Teacher-Student 学習を用いた Wave-U-Net による 低遅延リアルタイム音声強調 *

☆中岡 想太郎, 李 莉, 井上 翔太, 牧野 昭二
筑波大学

1 はじめに

本稿では時間領域におけるモノラル音声強調問題を扱う．音声認識や補聴器や車室内コミュニケーションをはじめ音声を用いるアプリケーションが普及している．入力音声に雑音が含まれる場合に音声の明瞭度や品質が著しく低下するため，これらのアプリケーションにおいては雑音を抑圧し，音声を強調する音声強調技術が不可欠である．その中には，音声認識などバッチ処理で実装されるものがある一方で，補聴器や車室内コミュニケーションをはじめとする対話を前提としたアプリケーションでは低遅延なリアルタイム動作が要求される．例えば，補聴器において，処理遅延が 10ms 以上の場合に会話相手の発した音声と処理後の音声が二重に聞こえ，聴覚上では不快感が生じる傾向にあることが知られている [1].

近年，深層学習（Deep Neural Network: DNN）の急速な発展により，モノラル音声強調の性能が飛躍的に向上した [2]．多くの手法では，短時間フーリエ変換（Short-Time Fourier Transform: STFT）で得られた振幅スペクトログラムを用いて DNN を学習し，時間周波数領域において雑音と音声を分離する [3, 4]．これらの手法は従来手法を凌駕する性能を示す一方で，推定信号の位相に混合信号の位相を利用するため，強調音声の明瞭度と品質に限界があるという問題点が残されている [5, 6]．位相を考慮する DNN を学習する手法も多数提案されているが [7, 8]， STFT の冗長性に対する制約が課されておらず，推定された複素スペクトログラムに対応する時間領域の音声信号が存在しない可能性がある [9]．この問題を解決するため，時間領域で音源分離を行う手法が提案されている [10–12]．Wave-U-Net はそのような手法の一つであり，最初は歌声分離の手法として提案され [12]，後に音声強調に応用されている [13].

Wave-U-Net による音声強調は，雑音を含む観測（混合）音声を入力とし，1 次元畳み込みニューラルネットワーク（Convolutional Neural Network: CNN）で構成される Down sampling (DS) と Up sampling (US) のブロックによって雑音成分の抑圧された音声信号を出力する手法である．推定された音声信号と混合信号の差分を推定の雑音信号とすることで，分離された各信号の和が入力信号と一致し，分離前後での整合性の保持が可能である．これらの性質を保ちながら Wave-U-Net は音声強調タスクにおいて高い性能が示されている．更に、長い窓を用いた STFT による高い周波数分解能が必要である周波数領域での高品質な音声強調に比べ，Wave-U-Net を含む時間領域での音声強調手法には窓長の制約が生じず、低遅延での処理が可能である．特に遅延の少ないオンライン処理が求められる会話型アプリケーションに適していると考えられる．従来の Wave-U-Net では，発話全体を入力とするオフライン処理の手法であるが，全ての層が CNN 層のみで結合された全畳み込みニューラルネットワーク（Fully Convolutional Network: FCN）であるため，任意長の混合信号を入力として扱うことができる．従って，ブロック処理の導入によるオンライン処理への拡張が考えられる．しかしながら，この拡張では低遅延化のためにブロック長の短縮が必要があり，推定に利用可能な入力セグメントからの情報量の減少によって分離性能が著しく低下する傾向にある．本研究では，オフライン処理の Wave-U-Net を低遅延かつリアルタイム処理に適用可能なオンライン手法に拡張する．更に，ブロック長の短さを原因とする性能低下を防ぐため，Teacher-Student 学習 (知識蒸留) [14] を利用したモデル学習法を提案する．車室内コミュニケーションを想定し，実測した走行雑音を用いて音声強調実験を行い，オンライン Wave-U-Net と Teacher-Student 学習の有効性を確認する.

