人工知能学会第二種研究会資料
AIチャレンジ研究会
SIG-Challenge-063-02

Parallel Adapter Model と Near-Identity 初期化を用いた音声認識の雑音耐性向上

Improving Noise Robustness of Automatic Speech Recognition based on a Parallel Adapter Model with Near-Identity Initialization

大﨑 崇博[1] *　周藤 唯[2]　糸山 克寿[1,2]　西田 健次[1]　中臺 一博[1]

[1] 東京工業大学
[1] Tokyo Institute of Technology
[2] (株) ホンダ・リサーチ・インスティチュート・ジャパン
[2] Honda Research Institute Japan, Co. Ltd.

Abstract: 本論文では, 音声認識の雑音耐性を少量のトレーニングで改善するために Parallel Adappter Model (PAM) を提案する．音声認識の雑音耐性を向上するために音声強調を使用すると，強調音声による歪みや目的音の情報除去により，認識精度が十分に向上しない．提案手法の PAM では，学習済み音声認識と音声強調に加え，アダプターと呼ばれる小規模ネットワークを適切な初期化手法でモデルに組み込み少量の再学習を行うことで，音声強調と音声認識の親和性を高め，認識精度を向上する．検証実験では，わずか 10 epoch の再学習で，音声認識の精度が 26.1 ポイント向上した．

1　はじめに

音声認識（Automatic Speech Recognition, ASR）は, 人間の発話音声をテキストに変換するシステムである．これまでの音声認識は，GMM(Gaussian mixture model) と HMM(Hidden Malkov model) を組み合わせた GMM-HMM モデル [1] や，GMM を DNN(Deep neural network) に置き換えた DNN-HMM モデル [2, 3] が主流であった．しかし，計算機の性能向上や大規模言語コーパスの拡充に伴い，近年では音響モデルと言語モデルを 1 つのネットワークで構成した End-to-end ASR が盛んに研究されている [4, 5, 6, 7]．このような ASR は，一般的に雑音が少ない環境で録音された教師用音声が用いられる．それに対して，ASR を実環境で用いる場合，入力音声には目的音声に加えて背景ノイズが混ざるため, 性能が低下する．特に，工場内やレストラン等の騒がしい場所で ASR を用いる場合には，音声認識の雑音耐性を向上する手法が必要不可欠となる．

ASR の雑音耐性を高める研究はこれまで行われており，代表的なものに音声強調 (Speech Enhancement, SE) を ASR フロントエンドとして用いる手法がある．SE はノイズが混ざった音声から，ノイズのパワーを弱め目的音声のみを強調する技術である．SE 手法にはマイクロホンアレイ処理 [8, 9] や深層学習 [10, 11] を用いたものがあり，近年では単一チャネル入力への適用が可能である深層学習ベースの手法が広く用いられている [12, 13]．しかし，音声強調がノイズ音だけではなく話し声の情報まで除去することや，出力音声に歪みが生じたりすることが原因となり，SE と ASR の単純な組み合わせでは，期待した性能向上が得られないことが知られている．

この歪みによる ASR と SE 間のミスマッチを解消するには 2 つのアプローチがある．1 つ目は，再トレーニングを行わない手法である．Takeda らの研究 [14, 15] では，モンテカルロ推定に基づく特徴量推定により，再学習なしで認識精度を向上している．しかし，推定には，潜在特徴量の繰返し計算が必要であり，時間を要する割に性能向上が小さい．2 つ目は，モデルの再トレーニングによるミスマッチの緩和である．モデル全体を再学習する手法や SE 部のみを再学習する手法が提案されている [16, 17, 18] が，いずれの手法も膨大なパラメータ更新が必要であり，学習コストが大きい．

そこで本研究では，アダプターを使用した少量の再トレーニング手法を提案し，ミスマッチ問題と再学習コスト問題を解決する．アダプターは小規模のネットワークであり，学習済みのモデルと同時に用いる．アダプターが組み込まれたモデルを学習する際は，ASR や SE のパラメータは固定し，アダプターのパラメータのみ更新することで，更新パラメータ数を抑えることができる．また，アダプター学習にはノイズを付与

