6H-05

品質劣化したラジオ音声を対象とした音声強調手法の検討

小林 彰夫 † 安 啓一 †
筑波技術大学 産業技術学部 †

1 はじめに

中波放送（AM 放送）は，低廉かつ安価な受信機を用いて送信所よりはるか遠方でも受信可能であるという特長があり，国内各局の FM 補完放送転換の時流にあっても，災害時のメディアとして位置づけられる．一方，中波放送は振幅変調を採用していることから，デジタル放送はもとより FM 放送と比べても伝播経路における雑音やフェージングといった妨害を受けやすく，受信音声の品質が容易に低下するといった問題がある．このため，受信環境によっては難聴者にとってきわめて聞き取りづらくなるケースが想定される．また，品質の大幅に劣化した音声に対する音声認識は認識率の低下を招くため，情報保障が困難となる．品質の大幅に劣化した音声から，原音声を復元することができれば，難聴者に聞き取りやすい音声を加工したり，音声認識を用いて情報保障が可能となる．そこで本稿では，品質劣化した AM 放送の復調音声に対して，雑音除去およびフェージングの緩和を指向した音声強調手法を適用し，品質改善の検討を行った．

2 ラジオの通信路と品質

わが国のラジオ放送では、振幅変調 (AM) もしくは周波数変調 (FM) 方式が採用されている．AM による中波放送は，FM に比べて変復調回路が単純であり，遠距離まで伝搬することから，特に災害時のメディアとしての強みを持つ．一方，FM に比べると雑音等による品質の低下の影響を受けやすい．本稿で取り上げる振幅変調による通信路の概略を図 1 に示す．振幅変調された信号は，加算性ガウス雑音のあるフェージング通信路を通過し復調される．ただし本稿では簡便のため通信路の伝搬損失は考慮しない．

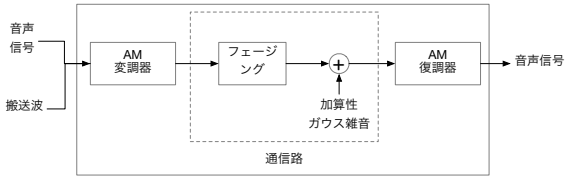

図 1: ラジオの通信路

無線通信では，送信点から発せられた電波が複数の経路をたどって受信点に到達することがある．例えば，建物，地表，電離層などによる反射である．このとき，受信した信号の振幅・位相差によって信号が減衰することがある．この現象をマルチパスフェージングと呼ぶ．

受信点で復調された音声信号は，通信路におけるフェージング・雑音により大きく品質が劣化する．

3 音声強調手法

劣化した信号を復元し品質を向上させる際には，雑音抑圧オートエンコーダーが使われることが多い．多くの手法では，雑音環境下における雑音抑圧や原信号の復元を試みているが，音声のレベルが大きく変動するフェージング環境下における雑音抑圧の研究はあまりない．そこで，本稿では従来法である Denoiser[1] および CleanUNet[2]，また Clean-UNet の自己注意機構に Conformer エンコーダーを採用した音声強調モデル（以下 *modified* CleanUNet）を用いて，フェージング環境下での音声強調の実験を行った．3 種類のモデルはいずれも波形サンプルを入力とするエンコーダー/デコーダー型のニューラルネットワークであり，エンコーダー出力に LSTM を組み合わせたモデルが Denoiser，自己注意機構を組み合わせたモデルが CleanUNet，Conformer を組み合わせたモデルが modified CleanUNet である．

4 実験

4.1 学習・評価データと評価方法

実験では 2 種類の学習データを用意した．

セット A Software defined radio のひとつ GNURadio を用いて図 1 の通信経路を模擬し，擬似的に品質劣化した受信音声を生成した [3]．擬似音声の原音には新聞記事読み上げコーパス (JNAS) を用い，合計 100 時間（55.7k 文）の音声を生成して学習データとした．音声データのサンプリング周波数は 16kHz である．

A Study on Speech Enhancement Methods for Degraded Radio Signals
†Akio KOBAYASHI, Keiichi YASU
Faculty of Industrial Technology, Tsukuba University of Technology

