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-066-01

音声強調と雑音特徴量を用いた音声認識の雑音耐性向上

Improving Noise Robustness of Automatic Speech Recognition with Speech Enhancement and Noise Feature Extractor

大﨑 崇博[1]*　周藤 唯[2]　中臺 一博[1]

[1] 東京科学大学
[1] Institute of Science Tokyo
[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: 本稿では，音声認識 (Automatis Speech Recognition, ASR) における雑音耐性の改善について述べる．ASR の性能は雑音環境下で低下するため，音声強調（Speech Enhancement, SE）をフロントエンド処理として採用することが多いが，SE と ASR の間にミスマッチ問題が発生してしまう．この問題を緩和するためには，SE と ASR を含むモデル全体を再学習する必要があり，時間のかかるプロセスである．そこで我々は，雑音特徴抽出器を用いて，その出力を音声特徴変換の重みとして用いる「雑音特徴 Biasing モデル」を提案する．提案手法により，わずか 10 エポックで ASR 性能を 11.7 ポイント向上することができ，提案手法の有効性を示すことができた．

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．深層学習 (Deep Newral Network, DNN) を用いて実装される ASR の学習には，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 [1, 2] や深層学習 [3, 4] 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 [5, 6]．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．

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

本稿では，このような ASR と SE 間のミスマッチ問題を解決し，雑音環境下での音声認識性能を向上することを目的とする．

2　関連研究

ASR と SE 間のミスマッチを解消する手法は，これまで主に 2 つの方向で研究されてきた．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 [7, 8] では，モンテカルロ推定に基づく特徴量推定により，再学習なしで認識性能を向上している．しかし，文字推定には潜在特徴量の繰り返し計算が必要で，計算コストが高い割に性能向上が限定的である．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 [9, 10, 11] 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．

これらの問題を解決するため，われわれは小規模なニューラルネットワークである「アダプター」を用いた雑音適応手法「Parallel Adapter Model」を提案し，アダプター学習の性能と収束速度を向上させる「Near-Identity 初期化」を併用することで，少量のパラメータ再学習で ASR の雑音耐性を向上させることに成功した [12, 13]．またこの手法では，音声強調を適用した音声だけではなく，適用していない音声も用いる．これにより，SE によって削除された情報の補間を行い，強

