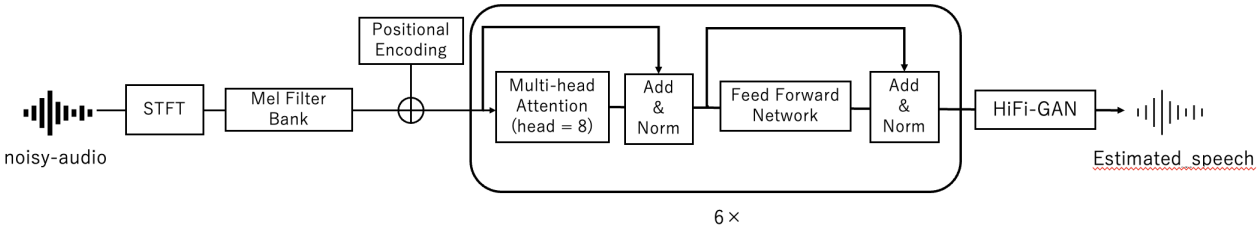

図 2　提案手法のモデル構造

とし, メルスペクトログラム空間での Transformer 処理とニューラルボコーダを統合する.

2.1　Transformer

Transformer は自己注意機構（Self-Attention）に基づく系列変換モデルであり, 系列中の全要素間の依存関係を同時に考慮できる点に特徴がある. 本研究では, 入力メルスペクトログラムを時間方向の系列として扱い, エンコーダ型 Transformer により雑音抑圧された特徴表現を推定する.

各エンコーダ層では，Multi-Head Self-Attention により各時間フレームが他のすべてのフレームを参照し，時間的に離れた発話成分間の相関関係を学習する．これにより，局所的な時間情報に依存する従来手法と比較して，広域な時間文脈を考慮した特徴更新が可能となる．Attention 出力は残差接続および Layer Normalization により安定化され，その後，位置ごとの Feed-Forward Network により非線形変換が施される．また，位置エンコーディングを付加することで時間順序情報を保持する．本モデルはエンコーダ層を 6 層，自己注意ヘッド数を 8 として構成され，最終的に推定されたメルスペクトログラム $\hat{M}$ と参照クリーンメルスペクトログラム M との平均二乗誤差（MSE）を最小化するように学習を行う．損失関数は次式で定義される.

$$\mathrm{L_{MSE}} = \frac{1}{N}\sum_{i=1}^{N}\left\|\hat{M}_i - M_i\right\|^2 \tag{1}$$

2.2　HiFi-GAN

HiFi-GAN は, メルスペクトログラムから高音質な音声波形を生成するニューラルボコーダである [4]. 本研究では, Transformer により推定されたメルスペクトログラムを, 事前学習済みの HiFi-GAN に入力することで音声波形を生成する. この生成過程により, 聴感的に自然な音声復元を優先する出力を得ることができる. HiFi-GAN のジェネレータは，メル

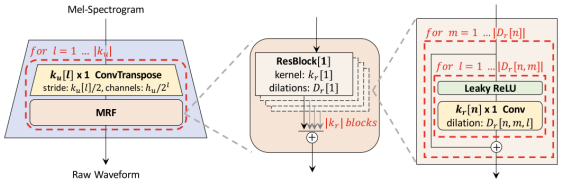

図 3　HiFi-GAN のモデル構造 [4]

スペクトログラムを入力とし，転置畳み込みにより時間方向に段階的なアップサンプリングを行うことで，最終的に音声波形と同じ時間分解能へと拡大する構造を持つ（図 3）．各アップサンプリング段には Multi-Receptive-Field モジュールが挿入されており，異なるカーネルサイズおよび膨張率をもつ複数の residual block から特徴量を統合することで，短時間成分から長時間成分まで多様な時間スケールの特徴を同時に扱うことが可能となっている．さらに，各 residual block 内では，膨張畳み込みと残差接続を組み合わせることにより，位相情報を含む微細構造を生成することができる.

3　実験

提案手法の有効性を検証するため，公開データセットのクリーン音声を用い，異なる雑音特性下におけるモデルの挙動を比較・分析した.

3.1　データセットおよび学習条件

本研究では，公開データセット VoiceBank-DEMAND に含まれる音声を用いて評価を行った．実験では，ホワイトノイズを付与した定常雑音条件と，FSD50K に基づく環境音を付与した非定常雑音条件の 2 種類を設定した.

