Proximal Deep Recurrent Neural Network for Monaural Singing Voice Separation

Weitao Yuan, Shengbei Wang, Xiangrui Li — Tianjin Polytechnic University
Masashi Unoki — JAIST
Wenwu Wang — University of Surrey

Contact Information:
School of Computer Science and Technolody
Tianjin Polytechnic University,
Tianjin, China

Phone: +86 135 1282 1367
Email: weitaoyuan@hotmail.com

Abstract

The recent deep learning methods can offer state-of-the-art performance for Monaural Singing Voice Separation (MSVS). In these deep methods, the recurrent neural network (RNN) is widely employed. This work proposes a novel type of Deep RNN (DRNN), namely Proximal DRNN (P-DRNN) for MSVS, which improves the conventional Stacked RNN (S-RNN) by introducing a novel interlayer structure. The interlayer structure is derived from an optimization problem for Monaural Source Separation (MSS). Accordingly, this enables a new hierarchical processing in the proposed P-DRNN with the explicit state transfers between different layers and the skip connections from the inputs, which are efficient for source separation. Finally, the proposed approach is evaluated on the MIR-1K dataset to verify its effectiveness. The numerical results show that the P-DRNN performs better than the conventional S-RNN and several recent MSVS methods.

Introduction

Monaural Singing Voice Separation (MSVS), as an important examplar of Monaural Source Separation (MSS), aims to separate the singing voice (vocal) from the background music components in a single channel mixture signal. Compared to traditional shallow methods, deep learning methods such as Deep Neural Network (DNN) have recently emerged as powerful alternatives and provided state-of-the-art performance for MSVS with the help of large datasets. There are three basic structures to construct DNN for MSVS: (i) Feed-Forward Network (FFN); (ii) Convolutional Neural Network (CNN); (iii) Recurrent Neural Network (RNN). The advantage of employing deep methods is built on a hypothesis that "a deep, hierarchical model can be exponentially more efficient at representing some functions than a shallow one'. This research concentrates on constructing a more effective deep architecture of RNNs for MSVS.

RNN can learn the temporal dynamics in audio signals, thanks to the recurrent (feedback) connections between the hidden units. However, the recurrent connections in RNN offer deep structures only in time, and lack hierarchical processing of the input at different scales. To address this problem, Deep Recurrent Neural Network (DRNN) is proposed, such as the Stacked RNN (S-RNN), which stacks multiple recurrent hidden layers on top of each other. However, the connection ('stacking') between layers of S-RNN is shallow, without intermediate, nonlinear hidden layers (interlayers) between different layers.

Main Objectives

We introduce an improved S-RNN, namely Proximal-DRNN (P-DRNN) for MSVS, which has a novel interlayer (Proximal Layer) to:

1. convey information between different layers via interlayers;
2. explicit state transfers between different layers;
3. have 'skip' connections from the inputs to each layer;
4. be customized for MSS and deepen RNNs effectively for MSVS.

Materials and Methods

The proposed interlayer architecture is derived from a proximal algorithm designed to solve a general MSS optimization problem:

$$\begin{aligned} \underset{\mathbf{x}_{t,j}}{\text{minimize}} \quad & \phi_1(\mathbf{x}_{t,1}) + \phi_2(\mathbf{x}_{t,2}) + ... + \phi_J(\mathbf{x}_{t,J}) \\ \text{subject to} \quad & \sum_{j=1}^{J} \mathbf{x}_{t,j} = \mathbf{m}_t, \end{aligned} \tag{1}$$

where the variable $\mathbf{x}_{t,j}$ corresponds to the j-th estimated source from the mixture $\mathbf{m}_t$, The goal of Eq. (1) is to decompose each frequency feature vector $\mathbf{m}_t$ into $\mathbf{x}_{t,j}$. The proposed method is derived from this optimization problem.