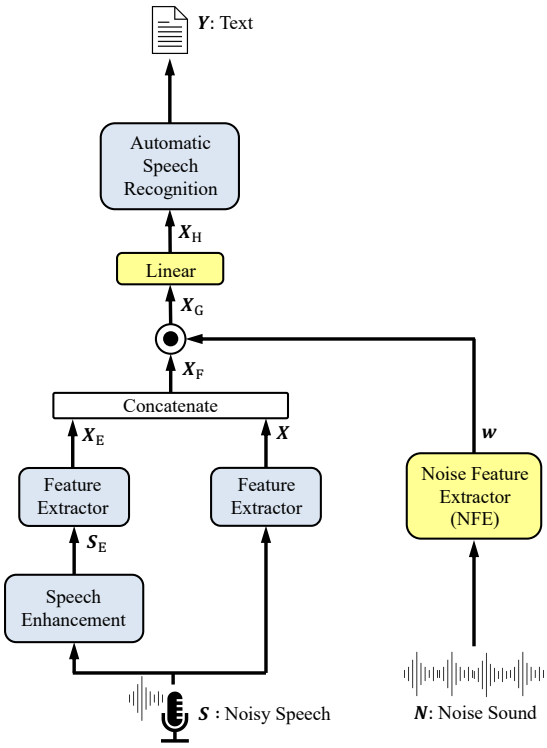

図 1: 雑音特徴 Biasing モデル

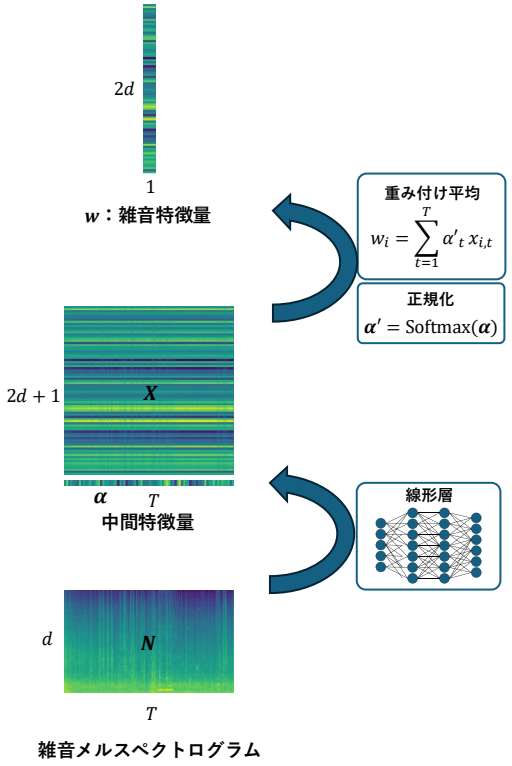

図 2: NFE で行われる雑音特徴量生成

調音声の歪みや過抑制を低減することができた．しかし，雑音適用を行うたびにモデルの再学習を必要とするため，音声認識を使用するまでに未だ数時間に及ぶ学習時間を要することが課題であった．

そこで本稿では，環境雑音を事前に録音し雑音特徴量を得ることで，新たに学習を行うことなく雑音性能を高める手法を提案する．

3 提案手法

本節では，提案手法である雑音特徴抽出器（Noise Feature Extractor, NFE），ならびに雑音特徴量を用いた ASR ネットワークである雑音特徴 Biasing モデルについて説明する．

3.1 雑音特徴 Biasing モデルの構造

図 1 に提案モデルの構造を示す．このモデルは，入力音声 $\boldsymbol{S}$ と，その音声に SE を適用した $\boldsymbol{S}_E$ を用いる．これは，強調音声の特徴量だけでなく，未処理音声の特徴量も利用することで，SE によって過剰に削除された音声特徴を補うことを目的としている．

モデルには 2 つの音声特徴抽出器 (Feature Extractor, FE) を使用し，それぞれ d 次元のメルフィルタバンクを適用して特徴量を抽出する．これにより，未処理音声から得られる特徴量 $\boldsymbol{X}$ と，SE 後の音声から得られる特徴量 $\boldsymbol{X}_E$ を生成する．次に，Concatenate 層で $\boldsymbol{X}$ と $\boldsymbol{X}_E$ を結合し，$2d$ 次元の混合音声特徴量 $\boldsymbol{X}_F$ を作成する．

その後，混合特徴量 $\boldsymbol{X}_F$ に対して，雑音特徴抽出器 (Noise Feature Extractor, NFE) により生成された雑音特徴量 $\boldsymbol{w}$ と $\boldsymbol{X}_F$ のアダマール積を計算し，新たな特徴量 $\boldsymbol{X}_G$ を生成する．

$$\boldsymbol{X}_G = \boldsymbol{X}_F \odot \boldsymbol{w} = \begin{bmatrix} X_{F,1,1}\cdot w_1 & X_{F,1,2}\cdot w_1 & \cdots & X_{F,1,T}\cdot w_1 \\ X_{F,2,1}\cdot w_2 & X_{F,2,2}\cdot w_2 & \cdots & X_{F,2,T}\cdot w_2 \\ \vdots & \vdots & \ddots & \vdots \\ X_{F,N,1}\cdot w_N & X_{F,N,2}\cdot w_N & \cdots & X_{F,N,T}\cdot w_N \end{bmatrix}$$

ここで，$\boldsymbol{w}$ は $2d$ 次元の時不変ベクトルであり，雑音の特徴を表現している（詳細は 3.2 節）．

最後に，入力次元 $2d$，出力次元 d の ReLU 関数つき線形層を適用し，次元削減を行うことで，元の特徴量と同じ次元数である混合特徴量 $\boldsymbol{X}_H$ を作成する．