VoiceBank-DEMAND データセットには，学習用 23,075 個およびテスト用 824 個のクリーン音声が含まれている．本研究では，学習用クリーン音声 23,075 個を対象とし，その 90% を学習データ，残り 10% をバリデーションデータとして用いた．テスト

用 824 個の音声は，最終評価のみに用いた．

学習データに付加する雑音の強さは，SI-SNR が −18, dB から 18, dB まで 6, dB 刻みとなるように設定した．この 7 段階の SI-SNR 条件を 23,075 個の各音声に適用することで，合計 23,075 × 7 = 161,525 個の学習用音声データを生成した．バリデーションデータおよびテストデータについても，同一の SI-SNR 条件でノイズ付加を行った．

非定常雑音条件では，FSD50K に含まれる環境音をノイズ源として用いた．FSD50K についても，公開されている学習用データを学習およびバリデーションに，評価用データをテストにそれぞれ対応させて使用した．これらのデータに対しても，7 段階の SI-SNR 条件でノイズ付加を行った．

実験条件は，入力特徴量のメルスペクトログラムは FFT サイズ 1024，ホップ長 256，メルバント数 80 とした．最適化手法のオプティマイザには AdamW を用い，学習率の設定は初期値 5.0×10-4 とし，Warmup ありの cosine Annealing を用いた．また，学習の設定として，バッチサイズ 32，エポック数 100 とした．評価方法としては，音声品質を評価する PESQ（Perceptual Evaluation of Speech Quality）および，音声の可聴性を評価する STOI（Short-Time Objective Intelligibility）を用いた．PESQ は人間の聴覚モデルに基づき主観的音質を推定する指標であり，一般に −0.5 から 4.5 の範囲の値を取り，値が大きいほど知覚的品質が高いことを意味する [5]．STOI は 0 から 1 の範囲の値を取り，値が大きいほど発話内容の理解しやすさが高いことを意味する [6]．本研究では比較対象として，バンドパスフィルタ（BPF），DNN[7], LSTM[8] を作成し，性能評価を行なった．表に示す各指標の値は，デノイズ処理後のテストデータに対して算出したスコアの平均値である．

3.2 定常雑音条件の結果と考察

定常雑音条件におけるデノイズ結果について，知覚的評価指標である PESQ および STOI の結果を表 1 および 表 2 に示す．

定常雑音条件において，提案手法は特に SI-SNR 条件が −18 dB の強雑音環境下では，PESQ および STOI のいずれの指標においても最も高い値を示し，聴感的品質および明瞭度の向上が確認された．一方で，中〜高 SI-SNR 条件においては指標ごとに優位なモデルが異なった．PESQ では，SI-SNR 条件が −6 dB 以上の条件において DNN が最も高い値を示し，知覚音質の観点では DNN が優位であった．これに対して，STOI では SI-SNR 条件が −12 dB 以上の広い条件で LSTM が最も高い値を示し，可聴性の観点では LSTM が優位であった．

表 1 各手法の PESQ 値（定常雑音条件）

Input	元音源	BPF	DNN	LSTM	提案手法
-18	1.04	1.10	1.10	1.09	**1.18**
-12	1.04	1.11	1.27	1.17	**1.30**
-6	1.04	1.14	**1.45**	1.27	1.41
0	1.05	1.22	**1.70**	1.40	1.52
6	1.10	1.38	**1.97**	1.56	1.63
12	1.21	1.67	**2.20**	1.76	1.74
18	1.44	2.02	**2.28**	1.97	1.79

表 2 各手法の STOI 値（定常雑音条件）

Input	元音源	BPF	DNN	LSTM	提案手法
-18	0.51	0.49	0.52	0.60	**0.61**
-12	0.58	0.56	0.67	**0.70**	0.69
-6	0.67	0.63	0.75	**0.76**	0.73
0	0.75	0.71	0.80	**0.82**	0.76
6	0.83	0.78	0.84	**0.87**	0.78
12	0.89	0.84	0.87	**0.91**	0.80
18	0.93	0.88	0.89	**0.94**	0.81

以上の結果より，提案手法は強雑音条件で一定の有効性を示した．ただし，−6 dB 以上の SI-SNR 条件では指標ごとに優位なモデルが異なり，必ずしも提案手法が最良となるわけではないことが確認された．

3.3 非定常雑音条件の結果と考察

FSD50K に基づく環境ノイズを付与した非定常雑音条件におけるデノイズ結果について，知覚的評価指標である PESQ および STOI の結果を表 3 および表 4 に示す．

非定常雑音条件において，PESQ に関しては，SI-SNR 条件が −18 dB から 0 dB の範囲では提案手法が最も高い値を示す条件が確認された．一方で，SI-SNR 条件が 6 dB 以上の条件では，BPF が最良の値を示した．

STOI に関しては，SI-SNR 条件が −18 dB の場合には提案手法が最良の値を示した一方，SI-SNR 条件が −12 dB から 0 dB の範囲では LSTM が最も高い値を示した．さらに，SI-SNR 条件が 6 dB の場合には元音声と LSTM が同等の値を示し，SI-SNR 条件が