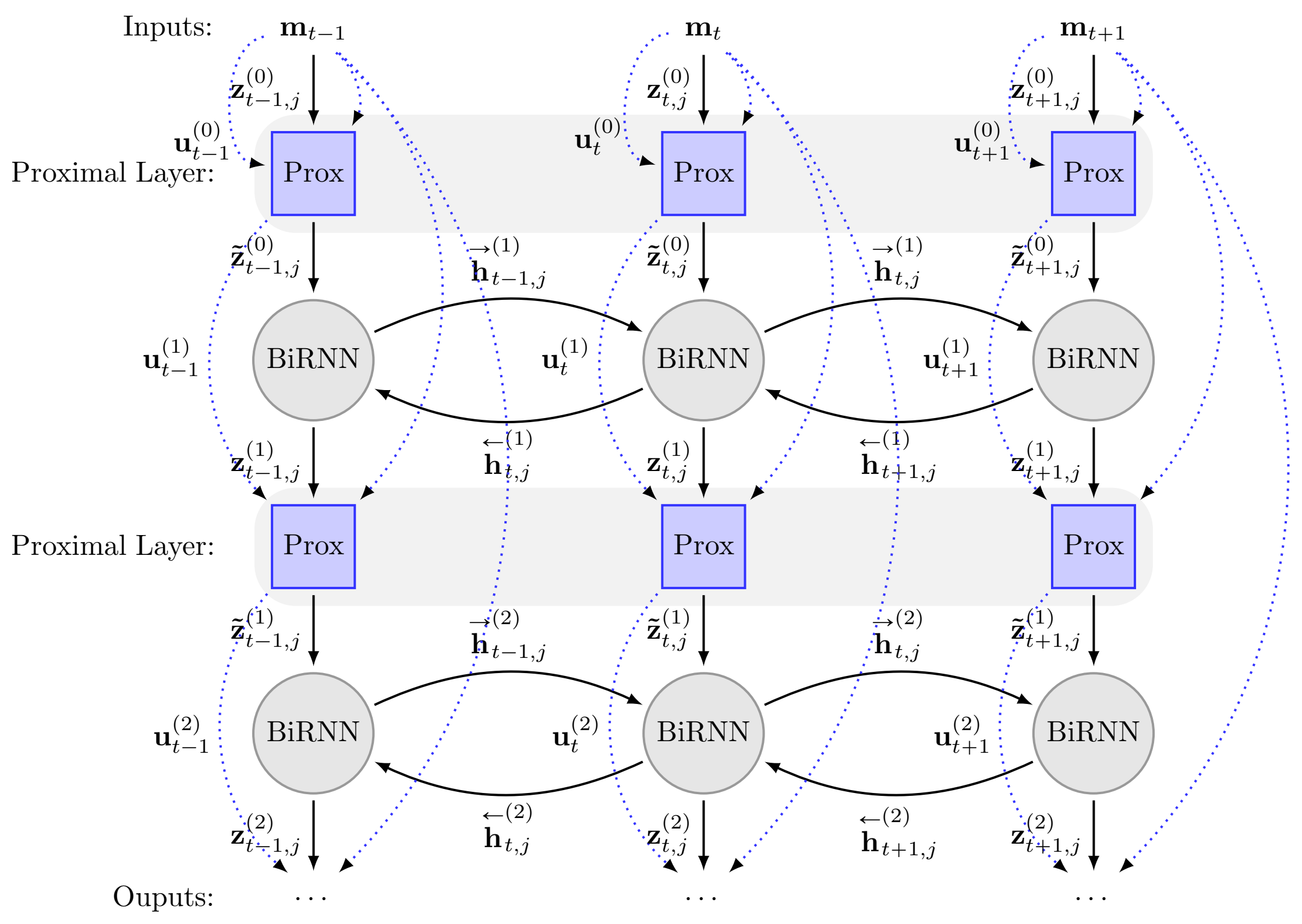
Inputs:
$\mathbf{m}_{t-1}$
$\mathbf{m}_t$
$\mathbf{m}_{t+1}$
$\mathbf{z}^{(0)}_{t-1,j}$
$\mathbf{z}^{(0)}_{t,j}$
$\mathbf{z}^{(0)}_{t+1,j}$
$\mathbf{u}^{(0)}_{t-1}$
$\mathbf{u}^{(0)}_{t}$
$\mathbf{u}^{(0)}_{t+1}$
Proximal Layer:
Prox
$\tilde{\mathbf{z}}^{(0)}_{t-1,j}$
$\overrightarrow{\mathbf{h}}^{(1)}_{t-1,j}$
$\tilde{\mathbf{z}}^{(0)}_{t,j}$
$\overrightarrow{\mathbf{h}}^{(1)}_{t,j}$
$\tilde{\mathbf{z}}^{(0)}_{t+1,j}$
$\mathbf{u}^{(1)}_{t-1}$
BiRNN
$\mathbf{u}^{(1)}_{t}$
$\mathbf{u}^{(1)}_{t+1}$
$\mathbf{z}^{(1)}_{t-1,j}$
$\overleftarrow{\mathbf{h}}^{(1)}_{t,j}$
$\mathbf{z}^{(1)}_{t,j}$
$\overleftarrow{\mathbf{h}}^{(1)}_{t+1,j}$
$\mathbf{z}^{(1)}_{t+1,j}$
Proximal Layer:
$\tilde{\mathbf{z}}^{(1)}_{t-1,j}$
$\overrightarrow{\mathbf{h}}^{(2)}_{t-1,j}$
$\tilde{\mathbf{z}}^{(1)}_{t,j}$
$\overrightarrow{\mathbf{h}}^{(2)}_{t,j}$
$\tilde{\mathbf{z}}^{(1)}_{t+1,j}$
$\mathbf{u}^{(2)}_{t-1}$
$\mathbf{u}^{(2)}_{t}$
$\mathbf{u}^{(2)}_{t+1}$
$\mathbf{z}^{(2)}_{t-1,j}$
$\overleftarrow{\mathbf{h}}^{(2)}_{t,j}$
$\mathbf{z}^{(2)}_{t,j}$
$\overleftarrow{\mathbf{h}}^{(2)}_{t+1,j}$
$\mathbf{z}^{(2)}_{t+1,j}$
Ouputs:
...