*連絡先：東京工業大学
〒 152-8552 東京都目黒区大岡山 2-12-1
osaki@ra.sc.e.titech.ac.jp

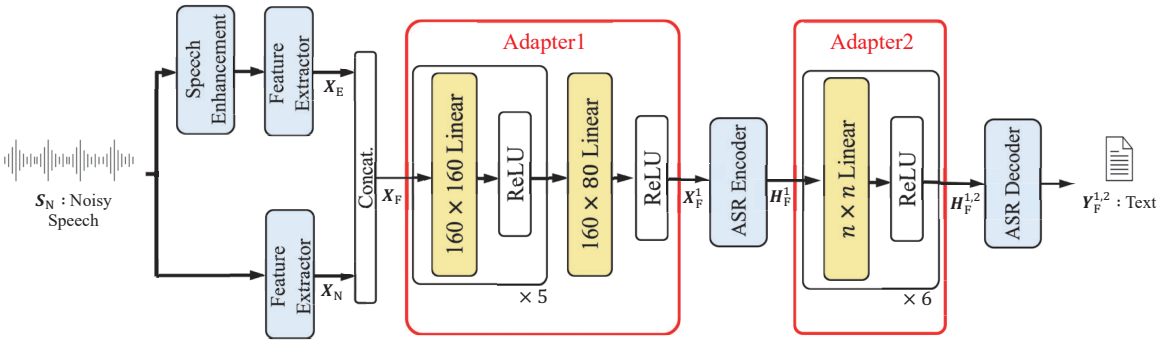


図 1: Parallel Adapter Model (PAM)．Adapter2 の次元数 n は潜在特徴量の次元数に準拠し，100 時間の学習用音声を用いる場合は 256，960 時間の学習用音声を用いる場合は 512 に設定．

した音声を用いることで，ノイズ耐性の高いモデルが学習できる．以上より，少ない再学習量でモデルの耐雑音性向上が期待できる．

2 提案手法

本章では，提案手法である Parallel Adapter Model (PAM) とその初期化手法を，ASR や SE と共に説明する．

2.1 音声認識 (ASR)，音声強調 (SE)

ASR には，Encoder, Decoder からなる End-to-End 音声認識モデルを用いるものとする．ASR への音声入力を $\boldsymbol{S}$ とすると，まず，$\boldsymbol{S}$ を特徴抽出器 (Feature Extractor, FE) によって音声特徴量 $\boldsymbol{X}$ に変換する．次に，変換した音声特徴量 $\boldsymbol{X}$ を Encoder 部に送出し，潜在特徴量 $\boldsymbol{H}$ に変換，その後，Decoder 部で，テキストに変換する．音声認識モデル学習では，音声 $\boldsymbol{S}$ が入力されたときに正しいテキスト $\boldsymbol{Y}$ が出力される確率が最大になるように，Encoder, Decoder のパラメータ $\boldsymbol{\theta}_{enc}, \boldsymbol{\theta}_{dec}$ の推定を行う．

SE は，雑音が混入した音声 $\boldsymbol{S}_{\mathrm{N}}$ から，雑音を抑圧した音声 $\boldsymbol{S}_{\mathrm{E}}$ に変換する．本稿では，入力音声が単チャネルであることを想定しているため，深層学習ベースの SE 手法を用いる．SE のパラメータ $\boldsymbol{\theta}_{\mathrm{se}}$ は，ASR と組み合わせて用いる前に，クリーン音声と強調音声の類似度が高くなるように事前学習を行うものとする．

2.2 Parallel Adapter Model

図 1 に，提案モデルである Parallel Adapter Model (PAM) を示す．このモデルは，SE を適用するフローに加えて，SE を行わない未処理音声を用いるフローを持つ．それぞれのアダプターは，学習可能なパラメータ $\boldsymbol{\theta}_{\mathrm{adp1}}$，$\boldsymbol{\theta}_{\mathrm{adp2}}$ を持った，ASR や SE よりも十分に小さいネットワークである．