表 3　各手法の PESQ 値（非定常雑音条件）

Input	元音源	BPF	DNN	LSTM	提案手法
-18	1.09	1.17	1.16	1.16	**1.28**
-12	1.12	1.23	1.22	1.22	**1.37**
-6	1.14	1.30	1.32	1.31	**1.47**
0	1.21	1.45	1.45	1.42	**1.59**
6	1.41	**1.75**	1.63	1.58	1.70
12	1.65	**2.00**	1.78	1.74	1.78
18	2.06	**2.34**	1.94	1.91	1.83

表 4　各手法の STOI 値（非定常雑音条件）

Input	元音源	BPF	DNN	LSTM	提案手法
-18	0.55	0.53	0.55	0.64	**0.65**
-12	0.63	0.60	0.63	**0.71**	0.70
-6	0.71	0.68	0.71	**0.78**	0.75
0	0.79	0.75	0.77	**0.82**	0.78
6	**0.86**	0.81	0.82	**0.86**	0.80
12	**0.90**	0.85	0.84	0.89	0.81
18	**0.94**	0.89	0.86	0.91	0.82

12 dB 以上の条件では元音声が最良となる傾向が確認された．

環境ノイズは時間的に非定常であり，特定の時間区間や周波数帯域に集中して出現する傾向を持つ．このため，発話成分が一時的に雑音に覆われる区間が生じやすく，時間方向の文脈情報を考慮した復元が重要となる．

定常雑音条件および非定常雑音条件を合わせた考察として，PESQ および STOI による知覚的評価指標において，特に低 SI-SNR 条件下で提案手法が有効に機能することが確認された．これは，メルスペクトログラムを介した特徴再構成およびニューラル音声生成モデルによる波形生成過程において，位相情報や微細な振幅構造が再推定されることにより，雑音成分が知覚的に緩和されるためであると考えられる．一方で, 両実験において得られた提案手法の結果から，STOI に着目すると，SI-SNR 条件が −12 dB から 18 dB における改善幅については限定的であり，既存手法と同等となる場合も確認された．この要因として，MSE 損失を用いた学習により，時間的に長く持続する母音成分が相対的に重視され，発話明瞭度に重要な破裂音や摩擦音などの子音成分の再現性が十分でない可能性が考えられる．

以上より，提案手法は定常雑音および非定常雑音という性質の異なる雑音条件下において共通して，聴感的品質および可聴性の改善を重視した生成的音声復元を実現しており，SI-SNR 条件が − 18 dB の場合において有効であることが示された．

4　まとめ

本研究の目的は，メルスペクトログラムを入力とする深層学習モデルを用いて，低 SNR 条件における音声の聴感的品質を向上させることにある．本研究では，メルスペクトログラムを入力とする Transformer と HiFi-GAN を組み合わせた生成的音声復元手法を提案し，VoiceBank-DEMAND を用いた定常および非定常雑音下での実験評価を行った．

実験の結果，提案手法は PESQ および STOI において，特に低 SI-SNR 条件下で音質の改善が確認された．このことから，本手法は強雑音環境において聴感的品質および可聴性の向上に有効であることが示された．一方で，子音など時間的に短い発話成分に対する再現性は十分ではなく，微細な時間構造の表現には依然として課題が残されている．

参考文献

[1] S. Latif, et al. Transformers in speech processing: A survey. **arXiv preprint**, 2022. http://arxiv.org/abs/2303.11607v1.

[2] Z. Kong, et al. Speech denoising in the waveform domain with self-attention. In **Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, pp. 7867–7871, 2022.

[3] N. Shao, et al. Cleanmel: Mel-spectrogram enhancement for improving both speech quality and asr. **arXiv preprint**, 2025. https://arxiv.org/abs/2502.20040.

[4] J. Kong, et al. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In **Advances in Neural Information Processing Systems (NeurIPS)**, pp. 17022–17033, 2020.

[5] Antony W. Rix, John G. Beerends, Michael P. Hollier, and Andries P. Hekstra. Perceptual evaluation of speech quality (pesq)—a new method for speech quality assessment. **IEEE Transactions on Audio, Speech, and Language Processing**, Vol. 10, No. 2, pp. 70–82, 2001.

[6] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. **IEEE Transactions on Audio, Speech, and Language Processing**, Vol. 19, No. 7, pp. 2125–2136, 2011.

[7] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee. A regression approach to speech enhancement based on deep neural networks. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, Vol. 23, No. 1, pp. 7–19, 2015.

[8] Felix Weninger, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R. Hershey, and Björn Schuller. Speech enhancement with lstm recurrent neural networks. In **Proceedings of the International**

Conference on Latent Variable Analysis and Signal Separation (LVA/ICA), pp. 91–98, 2015.