モデル学習を行う際には，雑音つき学習用音声 $\boldsymbol{S}$ と対応する雑音音声 $\boldsymbol{N}$ を用いてパラメータ調整を行う．また，モデルの学習では，ASR，FE，SE は事前に学習されたものを用い，NFE と Linear 層のみパラメー

タを更新する．これにより，少ないパラメータ更新でさまざまな環境雑音に対応できるモデルを習得できる．また，実際にネットワークを用いる際には，NFE に事前に録音した雑音音声を入力するだけでよく，追加のパラメータ更新を必要としない．これにより，ASR を新たな雑音環境に即時に適応させることが可能となる．

3.2 雑音特徴量生成

本節では，雑音特徴量を生成する NFE の構造について説明する．

これまで，音声データから時不変の特徴量を抽出する手法は，主に話者認識の分野で広く研究されてきた．i–Vectors [14] や x–vector [15] がその代表的なものであるが，いずれも時系列を含むニューラルネットワークを含んでおり，計算量が多いというデメリットがある．それに対し，音声分離技術である SpeakerBeam [16] では，Sequence Summarizing Network と呼ばれる線形層をもとにした特徴抽出器が用いられ，単純なネットワークながら高い性能で雑音の特徴を抽出できることが示されている．そこで本稿では，Sequence Summarizing Network の構造をもとに，雑音の特徴を抽出するネットワークを設計する．この特徴量を音声合成の重みパラメータとして用いることで，環境雑音に特化した特徴量を合成することを目的とする．

図 2 に，NFE の構造を示す．このネットワークは，環境雑音に対して FE を適用して得られたメルスペクトログラム $\boldsymbol{N}$ を入力とする．$\boldsymbol{N}$ の次元数は d，フレーム数は T とする．まず，$\boldsymbol{N}$ に入力次元 d，出力次元 $2d+1$ の多層線形層を適用する．得られた出力結果の 1 から $2d$ 次元までを時系列特徴量 $\boldsymbol{X}$，最後の $2d+1$ 次元目を重み値 $\boldsymbol{\alpha}$ として用いる．次に，$\boldsymbol{\alpha}$ に softmax 関数を適用することで，正規化された重み $\boldsymbol{\alpha}'$ を得る．最後に，それぞれの特徴量次元に対して，重み $\boldsymbol{\alpha}'$ を用いた重みつき時間平均を計算する．これにより，$2d$ 次元の雑音特徴量ベクトル $\boldsymbol{w}$ を作成する．このベクトルを，雑音特徴量とする．

重み $\boldsymbol{\alpha}'$ を用いるのは，時間フレームの中から特に重要なものを強調する，Attention 機構の考えによるものである．入力される雑音音声は強弱があるため，雑音の情報を強くもつタイムフレームの影響を強める必要がある．重みパラメータ $\boldsymbol{\alpha}'$ を用いることで，$\boldsymbol{w}$ に対するそれぞれのフレームごとの影響度を可変にすることができる．また，$\boldsymbol{w}$ の次元を音声特徴量の次元 d の 2 倍にしているのは，強調音声の特徴量と未処理音声の特徴量の両方に $\boldsymbol{w}$ を適用するためである．

3.3 初期化手法

我々の先行研究 [12, 13] では，アダプターを恒等写像に近い形で初期化する Near-Identity 初期化 (NI 初期化) を適用することで，音声認識性能と収束速度を向上させることに成功した．本稿では，この初期化手法を雑音特徴 Biasing モデルにも適用する手法を検討する．

NFE に含まれる多層線形層については，出力層のバイアスを 1 に，それ以外の重みとバイアスを 0 に近い値に初期化する．これにより，雑音特徴量の初期値は等しい値をもつ $2d$ 次元のベクトルとなる．