表 1: セット A で学習したモデルによる性能評価

	PESQ	STOI	CER(%)
w/o SE	1.084	0.558	**18.54**
Denoiser	**1.128**	0.582	31.28
CleanUNet	1.118	**0.593**	36.66
modified	1.126	0.585	36.87

表 2: セット B で学習したモデルによる性能評価

	PESQ	STOI	CER(%)
Denoiser	2.281	0.837	13.28
CleanUNet	2.443	0.846	**12.70**
modified	**2.491**	**0.849**	13.02

表 3: セット B のデータサイズと性能の関係

データ量 (時間)	PESQ	STOI	CER(%)
1h	1.379	0.729	25.43
5h	1.679	0.777	21.45
10h	1.841	0.798	17.91

セット B　筆者の研究室にて NHK ラジオ第 1 放送 (594kHz) のニュース番組を対象として 75 時間収録して学習データとした．ただし，アンテナと受信機との間に計算機を設置することにより雑音を混入させ、音声品質を故意に低下させた．また，中波放送の帯域は 7.5kHz であることから，16kHz のサンプリング周波数を用いてラジオ音声を収集した．

リファレンスは，NHK ネットラジオらじる★らじるから音声を収集し，16kHz にダウンサンプリングしたデータを用いた．また，ラジオ音声とネットラジオ音声とのラグを計算することによりアラインメントを行った．学習はセット A のみを用いた学習ののち，セット B を用いて追加学習を行った．

評価データは，セット B と同様の環境で収録した NHK ラジオ第 1 放送の 471 発話とした．ただし，セット B に対して発話内容はオープンである一方，受信環境はクローズドであることに注意する．評価は客観評価指標である PESQ および STOI を用いた．また，強調後の音声に対して Whisper(large)[4] による音声認識を行い，文字誤り率による性能評価を行った．

4.2　実験結果

セット A を学習データとしたモデルの性能を比較すると (表 1)，擬似データでは PESQ，STOI の両指標とも音声強調による改善がわずかであった．また，音声認識結果はオリジナルの劣化したラジオ音声に対する結果からの改善がみれらなかった．図 2 からわかるように，擬似データによるモデルではスペクトログラム中のノイズが除去できていない．

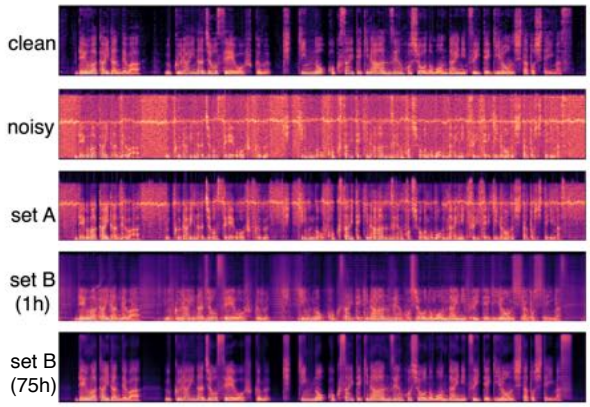

図 2: modified CleanUNet による音声強調結果

セット B を用いて追加学習したところ (表 2)，PESQ，STOI では modified CleanUNet が最も高い値となった．Whisper による音声認識結果は CleanUNet が最も性能が高かった．セット B の学習データ量を変えて modified CleanUNet を学習したところ (表 3)，音声認識性能でオリジナルのラジオ音声の結果を超えるのはデータ量が 10 時間の場合であった．

5　おわりに

品質劣化したラジオ音声に対する音声強調を試みた．今後の課題は，未知の受信環境に対して頑健な音声強調を行うことである．

謝辞

本研究の一部は JSPS 科研費 20H01716，21H00901 の助成を受けたものである．

参考文献

[1] A. Defossez et al., “Real Time Speech Enhancement in the Waveform Domain,” in *Proc. Interspeech*, 2020, pp. 3291–3295.

[2] Z. Kong et al., “Speech Denoising in the Waveform Domain with Self-Attention,” in *Proc. ICASSP*, 2022, pp. 7867–7871.

[3] A. Kobayashi, “Speech Enhancement for Demodulated Signals under Multipath Fading Communication Channels”, in *Proc. APSIPA*, 2020, pp. 460-464.

[4] A. Radford et al., “Robust Speech Recognition via Large-Scale Weak Supervision”, *arXiv:2212.04356*, 2022.