2 Wave-U-Net による音声強調

本章では，オフライン Wave-U-Net を用いた音声強調 [13] の問題設定とネットワーク構造について説明する．音声信号，雑音信号と観測信号をそれぞれ $s(t) \in [-1, 1]$, $n(t) \in [-1, 1]$, $m(t) \in [-1, 1]$ で表す．Wave-U-Net は観測された信号を分割せずに入力し，ニューラルネットワークを用いて雑音成分の抑圧された音声信号 $\hat{s}(t),\ t = 1, \ldots, T$ を求める．

$$\hat{s}(1), \ldots, \hat{s}(T) = \mathcal{F}_\theta(m(1), \ldots, m(T)) \qquad (1)$$

ここで、$\mathcal{F}_\theta(\cdot)$ はパラメータ θ を持つニューラルネットワークによる混合信号から推定された音声信号への非線形写像を表し，$t = 1, \ldots, T$ は時間のインデックスである．混合信号と分離信号間の整合性を満たすために，推定雑音信号は以下のように計算される．

$$\hat{n}(t) = m(t) - \hat{s}(t) \qquad (2)$$

ネットワークパラメータ θ は以下のように，推定された音声信号と雑音信号の組と，混合される前の音声信号および雑音信号の組の最小二乗誤差を最小化する基準の下で学習される．

$$\mathcal{L} = \mathbb{E}\Big[|\hat{s}(t) - s(t)|^2 + |\hat{n}(t) - n(t)|^2\Big] \qquad (3)$$

*Wave-U-Net with teacher-student learning for low latency online speech enhancement. Sotaro Nakaoka, Li Li, Shota Inoue, Shoji Makino(University of Tsukuba).

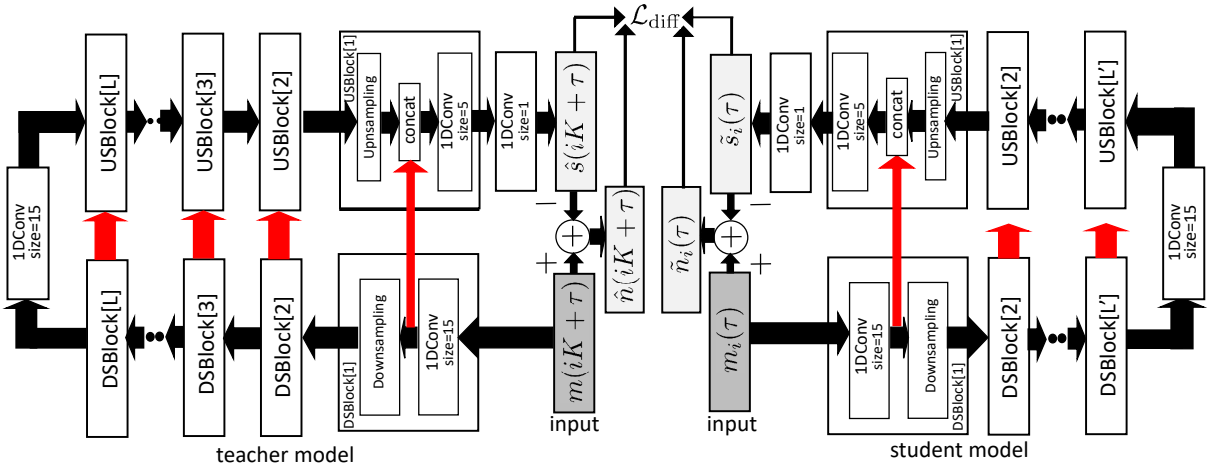

Fig. 1　Teacher-Student 学習に用いられるオフラインおよびオンライン Wave-U-Net モデルのネットワーク構造と学習法.