また，$\boldsymbol{X}_G$ に適用する Linear 層は，$\boldsymbol{X}_G$ と $\boldsymbol{X}$ の平均値に近い値が初期値となるように初期化する．具体的には，線形層の重みの初期値 $\boldsymbol{W}_{\text{init}}$ を対角成分が 0.5，それ以外の成分が 0 に近いランダムな値となる正方行列を 2 つ繋げたような値で初期化し，またとバイアス項の初期値 $\boldsymbol{b}_{\text{init}}$ についても 0 に近いランダムな値で初期化することで，これを実現する．

$$\boldsymbol{W}_{\text{init}} = \begin{bmatrix} 0.5 & e_{1,2} & \cdots & e_{1,d} & 0.5 & e_{1,d+2} & \cdots & e_{1,2d} \\ e_{2,1} & 0.5 & \cdots & e_{2,d} & e_{2,d+1} & 0.5 & \cdots & e_{2,2d} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ e_{d,1} & e_{d,2} & \cdots & 0.5 & e_{d,d+1} & e_{d,d+2} & \cdots & 0.5 \end{bmatrix}$$

$$\boldsymbol{b}_{\text{init}} = \begin{bmatrix} e_1 & e_2 & \dots & e_d \end{bmatrix}^T$$

この初期化手法を，雑音特徴 Biasing モデルにおける NI 初期化と定義する．

4 評価実験

本実験では，雑音特徴 Biasing モデルの有効性を確かめるために，LibriSpeech 音声コーパス [17] を用いた学習を行い，音声認識性能に関する比較実験を行った．モデルの学習では，train-clean-100 subset に，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，10 epoch 分の学習を行った．入力雑音 $\boldsymbol{N}$ には，音声に足し合わせた雑音音声と時間同期は行わずに同一の雑音を入力した．なお，この雑音は，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，test-clean を用いた．雑音耐性を評価するために，同一の雑音音声を SNR が，-15 dB から 0 dB の範囲となるよう足し合わせて評価用音声とした．評価指標には単語誤り率 (word error rate, WER) を用いた．

4.1 使用モデル

入力音声は，16 bit, 16 kHz サンプリングの信号であり，Feature Extractor では，これを窓長 512，シフト

表 1: 学習結果（WER）．「C」は Conv-TasNet [5]，「NFBiasing」は提案モデルである 雑音特徴 Biasing モデルを表す．「Layers in NFE」は NFE に含まれる線形層の数を，「Hidden Dim.」は NFE の中間層の次元数を表す．

No.	SE	Model Type	Layers in NFE	Hidden Dim.	Init.	clean	15dB	10dB	5dB	0dB
1	-	only ASR	-	-	-	**6.7**	11.3	20.3	43.0	75.3
2	C	ASR + SE	-	-	-	7.1	**10.3**	**15.3**	27.7	56.1
3	C	NFBiasing (proposed)	3	200	random	11.2	15.8	20.2	30.1	55.1
4	C	NFBiasing (proposed)	3	200	NI(proposed)	8.9	12.0	**15.3**	**23.4**	**45.2**

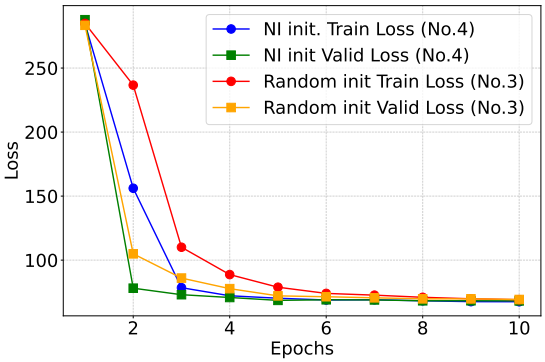

図 3: 学習損失の推移

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．

ASR には，ESPnet の hybrid CTC/attention モデル [18] 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 CTC ブロックで構成されるデコーダーをもつ．エンコーダーは 80 次元の音声特徴量から潜在特徴量を生成してデコーダーに渡し，デコーダーによって潜在特徴量がテキストへと変換される．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 [5] 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 [19] からなる．SE ネットワークは CHiME-4 データセット [20] で事前学習されたモデルを用いた．