アダプター 1 への入力 $\mathrm{Cat}(\boldsymbol{X}_E, \boldsymbol{X}_N)$ は，強調音声と未処理音声の特徴量を結合した特徴量であり，$\boldsymbol{X}_{\mathrm{F}}^1$ が出力される．

$$\boldsymbol{X}_{\mathrm{E}} = \mathrm{FE}(\boldsymbol{S}_{\mathrm{E}}) \quad (1)$$

$$\boldsymbol{X}_{\mathrm{N}} = \mathrm{FE}(\boldsymbol{S}_{\mathrm{N}}) \quad (2)$$

$$\boldsymbol{X}_{\mathrm{F}}^1 = \mathrm{Adapter1}\left(\mathrm{Cat}(\boldsymbol{X}_E, \boldsymbol{X}_N); \boldsymbol{\theta}_{\mathrm{adp1}}\right) \quad (3)$$

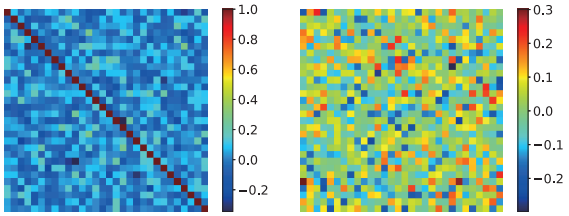
アダプター 2 への入力は Encoder の出力 $\boldsymbol{H}_{\mathrm{F}}^1$ であり，Decoder への入力 $\boldsymbol{H}_{\mathrm{F}}^{1,2}$ を出力する．

$$\boldsymbol{H}_{\mathrm{F}}^{1,2} = \mathrm{Adapter2}\left(\boldsymbol{H}_{\mathrm{F}}^1; \boldsymbol{\theta}_{\mathrm{adp2}}\right) \quad (4)$$

モデルを学習する際は，アダプターのパラメータのみ更新し，ASR や SE のパラメータは固定した状態で，認識精度が最も良くなるようなアダプターのパラメータを学習する．

PAM を構築する上で，アダプターには 3 つの要件，すなわち「選択」「復元」「混合」が求められる．「選択」は，音響特徴量から音声成分を抽出し，雑音成分を除去するとともに，SE による歪みを緩和し，ミスマッチ問題に対処する機能である．「復元」は，SE によって過度に欠落した情報を復元する機能である．例えば，倍音成分同士が持つ強い相関関係を基に，アダプターが損なわれた特徴量成分を復元できると考えられる．「混合」は，アダプター 2 に必要とされる，2 つの特徴量から認識に適した特徴量を生成することで，ASR の性能向上を実現する機能である．

これらのうち，「復元」と「混合」は線形層で，「選択」は活性化関数として ReLU を用いることで実現する．またアダプターの表現力を向上させるために，本稿では，ReLU 関数をもつ線形層を 6 層スタックすることで，アダプターを形成する (図 1)．これにより，軽量なモデルでアダプターを表現できる．アダプター 1



(a) near-identity 初期化　(b) ランダム 初期化

図 2: アダプターの初期化方法

により，SE と ASR のミスマッチ問題を緩和するとともに，SE によって過度に失われる音声情報を未処理音声や他の周波数帯の音声情報で補うことが期待できる．アダプター 2 により，アダプター 1 で緩和しきれなかった雑音や歪み成分に由来する潜在特徴量のミスマッチが緩和されることが期待される．

2.3 Near-identity 初期化 (NI 初期化)

一般的に，ニューラルネットワークのパラメータはランダム初期化を行うことが多い (図 2(b)) が，パラメータ値が最適値と大きく異なるため，学習に時間がかかる傾向がある．この問題に対し，アダプターが恒等写像に近くなるような near-Identity 初期化 (NI 初期化) が有効であることが報告されている [19]．そこで，NI 初期化をアダプターに適用する．この場合，図 2(a) に示すように，線形層の対角成分を 1 に，他の重みとバイアスを 0 に近い値とすることで，アダプタの入力と出力がほぼ同じになるように初期化することができる．なお，アダプター 1 の 最終線形層の NI 初期化では，未処理音声特徴量と強調音声特徴量の平均が出力となるように，対角成分が 0.5 の対角行列を 2 つ横に並べた値で初期化する．