Wave-U-Net の基本的なネットワーク構造を Fig. 1 中の teacher model に示す．Wave-U-Net は DS ブロックと US ブロックによって構成されており，DS ブロックと US ブロックの間にボトルネック層，US ブロックの直後に出力層である 1 次元 CNN 層が設けられている．DS ブロックは 1 次元 CNN 層とそれに続く DS 層によって構成されており，CNN 層における畳み込み処理によってチャンネル数を増加させた後，DS 層でデータを 1 サンプルおきに排除する事で時間スケール上の分解能を半減させる．US ブロックは US 層とそれに続く 1 次元 CNN 層によって構成されており、US 層で各サンプルの間に中間値を補間してデータ長を倍にし，CNN 層において畳み込みによってチャンネル数を減少させる．なお，US 層ではゼロパディングを伴う転置畳み込みではなく線形補間を用いることで，ゼロパディングによるエイリアシングの回避が可能である [12]．DS ブロックと US ブロックは対称であり，それぞれ L 個存在する．従って，出力層に入力される特徴量の時間分解能は入力データと同一である．また，出力層における非線形活性化関数としては tanh 関数が用いられており，それ以外では Leaky ReLU を用いられている．

3　低遅延オンライン Wave-U-Net モデル

Wave-U-Net は音声強調タスクにおいて優れた性能を発揮することが示されている一方で，発話データの全区間を入力とするオフライン処理での利用が前提となっている．本研究では，Wave-U-Net の高い音声強調性能を維持しつつ，10ms 以上の遅延がユーザにとって好ましくない会話型のアプリケーションを対象としたコンパクトなサイズのオンライン Wave-U-Net モデルを提案する．具体的には，Teacher-Student 学習を適用したモデル学習法を提案する．本節では，まず，ブロック処理を用いたオンラインアルゴリズムについて述べる．次に，Teacher-Student 学習を導入した Wave-U-Net のモデル学習法について述べる．

3.1　オンライン Wave-U-net

オンライン処理を実現する最も単純な方法としてはブロック処理の導入が考えられる．ブロック処理では固定長の入力セグメント $m_i(\tau) = m(iK+\tau)$ に対して逐次的に音声強調を行う．ここで，$\tau = 1, \ldots, K$ と i はそれぞれ時間とセグメントのインデックスであり，K はセグメント長である．セグメント長 K はシステム遅延の下限であるため，可能な限りの K の短縮が求められる．しかしながら，K の短縮によって起こるセグメント内の利用可能な情報量の減少は音声強調性能の低下を招く可能性がある．したがって，音声強調性能とセグメント長の間に生じるトレードオフを考慮する必要がある．また，オンライン処理のアルゴリズムにおいては現時刻以降の情報を利用できないため，データの左側にのみゼロパディングを施す Causal CNN が広く利用されている．通常の CNN に比べ，Causal CNN はセグメント境界に生じる人工的な歪みを抑えられることが知られている．本稿では，Wave-U-Net において Causal CNN と通常の CNN を比較した実験結果を 4 章に報告する．

3.2　Teacher-student 学習による知識伝達

Teacher-Student 学習は，事前に学習された教師モデルが獲得した推論に関する知識を生徒モデルに伝達する学習方法であり，様々な場面に適用されている．例えば，ドメイン適応 [15, 16] や，深いネットワークから浅いネットワーク [17]，異なるアーキテクチャのネットワーク [18] への知識の伝達に用いられる．教師モデル $\mathcal{F}_\theta(\cdot)$ のパラメータ θ は事前学習で最適化され，生徒モデルの学習時には固定される．生徒モデル $\mathcal{G}_\phi(\cdot)$ のパラメータ ϕ は以下の損失関数を最小化する基準の下で学習される．

$$\mathcal{L}_{\text{stu}} = \mathcal{L} + \beta \mathcal{L}_{\text{diff}} \tag{4}$$

ここで，$\mathcal{L}$ は正解データである雑音信号 $n_i(\tau) = n(iK+\tau)$ と音声信号 $s_i(\tau) = s(iK+\tau)$ の組と

Table 1　異なるブロック数の生徒モデルのパラメータ数の比較.