NFE に含まれる多層線形層の次元数は，入力が $d = 80$ 次元，出力が $2d + 1 = 161$ 次元とし，中間層の次元を $100, 200, 300$，層数を $3, 5, 7$ と変化させて実験を行った．時間フレームごとの重み $\boldsymbol{\alpha}$ は時間フレームと同じ次元のベクトルで，出力される特徴量ベクトル $\boldsymbol{w}$ も 160 次元である．$\boldsymbol{X}_G$ に適用する Linear 層は，入力を 160 次元，出力を 80 次元とすることで，80 次元の混合特徴量 $\boldsymbol{X}_H$ を ASR ネットワークに渡す．ネットワークの初期化については，一般的なランダム初期化と，3.3 節で述べた NI 初期化を比較する．0 に近い小さな値については，平均 0，分散 10^{-4} の正規分布に従うランダム値とした．

5 実験結果・考察

本章では，実験結果について，初期化手法とモデル設定の観点から考察する．表 1 – 3 に，それぞれの実験設定での WER を示す．

5.1 提案モデルと初期化手法について

表 1 に，事前学習モデル，音声強調適用モデル，初期化手法ごとの雑音特徴 Biasing モデルの学習結果を示す．雑音が多く含まれる音声について，もともとの ASR (No. 1) や音声強調のみ併用した場合 (No. 2) に比べ，提案手法である雑音特徴 Biasing モデルの WER が低く，認識性能が高いことが確認された．これにより，雑音特徴量による特徴量の重み付けが，認識性能向上に寄与することが示された．

さらに，初期化手法についても，ランダム初期化 (No. 3) に比べて NI 初期化を用いた場合 (No. 4) のほうが，認識性能が高い結果となった．ランダム初期化では雑音の少ない音声の認識性能も大きく下がっていることから，10epoch では学習が収束していないことがわかる．図 3 に，それぞれの初期化手法を用いた際の損失推移を示す．この結果からも，NI 初期化を用いたほうが学習収束が早いことが確認できる．先行研究のモデルと同様，モデルの特徴を崩さない状態から学習を開始できるため，限られた学習時間の中でより早く学習が収束できたと考えられる．以上より，雑音特徴 Biasing モデルにおいても，NI 初期化が有効であることが確認できた．

よって，以降の比較実験では，すべてのケースにおいて NI 初期化を用いることとする．

表 2: NFE がもつ線形層の数を変えたときの学習結果（WER）.

No.	SE	Model Type	Layers in NFE	Hidden Dim.	Init.	clean	15dB	10dB	5dB	0dB
5	C	NFBiasing	2	200	NI	9.2	12.3	15.5	24.5	47.1
6	C	NFBiasing	3	200	NI	**8.9**	**12.0**	15.3	23.4	45.2
7	C	NFBiasing	5	200	NI	10.6	12.1	**15.1**	23.6	45.2
8	C	NFBiasing	7	200	NI	11.4	12.7	15.4	**23.2**	**44.4**

表 3: NFE の中間層の次元数を変えたときの学習結果（WER）.

No.	SE	Model Type	Layers in NFE	Hidden Dim.	Init.	clean	15dB	10dB	5dB	0dB
9	C	NFBiasing	3	80	NI	9.1	12.0	15.3	23.6	45.5
10	C	NFBiasing	3	100	NI	9.2	12.1	15.4	23.9	45.9
11	C	NFBiasing	3	200	NI	8.9	12.0	15.3	23.4	45.2
12	C	NFBiasing	3	300	NI	9.7	12.2	15.0	23.2	44.8