3 実験設定

本章では，提案モデルである PAM と NI 初期化の有効性を確認するために，LibriSpeech 音声コーパス [20] を用いたモデル学習を行い，認識性能に関する比較実験を行う．アダプター層の学習では，train-clean-100 subset（100 時間）を音声を用いる実験と，train-clean-100, train-clean-360, train-other-500（合計 960 時間）の音声を用いる実験を行う．学習を行う際には，自動車製造工場で録音した雑音を SNR (信号対雑音比) が 0 dB になるよう足し合わせた学習データを作成し，25 epoch 分の学習を行った．なお，使用した雑音は，定常的な低周波成分と不定期な高周波を含む機械音が含まれる．評価用音声には，LibriSpeech データセットの dev-clean と test-clean を用いた．雑音ロバスト性を評価するために，これらの評価音声に上記工場で別途収録した雑音を SNR が，15 dB，10 dB，5 dB, 0 dB になるように足し合わせたものを用意し，評価用音声とした．

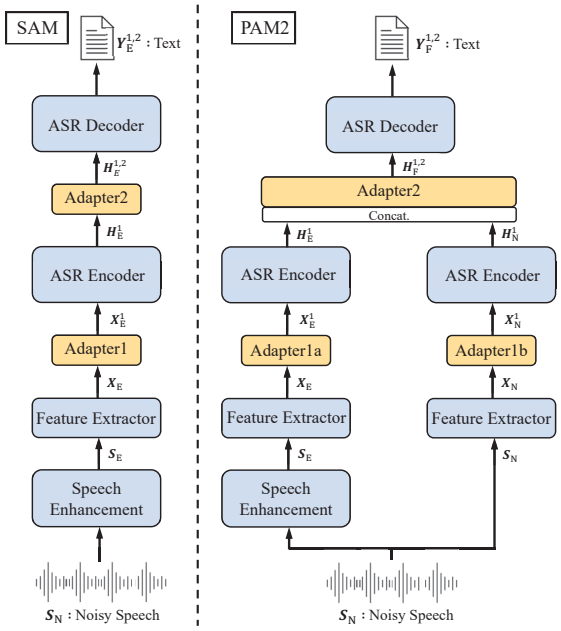


図 3: 比較モデル ; SAM,PAM2

3.1 比較モデル

PAM の有効性を確かめるために，他の 2 種のアダプターモデルである，Serial Adapter Model (SAM) と，Parallel Adapter Model 2 (PAM2) を比較に用いた．図 3 にこれらの比較モデルの構造を示す．SAM では，SE と ASR の直列モデルに対して，2 箇所にアダプターを挿入する．このモデルは他のモデルに比べて軽量であるが，未処理音声を推論に用いないモデルである．PAM2 は，提案手法である PAM と同様に，SE を適用しない音声を用いる．2 つの音響特徴量を，別々にアダプタ 1a とアダプタ 1b に送出し，共通の学習済みパラメータをもつエンコーダーによって潜在特徴量に変換し，これらの特徴量をアダプター 2 によって結合・混合する．PAM2 の Adapter2 の構造は，図 1 の Adapter1 と同様に，入力層は潜在特徴量の 2 倍の次元数を持ち，最後の線形層でもとの次元に戻される．

また，これらの比較モデルとは別に，PAM において片方のアダプターのみを用いる場合の性能も検証する．

3.2 モデル設定

入力音声は，16 bit, 16 kHz サンプリングの信号であり，FE は，これを窓長 512，シフト長 160 でフレーム化したのち，80 次元のメルフィルタバンクを適用した結果得られる音響特徴量を出力する．

表 1: 100 時間の教師用音声を用いてモデル事前学習・アダプター学習を行った場合の音声認識精度 (WER). 「H」は「hybrid CTC/Attention ASR [4]」, 「C」は「Conv-Tasnet [12]」, 「N」は「NI 初期化」, 「R」は「ランダム初期化」, 「Tunable Param.」は, アダプターのパラメータ数の合計を表す.