ブロック数	parameter 数
$L'=6$	1,079,302
$L'=7$	1,625,602
$L'=8$	2,329,942

Table 2　K=1024 における Causal CNN と通常の CNN を用いたモデルの平均 SDR, SIR, SAR の比較.

model	SDR[dB]	SIR[dB]	SAR[dB]
regular CNN	7.82	14.31	9.55
causal CNN	7.51	13.15	9.38

生徒モデルにより推定された雑音信号 $\tilde{n}_i(\tau)$ と音声信号 $\tilde{s}_i(\tau)$ の組の間の最小二乗誤差であり，$\mathcal{L}_{\text{diff}}$ は教師モデルと生徒モデルでそれぞれ推定された信号の組の間の最小二乗誤差である．$\beta \geq 0$ は後者の項に対する重み係数を表す．短い入力セグメント $m_i(\tau)$ を入力とした生徒モデル $\mathcal{G}_\phi(\cdot)$ の出力を $\tilde{s}_i(1),\ldots,\tilde{s}_i(K)=\mathcal{G}_\phi(m_i(1),\ldots,m_i(K))$ とし，推定雑音を $\tilde{n}_i(\tau)=m_i(\tau)-\tilde{s}_i(\tau)$ とする．式 (4) の第一項および第二項は以下のように定義できる．

$$\mathcal{L}=\mathbb{E}\Big[|\tilde{s}_i(\tau)-s(iK+\tau)|^2 + |\tilde{n}_i(\tau)-n(iK+\tau)|^2\Big] \quad (5)$$

$$\mathcal{L}_{\text{diff}}=\mathbb{E}\Big[|\tilde{s}_i(\tau)-\hat{s}(iK+\tau)|^2 + |\tilde{n}_i(\tau)-\hat{n}(iK+\tau)|^2\Big] \quad (6)$$

Fig. 1 にオフライン処理の教師モデルとオンライン処理の生徒モデルを用いた Wave-U-Net の Teacher-Student 学習を表す．教師モデルに対する生徒モデルの相違点は，DS ブロックと US ブロックの数 L' が異なる点と，入力セグメント長が限られている点である．

4　評価実験

本研究では，車内コミュニケーションアプリケーションを想定し，走行雑音を用いて車内環境を模して，オフライン Wave-U-Net，単純なブロック処理によるオンライン Wave-U-Net，および提案法である Teacher-Student 学習を適用したオンライン Wave-U-Net の音声強調性能を評価する．音声強調性能の評価指標として，signal-to-distortions ratio (SDR), signal-to-interferences ratio (SIR), および signals-to-artifacts ratio(SAR) [19] を用いる．また，強調された音声信号の品質を評価する指標として，perceptual evaluation of speech quality (PESQ) [20] と short-time objective intelligibility (STOI) [21] を用いる．

4.1　データセット

教師モデルおよび生徒モデルのネットワークパラメータの学習には，ドライソースの音声信号と車室内で収録された雑音信号を正解データとし，それらの混合信号をネットワークの入力データとして用いた．ドライソースの音声信号は CMU Arctic database [22] より，10 話者の音声発話を用いた．学習と検証には “aew”, “ahw”, “aup”, “axb”, “eey”, “fem” の 6 話者（男性 4 人, 女性 2 人）の音声発話を用い，評価には “awb”, “bdl”, “clb”, “slt” の 4 話者（男性 2 人, 女性 2 人）を用いた．各発話の長さは 3-7 秒である．車室内雑音は JEIDA-NOISE データベース [23] より，2 種類の車両で収録された走行中の車室内における雑音信号を用いた．各車両で 1 時間の走行雑音が収録されており，1 台の走行雑音を学習データとし，別の 1 台の走行雑音を評価に用いた．ネットワークパラメータの学習と検証に用いた混合信号は，音声と雑音の振幅の比率を [0.2, 0.9] の間でランダムに選択し，無作為に選択された区間の雑音信号と発話音声を加算して生成した．音声品質の評価に用いた混合信号は signal-to-noise ration（SNR）を {-3,0,3}dB の条件で各 50 発話，合計 150 発話を生成した．入力混合信号の平均 SDR は 0.07 dB であった．各信号の標本化周波数は 16kHz である．

