軽量な RNN を用いた音声強調 *

◎竹内大起, 矢田部浩平 (早大理工), 小泉悠馬, 原田登 (NTT), 及川靖広 (早大理工)

1 まえがき

音声強調は様々なタスクの前処理として応用される重要な音響信号処理である．深層ニューラルネットワーク (DNN) を用いた時間周波数マスキングによる音声強調が数多く研究されており，用いられる DNN 構造の一つに再帰的ニューラルネットワーク (RNN) がある．RNN は時系列データのモデリングが可能な一方で，誤差逆伝播時に勾配の消失や発散が起こり学習が困難になることがある．Long short-term memory (LSTM) を用いることで勾配の発散を緩和できるが，LSTM が持つ 3 つのゲート構造は，それぞれ 2 つの全結合層を必要とするため，パラメータ数が増大する (図–1(a))．本稿では，Equilibriated RNN (ERNN) を時間周波数マスク推定に適用し，Bidirectional LSTM (BLSTM) の 1/9 以下のパラメータ数で同等以上の性能を持つことを確認した [1].

2 DNN 音声強調

観測信号 x が目的信号 s と雑音 n が足し合わされたものと考えれば，観測信号 x は

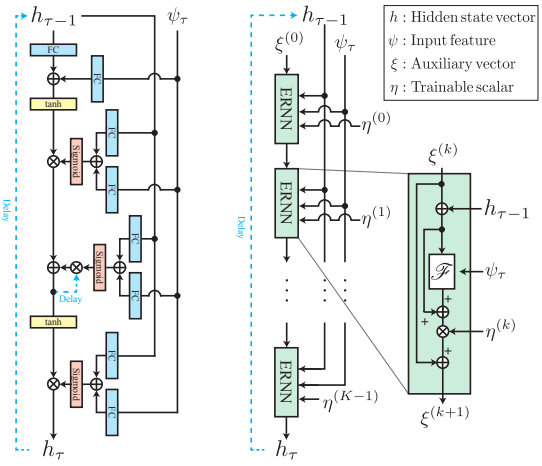
$$x_t = s_t + n_t$$

と書ける．ただし，t は時間のインデックスである．音声強調は観測信号 x から目的信号 s を取り出す音響信号処理である．時間周波数マスキングを用いた音声強調では，マスク $G_{\omega,\tau}$ を用いて，推定信号 $\hat{s}$ を

$$\mathcal{F}(\hat{s})_{\omega,\tau} = G_{\omega,\tau}\mathcal{F}(x)_{\omega,\tau}$$

と計算する．ただし，$\mathcal{F}(\cdot)$ は時間周波数変換を表し，短時間 Fourier 変換 (STFT) が広く用いられる．ω と τ は時間周波数領域における周波数と時間のインデックスである．DNN 音源強調では，DNN を用いた関数 $\mathcal{M}$ でマスク $G_{\omega,\tau}$ を $G_{\omega,\tau} = \mathcal{M}_\theta(\Psi)_{\omega,\tau}$ と推定する．ただし，θ はニューラルネットワークのパラメータ，Ψ は観測信号から得られる音響特徴量である．

RNN は様々なタスクに広く用いられる DNN 構造の一つであり，再帰的構造によって時系列データのモデリングが可能である．一方で，誤差逆伝播時に同じ層の勾配を複数回乗算するので，学習に勾配の消失や発散を伴うことが知られている．勾配の発散を緩和する RNN 構造の一つとして LSTM があり，LSTM を時間の順方向と逆方向の双方向に適用した BLSTM が DNN 音声強調に広く用いられている．LSTM は



(a) LSTM (b) ERNN

図–1 LSTM と ERNN の構造

勾配の発散を緩和ために 3 つのゲート構造を用いており，ゲート構造は全結合層を 2 つ必要とするためパラメータ数が増大する．

勾配の消失や発散を緩和する別の手法として，常微分方程式の陰解法を参考にした RNN 構造である ERNN が提案されている [2]．ERNN は一回の時間発展を

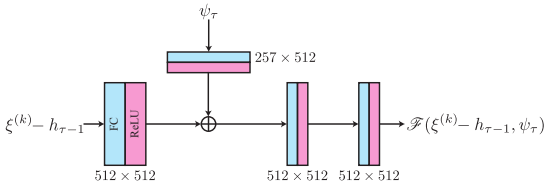
$$\xi^{(k+1)} = \xi^{(k)} + \eta^{(k)}[\mathscr{F}(\psi_\tau, \xi^{(k)} + h_{\tau-1}) - (\xi^{(k)} + h_{\tau-1})]$$