No.	Model				Tunable	clean		15dB		10dB		5dB		0dB		Ave.
	ASR	SE	Type	Init.	Param.	dev	test	dev	test	dev	test	dev	test	dev	test	
1	H	-	-	-	-	**8.3**	**8.6**	15.8	15.0	29.4	26.6	57.4	52.6	85.5	81.1	38.0
2	H	C	-	-	-	9.0	9.2	14.4	13.6	21.8	19.5	38.2	33.9	68.9	63.6	29.2
3	H	C	Takeda [14]	-	-	9.3	9.5	14.3	13.8	22.1	20.1	39.1	34.6	68.0	62.4	29.3
4	H	C	PAM (Proposed)	N	536K	12.3	12.0	**13.1**	**12.9**	**15.9**	15.7	**23.8**	**21.7**	**42.8**	**38.6**	**20.9**
5			PAM	R		25.2	23.6	18.9	18.3	22.8	21.7	32.2	30.2	54.3	50.4	29.8
6	H	C	SAM	N	434K	11.1	11.2	14.5	14.2	19.1	17.9	30.9	27.8	55.5	50.3	25.3
7			SAM	R		53.7	53.0	54.2	53.8	58.2	57.4	67.6	66.0	82.9	80.4	62.7
8	H	C	PAM2	N	1.52M	11.6	11.5	**13.1**	**12.9**	16.0	**15.6**	24.2	24.2	44.6	44.6	21.8
9			PAM2	R		92.9	93.0	87.9	88.5	88.3	88.3	90.5	89.9	95.6	94.9	91.0

表 2: 960 時間の教師用音声を用いてモデル事前学習・アダプター学習を行った場合の音声認識精度 (WER). 「H」は「hybrid CTC/Attention ASR [4]」, 「C」は「Conv-Tasnet [12]」, 「N」は「NI 初期化」, 「R」は「ランダム初期化」, 「Tunable Param.」は, アダプターのパラメータ数の合計を表す.

No.	Model			Tunable	clean		15dB		10dB		5dB		0dB		Ave.
	ASR	SE	Type	Param.	dev	test	dev	test	dev	test	dev	test	dev	test	
10	H	-	-	-	**2.6**	**2.6**	**3.8**	**3.5**	7.0	6.2	24.8	20.2	69.1	62.3	20.2
11	H	C	-	-	3.8	4.0	5.2	4.8	7.6	6.9	15.1	12.9	36.9	31.9	12.9
12	H	C	Takeda [14]	-	2.9	2.9	4.3	4.0	6.6	6.0	15.7	13.1	43.9	37.4	13.7
13	H	C	PAM (Proposed)	1.72M	4.2	4.3	4.6	4.6	**6.0**	**5.7**	**9.9**	**8.9**	**24.2**	**20.3**	**9.3**
14	H	C	SAM	1.61M	3.6	3.6	5.1	4.6	7.3	6.6	14.9	12.6	36.3	31.7	12.6
15	H	C	PAM2	5.85M	4.5	4.6	5.0	5.2	6.5	6.3	10.5	9.6	26.3	22.1	10.1

ASR には, ESPnet [21], の hybrid CTC/attention モデル [4] を用いる. このモデルは, 12 個の Conformer ブロックで構成されるエンコーダーと, 6 層の Transformer ブロックと線形層で構成されるデコーダーをもつ. エンコーダーが出力する潜在特徴量は, 100 時間の音声で学習する場合は 256 次元に, 960 時間の音声で学習を行う際には 512 次元にした. 損失関数には, CTC 損失と Attention デコーダーの損失の重み付き和を用い, 重みはそれぞれ 0.3, 0.7 である.