5.2 NFE の設定

表 2 に，NFE の線形層の数を変更して学習したときの結果を示す．結果より，雑音が強く混入する音声については，層を多くすることで認識性能をより高めることができた．特に，層を 7 層としたとき，WER が 44.4 % となっており，SE のみ用いた場合に比べて認識性能が 11.7 ポイント向上している．その一方で，雑音が少ない音声 (clean, 15dB) では，層が多いほど性能が悪化している．これは，NFE の表現力が増えることで，モデルが雑音つき音声に対してより適応し，もともとの学習済み ASR の特徴を失ったことが原因と考えられる．この現象は「破滅的忘却 (Catastrophic Forgetting)」[21] と呼ばれる現象で，アダプターモデルや転移学習における課題となっている．雑音特徴 Biasing モデルで破滅的忘却を避けるためには，NFE の線形層数を大きくしすぎず，表現力を抑制することが有効であると確認できた．

また，表 3 に，中間層の次元を増減してモデル学習を行った場合の認識性能を示す．結果より，層数を増やす場合と同様に，隠れ層の次元を増やすことでも，認識性能が向上することが確認できた．隠れ層の次元を増やすことで NFE の表現力が増し，より雑音影響の少ない特徴量を排出できるようになったことが要因と考えられる．一方，中間層次元を大きくすることでも，雑音が少ない音声の認識性能が低下し，破滅的忘却が確認された．中間層の次元についても，過度に増やすのではなく，適切に値を選択することが必要となる．

以上の結果より，NFE の線形層の表現力を増すことで，雑音成分が多い音声の認識性能をより高めることができることを確認できた．その一方で，雑音成分が少ない音声の認識性能とのトレードオフとなることも確認できた．これより，適切な総数と次元数を選択することで，用途に応じた ASR ネットワークを構築することができると考える．

6 まとめ

本稿では，音声認識の雑音耐性を向上する手法として，雑音特徴量を用いて音声特徴量の変換を行う「雑音特徴 Biasing Model」を提案した．このモデルには雑音特徴生成器 (NFE) が含まれ，事前に録音された環境雑音音声から時不変の雑音特徴量を生成する．この雑音特徴量を重みとして音声特徴量を変換することで，雑音影響の少ない特徴量を ASR ネットワークに渡すことができる．さらに，パラメータの初期値を恒等写像に近い形とする「NI 初期化」を提案モデルで実装し，学習の高速化と性能向上を図った．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，もとの SE のみを用いたモデルと比較して，最大で 11.7 ポイントの性能向上を達成した．今後の課題として，NFE の改良，複数種類の学習を用いた実験・評価があげられる．

謝辞

本稿は JSPS 科研費 JP22F22769，JP22KF0141，立石財団研究助成 (A) 2241011 および福島国際研究教育機構 (F-REI) の委託研究費 (JPFR23010102) の助成を受けた．また本稿は，東京科学大学のスーパーコンピュータ TSUBAME4.0 を利用して実施した．

参考文献

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 大﨑崇博, 周藤唯, 糸山克寿, 西田健次, 中臺一博. Parallel adapter model と near-identity 初期化を用いた音声認識の雑音耐性向上. 人工知能学会第二種研究会資料, Vol. 2023, No. Challenge-063, p. 02, 2023.

[13] Takahiro Osaki, Yui Sudo, Katsutoshi Itoyama, Kenji Nishida, and Kazuhiro Nakadai. Improving noise robustness of automatic speech recognition based on a parallel adapter model with near-identity initialization. In *Advances and Trends in Artificial Intelligence. Theory and Applications*, pp. 454–466, Singapore, 2024. Springer Nature Singapore.

[14] George Saon, Hagen Soltau, David Nahamoo, and Michael Picheny. Speaker adaptation of neural network acoustic models using i-vectors. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 55–59, 2013.

[15] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5329–5333, 2018.

[16] Kateřina Žmolíková, Marc Delcroix, Keisuke Kinoshita, Tsubasa Ochiai, Tomohiro Nakatani, Lukáš Burget, and Jan Černocký. Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures. *IEEE Journal of Selected Topics in Signal Processing*, Vol. 13, No. 4, pp. 800–814, 2019.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] Steven Vander Eeckt and Hugo Van Hamme. Using adapters to overcome catastrophic forgetting in end-to-end automatic speech recognition. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.