と計算する．ただし，h_τ は隠れ状態，$\xi^{(k)}$ は ERNN 内の反復における一時的な変数で $\xi^{(0)} = \mathbf{0}$，$\eta^{(k)}$ は学習可能なスカラー，$\mathscr{F}$ は ψ_τ と $\xi^{(k)}$ を入力とする DNN，$k = 0, \ldots, K-1$ は反復のインデックスである．反復回数 K は任意で，K 回の反復後に時間インデックス τ が 1 つ進む．ERNN ではこの反復によって ERNN の勾配のノルムが 1 に近づき，勾配の消失や発散を緩和する．本稿では，時間周波数マスキングを用いた音声強調に ERNN を適用し，従来手法である LSTM と BLSTM との比較を行った．

3 実験

時間周波数マスクの推定に ERNN を適用し，音声強調の性能とそのパラメータ数を従来手法と比較した．ERNN 内の DNN $\mathscr{F}$ には図–2 を適用し，反復回数 K は $1, 3, 5$ の 3 種類を用いた．従来手法として，LSTM と BLSTM を用いた時間周波数マスク推定を行った．また，ERNN の反復構造の有効性を確認するため，一般的な RNN，FastRNN [3]，FastGRNN [4] との比較も行った．それぞれの隠れ層の次元はすべて

* Speech enhancement with small RNN. By Daiki TAKEUCHI, Kohei YATABE (Waseda University), Yuma KOIZUMI, Noboru HARADA (NTT) and Yasuhiro OIKAWA (Waseda University).



図–2　ERNN 内の DNN $\mathscr{F}$ の構造

表–1　実験に用いた DNN 構造

Layer	Type	Size (activation)
LSTM/BLSTM		
Layer1	LSTM/BLSTM	257→512
Layer2	LSTM/BLSTM	512→512
output	Fully	512→257 (sigmoid)
RNN/FastRNN/FastGRNN		
Layer1	RNN/FastRNN/FastGRNN	257→512
Layer2	RNN/FastRNN/FastGRNN	512→512
output	Fully	512 → 257 (sigmoid)
ERNN		
Layer1	ERNN	257→512
output	Fully	512→257 (sigmoid)

512 とし，最終層の活性化関数をシグモイド関数としてマスクの値を 0 以上 1 以下に制限した．DNN 全体の構造を表–1 に示す．時間周波数変換には Hann 窓 512 点，DFT 点数 512 点，時間シフト幅 256 点の STFT を適用し，逆 STFT にはその標準双対窓を用いた [5]．DNN への入力特徴量は観測信号のスペクトログラムの log 振幅を用いた．学習データと評価のためのテストデータは Voice Bank と Diverse Environments Multichannel Acoustic Noise Database (DEMAND) が目的音と雑音として混合されたデータセット [6] を利用した．データのサンプリング周波数はどちらも 16 kHz とした．ロス関数には時間領域での平均絶対誤差 (MAE) と圧縮された信号対歪み比 (SDR) [7] の 2 つを適用した．最適化は Adam を用いて行い，ミニバッチサイズは 16 とした．各バッチは音声を 32768 点 (約 2 秒) ずつ切り出した．評価指標には PESQ [8] と主観評価を予測するために評価指標を複合させた CSIG，CBAK，COVL [9] を用いた．