SE には, 深層学習ベースの手法である Conv-TasNet [12] を用いた. このモデルは, エンコーダー, セパレーター, デコーダーから構成される. エンコーダーは 1 層の畳み込み層, デコーダーは 1 層の転置畳み込み層からなる. セパレーターはマスク推定用のネットワークで, それぞれ 8 個の畳み込みブロックを含む, 4 個の Temporal Convolutional Network [22] からなる. SE は SI-SNR [23] を最大化するように, CHiME-4 データセット [24] で事前学習されたモデルを用いた.

PAM と PAM2 の学習では SpecAug を用いず, SAM の学習では SpecAug を用いた. NI 初期化を用いる場合, 対角成分は 1, それ以外の重みとバイアスは平均 0, 標準偏差 0.01 の正規分布で初期化した. またランダム初期化を用いる場合, 線形層の入力次元数を m とすると, 一様分布 $[-\sqrt{1/m}, \sqrt{1/m}]$ で初期化を行った.

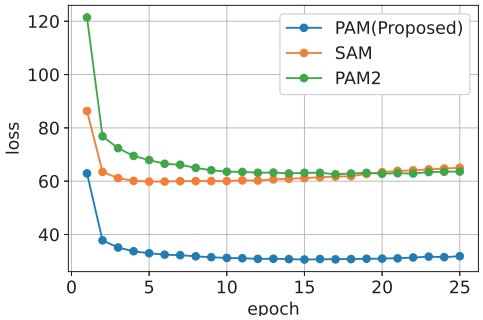


図 4: PAM,SAM,PAM2 の学習損失推移

4 実験結果

本章では, 実験結果について, 初期化手法, モデルの妥当性, アダプターの寄与の 3 点から考察する.

4.1 初期化手法について

表 1 に, 100 時間の教師用音声を用いてアダプター学習を行った結果を示す. これより, NI 初期化を用いると, 音声強調のみを用いた場合 (No. 2, 3) に比べて, すべてのアダプターモデル (No. 4, 6, 8) で雑音環境での認識性能が向上した. NI 初期化を用いることで, もとのモデルの性質を大きく崩さない初期値から学習が

表 3: 100 時間の教師用音声を用いてモデル事前学習・アダプター学習を行った場合の音声認識精度 (WER).「H」は「hybrid CTC/Attention ASR [4]」,「C」は「Conv-Tasnet [12]」,「Adapter」ではそれぞれのアダプターの使用有無を示している．No. 16, 19 はそれぞれ表 1 の No. 2, 4 と同一の実験結果である．

No.	Model			Adapter		clean		15dB		10dB		5dB		0dB		Ave.
	ASR	SE	Type	1	2	dev	test	dev	test	dev	test	dev	test	dev	test	
16	H	C	-	-	-	**9.0**	**9.2**	14.4	13.6	21.8	19.5	38.2	33.9	68.9	63.6	29.2
17	H	C	PAM (Proposed)	✓	-	13.0	12.9	14.2	13.8	17.1	16.9	26.3	23.7	47.0	42.4	22.7
18				-	✓	10.5	10.8	14.0	13.8	19.3	18.2	32.1	28.9	58.8	53.1	26.0
19				✓	✓	12.3	12.0	**13.1**	**12.9**	**15.9**	**15.7**	**23.8**	**21.7**	**42.8**	**38.6**	**20.9**

開始でき，少ない学習量でも，パラメータが良好に学習できていることが確認できた．その一方で，ランダム初期化を用いた場合，性能が向上するケース (No. 5) と，劣化するケース (No. 7, 9) が見られた．ランダム初期化を用いて性能が向上したケースでも，NI 初期化を用いると，さらに性能が向上した．このため，ランダム初期化は，学習不足で十分な最適化ができていないことが推測される．以上の結果より，アダプター学習における NI 初期化の有効性が示された．以降では，NI 初期化を用いる場合のみ考える．