4.2　実験設定

オフライン Wave-U-Net モデルとして, 8 つの DS ブロックおよび US ブロックで構成したモデル ($L=8$) を用いた．1 次元 CNN 層におけるフィルタサイズは DS ブロック，US ブロック，ボトルネック層と出力層の順に 15，5，15，1 とした．l 番目の DS ブロックや US ブロックのチャンネル数は $20l$ であり，ボトルネック層と出力層のチャンネル数は $20(L+1)$ および d1 とした．学習データのサンプル長を 64000 とし，ゼロパディングまたは区間選択によって各発話のサンプル長を統一した．学習には学習率 0.0001 の Adam を用い，バッチサイズは 32 とした．学習済みのオフラインモデルを教師モデルとして，入力セグメント長 $K=\{64,128,256,512,1024,2048\}$ の生徒モデルの学習を実施した．$K=64$ では $L'=6$，$K=128$ では $L'=7$ であることを除いて，オフラインモデルのネットワーク構造を $L'=8$ に同一とした．異なる L' のネットワークパラメータの総数を Table 1 に示す．Teacher-Student 学習における重み β は 0.01 に設定した．

4.3　実験結果

まず，入力セグメント長 $K=1024$ のブロック処理を導入した Wave-U-Net に通常の CNN と Causal CNN を用いた場合の音声強調性能を比較する予備実験を実施した．Table 2 に実験結果として，混合信号 150 発話に対する SDR の平均値を示す．結果より，Causal CNN を用いた場合の音声強調性能は通常 CNN を用いた場合と比較してわずかに低下することが確認された．この実験結果を考慮して，以降全ての

Table 3　Teacher-Student 学習の有無による平均 SDR, SIR, SAR, PESQ, STOI の比較.

model	K	SDR [dB]		SIR [dB]		SAR [dB]		PESQ		STOI	
		$\beta=0$	$\beta=0.01$	$\beta=0$	$\beta=0.01$	$\beta=0$	$\beta=0.01$	$\beta=0$	$\beta=0.01$	$\beta=0$	$\beta=0.01$
online	64	4.79	8.80	11.62	20.36	6.34	9.34	2.14	2.34	0.85	0.87
	128	6.07	10.35	13.85	24.12	7.22	10.66	2.40	2.58	0.88	0.89
	256	7.83	11.29	14.68	25.63	9.17	11.55	2.54	2.73	0.91	0.91
	512	7.08	11.80	14.24	27.43	8.55	11.99	2.70	2.80	0.92	0.92
	1024	7.82	12.19	14.31	28.07	9.55	12.38	2.72	2.90	0.93	0.93
	2048	8.28	12.22	16.38	28.54	9.68	12.38	2.78	2.94	0.92	0.93
offline		13.84	—	29.81	—	13.98	—	2.56	—	0.92	—

Table 4　異なるセグメント長 K における実行時間およびシステム遅延.