3.1　実験結果

ロス関数に MAE を用いた結果を表–2，SDR を用いた結果を表–3 に示す．ロス関数が MAE，SDR のどちらの場合でも，ERNN，FastRNN，FastRNN は BLSTM，LSTM と同等かそれ以上の性能を BLSTM の 1/9，LSTM の 1/3 以下のパラメータ数で実現した．また，ロス関数を SDR とした時の $K = 3, 5$ の ERNN の性能は $K = 1$ の ERNN，FastRNN，FastGRNN のものより全ての評価指標で向上しており，DNN 音声強調においても ERNN の反復による勾配の消失や発散の緩和がより効果的な学習を可能としていると考えられる．一方で，$K = 3, 5$ のときの性能に大きな差はなく， 勾配の消失や発散の緩和は数回程度の ERNN 内の反復 K で十分に効果があると考えられる．

表–2　ロス関数を MAE としたときの結果

DNN	K	#param.	PESQ	CSIG	CBAK	COVL
LSTM	-	3.81M	2.46	3.64	2.61	3.04
BLSTM	-	9.72M	2.49	3.62	**2.63**	3.04
ERNN	1	1.05M	2.49	3.71	**2.63**	3.09
ERNN	3	1.05M	**2.51**	3.71	**2.63**	**3.10**
ERNN	5	1.05M	2.48	3.68	**2.63**	3.07
RNN	-	1.05M	2.32	3.46	2.54	2.87
FastRNN	-	1.05M	2.30	3.49	2.53	2.88
FastGRNN	-	1.05M	**2.51**	**3.72**	**2.63**	**3.10**

表–3　ロス関数を SDR としたときの結果

DNN	K	#param.	PESQ	CSIG	CBAK	COVL
LSTM	-	3.81M	2.48	3.65	2.62	3.05
BLSTM	-	9.72M	2.52	3.63	2.64	3.06
ERNN	1	1.05M	2.52	3.69	2.63	3.09
ERNN	3	1.05M	**2.54**	**3.77**	**2.65**	**3.14**
ERNN	5	1.05M	**2.54**	**3.77**	**2.65**	**3.14**
RNN	-	1.05M	2.46	3.26	2.59	2.84
FastRNN	-	1.05M	2.47	3.62	2.61	3.03
FastGRNN	-	1.05M	2.51	3.71	2.63	3.10

4　むすび

本稿では，RNN の勾配消失問題を緩和する手法の一つである ERNN を音声強調のマスク推定に適用した．ERNN と従来手法である LSTM の比較を行い，より少ないパラメータで同等以上の音声強調性能が実現できることを確認した．今後はより音声強調に適した ERNN 内の DNN 構造の検討を行う．

参考文献

[1] D. Takeuchi, K. Yatabe, Y. Koizumi, Y. Oikawa, and N. Harada “Real-time speech enhancement using equilibriated RNN.” (submitted)

[2] A. Kag, Z. Zhang, and V. Saligrama, “RNNs evolving in equilibrium: A solution to the vanishing and exploding gradients,” *arXiv preprint arXiv:1908.08574*, 2019.

[3] H. Jaeger, M. Lukosevicius, D. Popovici, and U. Siewert, “Optimization and applications of echo state networks with leaky-integrator neurons.,” Neural Networks, vol. 20, no. 3, 2007, 335-352.

[4] A. Kusupati, M. Singh, K. Bhatia, A. Kumar, P. Jainand, and M. Varma, “FastGRNN: A Fast, Accurate, Stable and Tiny Kilobyte Sized Gated Recurrent Neural Network.,” in *Adv. Neural Inf. Process. Syst. 31*, 2018, pp. 9017–9028.

[5] K. Yatabe, Y. Masuyama, T. Kusano and Y. Oikawa, “Representation of complex spectrogram via phase conversion,” *Acoust. Sci. Tech*, vol. 40, no. 3, 2019.

[6] C. Valentini-Botinho, X. Wang, S. Takaki, and J. Yamagishi, “Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech.,” in *9th ISCA Speech Synth. Workshop*, 2016, pp. 146–152.

[7] H. Erdogan and T. Yoshioka, “Investigations on data augmentation and loss functions for deep learning based speech-background separation,” in *Interspeech 2018*, 2018, pp. 3499–3503.

[8] *P.862.2: Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs*, ITU-T Std. P.862.2, 2007.

[9] Y. Hu and P. C. Loizou, “Evaluation of objective quality measures for speech enhancement,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, 2008.