4.2 モデルごとの性能について

表 1 より，提案手法である PAM が，ほとんどの雑音環境での認識性能，および平均性能で最も良好であるという結果となった．SNR が 0 dB のとき，SE のみを適用した場合に比べて，dev-clean は 26.1 ポイント，test-clean は 25.0 ポイントの性能向上が確認できた．さらに，平均性能では，8.3 ポイント性能が向上した．SAM では，どのケースも PAM に比べて明らかに性能が低いことがわかる．PAM2 については，15 dB や 10 dB の雑音環境において PAM と同等の性能が確認できた．一方で，高雑音環境では PAM の方が性能が高い．更新パラメータ数では，PAM のほうが少ないことを考慮すれば，PAM がより優秀なモデルであると言える．また，図 4 に学習損失の推移を示す．これより，ほとんどのモデルで，10 epoch 程度で学習損失が収束しており，その中でも PAM が最も学習損失が小さい結果となっている．学習損失の点からも，PAM 有効性が確認できた．

表 2 は，960 時間の教師用音声を用いてアダプター学習を行った結果を示している．こちらの実験でも，提案手法である PAM が，雑音の多い環境での認識性能，および平均性能で最も良好であるという結果となった．SNR が，0 dB のとき，SE のみを適用した場合に比べて，dev-clean は 12.7 ポイント，test-clean は 11.6 ポイントの性能向上が確認できた．さらに，平均性能では，3.6 ポイント性能が向上した．

4.3 アダプターの性能への寄与について

表 3 より，SE のみ用いた場合に比べ，いずれかのアダプターを用いた場合にも性能の向上が確認された．Adapter1 と Adapter2 のそれぞれ一方のみを用いた場合 (No. 17, 18) では，Adapter1 のみを用いたほう (No. 17) が大きな性能向上がみられた．Adapter1 のみ用いた場合，SNR が 0 dB の環境において，dev-clean では 21.9 ポイント，test-clean では 21.2 ポイントの向上であったのに対し，Adapter2 のみ用いた場合，dev-clean では 10.1 ポイント，test-clean では 10.5 ポイントの向上であった．また平均性能も，Adapter2 よりも Adapter1 のほうが 3.3 ポイント上回っている．Adapter1 では音響特徴量を混合しており，特徴量の変換がより有効に機能したと考えられる．

その一方で，アダプターを 2 つ用いる PAM(No. 19) では，さらに精度が向上している．以上より，2 つのアダプターが性能に寄与し，音声認識性能を改善していることが確認された．

5 むすび

本研究では，音声認識の雑音耐性を向上するために，音声強調を音声認識のフロントエンドとして用いる場合に問題となる音声認識と音声強調のミスマッチ問題を緩和するために，Parallel Adapter Model とその初期化に Near-Identity 初期化法を利用することを提案した．Parallel Adapter Model は，強調音声に加えて，未処理音声を用いてミスマッチ問題の緩和をはかる一つ目のアダプターと，一つ目のアダプターで緩和しきれなかった潜在特徴量に残存するミスマッチを緩和する二つ目のアダプターの 2 つのアダプターを用いる手法である．また，Near-Identity 初期化法は，恒等写像に近い値で初期化することで，効率的な学習を可能にする手法である．実験の結果，train-clean-100 を用い提案手法で学習を行ったモデルは，アダプターを用いない場合と比較して，性能が最大で 25.2 ポイント，平均で 8.4 ポイント性能が向上した．また，PAM においては，2 つのアダプターを用いることで性能がより

向上することが確認された．今後の課題として，アダプター構造のさらなる検討や，他の ASR/SE モデルへの提案手法の適用があげられる．

謝辞

本研究は JSPS 科研費 JP19KK0260, JP20H00475 および JP23K11160 の助成を受けた．

参考文献

[1] B. H. Juang and L. R. Rabiner. Hidden Markov models for speech recognition. *Technometrics*, Vol. 33, pp. 251–272, 1991.

[2] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, Vol. 29, No. 6, pp. 82–97, 2012.

[3] Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. Hybrid speech recognition with deep bidirectional lstm. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 273–278, 2013.

[4] Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R. Hershey, and Tomoki Hayashi. Hybrid CTC/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, Vol. 11, No. 8, pp. 1240–1253, 2017.