K	processing time[ms]	system latency[ms]
64	2.71	6.71
128	3.40	11.40
256	4.83	20.83
512	6.05	38.05
1024	8.62	72.62
2048	14.89	142.89

実験では通常の CNN を用いた．Table 3 に，それぞれの手法の平均 SDR，SIR，SAR，PESQ および STOI を示す．Teacher-Student 学習を用いないオンライン処理のモデルでは，セグメント長の短縮によって音声強調性能が低下し，特に $K=64$ の場合に SDR が 9 dB 以上低下する結果が示された．一方で，Teacher-Student 学習を適用することにより，いずれのセグメント長のモデルにおいても約 4dB の SDR の向上が可能であることを確認した．Table 4 に，各モデルのシステム遅延と，1 セグメントあたりの平均処理時間を示す．全ての処理を Python および PyTorch で実装し，Intel (R) Core (TM) i7-7800XCPU@3.50GHz で推論を実行した．本実験における最も短いシステム遅延は $K=64$ の場合の約 6.7 ms であり，対話型アプリケーションへの実用を目的とした，10 ms 未満のシステム遅延の要求を満たす．また，約 11.4 ms のシステム遅延で動作可能な $K=128$ のモデルは，$K=64$ のモデルと比較して SDR が約 1.5 dB 向上し，会話型アプリケーションへの実用が可能であると考えられる．

5　おわりに

本研究では，低遅延性が要求される音声を用いた会話型のアプリケーションを想定し，オフライン処理の Wave-U-Net を低遅延かつリアルタイム処理に適用可能なオンライン手法に拡張した．ブロック長の短さを原因とする性能低下を防ぐため，オフライン処理の Wave-U-Net を教師モデルとした Teacher-Student 学習を適用したモデル学習法を提案した．走行中の車室内を模した評価実験により，提案法ではシステムの遅延を約 6.7 ms に留めながら，8.8 dB の音声強調性能を実現した．

謝辞　本研究は JSPS 科研費 19H04131 及び戦略的基盤技術高度化支援事業の助成を受けて行われた。

参考文献

[1] J. Agnew, *et al.*, *JAAA*, 11(6), pp. 330–336, 2000.
[2] D. Wang, *et al.*, *IEEE/ACM Trans. ASLP*, 26(10), pp. 1702–1726, 2018.
[3] Y. Xu, *et al.*, *IEEE/ACM Trans. ASLP*, 23(1), pp. 7–19, 2014.
[4] K. Tan, *et al.*, *Interspeech*, pp. 3229–3233, 2018.
[5] M. S. E. Langarani, *et al.*, *ISSPA*, pp. 1446–1447, 2012.
[6] K. Paliwal, *et al.*, *Speech Communication*, 53(1), pp. 465-494, 2011.
[7] H. Erdogan, *et al.*, *ICASSP*, pp. 708–712, 2015.
[8] D. S. Williamson, *et al.*, *IEEE/ACM Trans. ASLP*, 24(3), pp. 483-492, 2016.
[9] S. Wisdom, *et al.*, *ICASSP*, pp. 900–904, 2019.
[10] S. Pascual, *et al.*, *Interspeech*, pp. 3642–3646, 2017.
[11] Y. Luo, *et al.*, *ICASSP*, pp. 696–700, 2018.
[12] D. Stoller, *et al.*, *ISMIR*, 2018.
[13] C. Macartney, *et al.*, *arXiv preprint:1811.11307*, 2018.
[14] G. Hinton, *et al.*, *NIPS*, 2014.
[15] Z. Meng, *et al.*, *ICASSP*, pp. 6445–6449, 2019.
[16] V. Manohar, *et al.*, *SLT*, pp. 250–257, 2018.
[17] A. Polino, *et al.*, *ICLR*, 2018.
[18] K. J. Geras, *et al.*, *ICLR*, 2016.
[19] E. Vincent, *et al.*, *IEEE Trans. ASLP*, 14(4), pp. 1462–1469, 2006.
[20] A. W. Rix, *et al.*, *ICASSP*, pp. 749–752, 2001.
[21] C. H. Taal, *et al.*, *ICASSP*, pp. 4214–4217, 2010.
[22] J. Kominek *et al.*, *SSW*, pp. 223–224, 2004
[23] JEIDA Noise Database, accessed in Oct. 21, 2020, http://research.nii.ac.jp/src/en/JEIDA-NOISE.html