[5] Yui Sudo, Muhammad Shakeel, Brian Yan, Jiatong Shi, and Shinji Watanabe. 4D ASR: Joint modeling of CTC, attention, transducer, and mask-predict decoders. In *in Proc. Interspeech*.

[6] Yifan Peng, Jinchuan Tian, Brian Yan, Dan Berrebbi, Xuankai Chang, Xinjian Li, Jiatong Shi, Siddhant Arora, William Chen, Roshan Sharma, et al. Reproducing whisper-style training using an open-source toolkit and publicly available data. *arXiv preprint arXiv:2309.13876*, 2023.

[7] Linhao Dong, Shuang Xu, and Bo Xu. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5884–5888, 2018.

[8] Takuya Higuchi, Nobutaka Ito, Takuya Yoshioka, and Tomohiro Nakatani. Robust mvdr beamforming using time-frequency masks for online/offline asr in noise. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5210–5214, 2016.

[9] Jahn Heymann, Lukas Drude, Christoph Boeddeker, Patrick Hanebrink, and Reinhold Haeb-Umbach. Beamnet: End-to-end training of a beamformer-supported multi-channel asr system. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5325–5329, 2017.

[10] Zhong-Qiu Wang, Peidong Wang, and DeLiang Wang. Complex spectral mapping for single- and multi-channel speech enhancement and robust asr. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 28, pp. 1778–1787, 2020.

[11] Panagiotis Tzirakis, Anurag Kumar, and Jacob Donley. Multi-channel speech enhancement using graph neural networks. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3415–3419, 2021.

[12] Yi Luo and Nima Mesgarani. Conv-tasnet: Surpassing ideal time – frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 27, No. 8, pp. 1256–1266, 2019.

[13] Julius Richter, Simon Welker, Jean-Marie Lemercier, Bunlong Lay, and Timo Gerkmann. Speech enhancement and dereverberation with diffusion-based generative models, 2022.

[14] Ryu Takeda, Yui Sudo, Kazuhiro Nakadai, and Kazunori Komatani. Empirical Sampling from Latent Utterance-wise Evidence Model for Missing Data ASR based on Neural Encoder-Decoder Model. In *Proc. Interspeech*, pp. 3789–3793, 2022.

[15] Ryu Takeda, Yui Sudo, and Kazunori Komatani. Flexible Evidence Model to Reduce Uncertainty Mismatch Between Speech Enhancement and ASR Based on Encoder-Decoder Architecture. In *Proc. APSIPA*, 2023.

[16] Jisi Zhang, Catalin Zorila, Rama Doddipatla, and Jon Barker. On monoaural speech enhancement for automatic recognition of real noisy speech using mixture invariant training. In *Interspeech 2022*. ISCA, sep 2022.

[17] Xuankai Chang, Takashi Maekaku, Yuya Fujita, and Shinji Watanabe. End-to-end integration of speech recognition, speech enhancement, and self-supervised learning representation, 2022.

[18] Yuchen Hu, Nana Hou, Chen Chen, and Eng Siong Chng. Dual-Path Style Learning for End-to-End Noise-Robust Speech Recognition. In *Proc. INTERSPEECH 2023*, pp. 2918–2922, 2023.

[19] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 09–15 Jun 2019.

[20] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015.

[21] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. Espnet: End-to-end speech processing toolkit, 2018.

[22] Colin Lea, René Vidal, Austin Reiter, and Gregory D. Hager. Temporal convolutional networks: A unified approach to action segmentation. In Gang Hua and Hervé Jégou, editors, *Computer Vision – ECCV 2016 Workshops*, pp. 47–54, Cham, 2016. Springer International Publishing.

[23] Yi Luo and Nima Mesgarani. Tasnet: Time-domain audio separation network for real-time, single-channel speech separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 696–700, 2018.

[24] Emmanuel Vincent, Shinji Watanabe, Aditya Arie Nugraha, Jon Barker, and Ricard Marxer. An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Computer Speech & Language*, Vol. 46, pp. 535–557, 2017.