Figure 1: The proposed P-DRNN made of alternating layers of Bidirectional RNN (BiRNN) Layer and Proximal (Prox) Layer.

Mathematical Section

In the following, we omit the index t in all the variables for simplicity. First, we rewrite Eq. (1) as an unconstrained minimization problem. We denote

$$\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_J] \in \mathbb{R}^{N\times J}, \tag{2}$$

$$f(\mathbf{X}) = \sum_{j=1}^{J} \phi_j(\mathbf{x}_j), \tag{3}$$

$$g(\mathbf{X}) = \mathbb{I}_C(\mathbf{X}), \tag{4}$$

where $\mathbb{I}_C$ is the indicator function of set C,

$$C = \left\{ \mathbf{X} \in \mathbb{R}^{M\times J} \middle| \sum_{j=1}^{J} \mathbf{x}_j = \mathbf{m}_t \right\}. \tag{5}$$

Thus Eq. (1) becomes

$$\underset{\mathbf{X}}{\text{minimize}} \; f(\mathbf{X}) + g(\mathbf{X}). \tag{6}$$

If f and g are closed convex functions with nonempty domains, and the solution of this minimization problem is not empty, the problem in Eq. (6) can be solved by the primal-dual proximal method. Given an auxiliary variable,

$$\mathbf{U} = [U_1, ..., U_J] \in \mathbb{R}^{N\times J}, \tag{7}$$

the primal-dual method gives the following iteration,

$$\mathbf{X}^{k-1/2} \leftarrow \text{Prox}_{\tau f}(\mathbf{X}^{k-1} - \tau \mathbf{U}^{k-1}), \tag{8}$$

$$\mathbf{U}^{k-1/2} \leftarrow \text{Prox}_{\sigma g^*}(\mathbf{U}^{k-1} + \sigma(2\mathbf{X}^{k-1/2} - \mathbf{X}^{k-1})), \tag{9}$$

$$\mathbf{X}^{k} \leftarrow \mathbf{X}^{k-1} + \rho_k(\mathbf{X}^{k-1/2} - \mathbf{X}^{k-1}), \tag{10}$$

$$\mathbf{U}^{k} \leftarrow \mathbf{U}^{k-1} + \rho_k(\mathbf{U}^{k-1/2} - \mathbf{U}^{k-1}), \tag{11}$$

where k represents the k-th iteration step, g^* is the conjugate of g, and $\text{Prox}_{\tau f}$ and $\text{Prox}_{\sigma g^*}$ are the proximal operators of f and g^*. The parameters ρ_k, τ, and σ are positive. Since Eq. (3) suggests that f is separable, according to Proposition 24.11 in [1], $\text{Prox}_{\tau f}$ in Eq. (8) can be broken into N smaller operations that can be carried out independently in parallel,

$$\text{Prox}_{\tau f}(\mathbf{Y}) = \left(\text{Prox}_{\tau \phi_j}(\mathbf{y}_j)\right)_{1\leq j\leq J}, \forall \mathbf{Y} = [\mathbf{y}_j] \in \mathbb{R}^{N\times J}. \tag{12}$$

The $\text{Prox}_{\sigma g^*}$ can be evaluated analytically. In fact, the proximal operator of an indicator function is a projection operator [1],

$$\text{Prox}_{\sigma g}(\mathbf{Y}) = \text{Proj}_C(\mathbf{Y}) \tag{13}$$

$$= \left(\mathbf{y}_j - \bar{\mathbf{Y}} + (1/J)\mathbf{m}_t\right)_{1\leq j\leq J}, \tag{14}$$

where $\bar{\mathbf{Y}} = 1/J \sum_{j=1}^{J} \mathbf{y}_j$. Suppose $\mathbf{S}$ is a temporary variable,

$$\mathbf{S} = \mathbf{U}^{k-1} + \sigma(2\mathbf{X}^{k-1/2} - \mathbf{X}^{k-1}), \tag{15}$$

according to the following Moreau identity [1]

$$t\text{Prox}_{t^{-1}g^*}(\mathbf{Y}/t) = \mathbf{Y} - \text{Prox}_{tg}(\mathbf{Y}), t > 0, \tag{16}$$

Eq. (9) can be simplified as follows,

$$\begin{aligned} \mathbf{U}^{k-1/2} \leftarrow & \quad \text{Prox}_{\sigma g^*}(\mathbf{S}) && \text{(using Eq. (15))} \\ = & \; \mathbf{S} - \sigma \text{Prox}_{\sigma^{-1}g}(\sigma^{-1}\mathbf{S}) && \text{(using Eq. (16))} \\ = & \; \mathbf{S} - \sigma \text{Proj}_C(\sigma^{-1}\mathbf{S}) && \text{(using Eq. (13))} \\ = & \; \left(\bar{\mathbf{S}} - (1/J)\sigma\mathbf{m}_t\right)_{1\leq j\leq J} && \text{(using Eq. (14))} \end{aligned}$$

which implies that all elements of $\mathbf{U}^{k-1/2}$ are equal. From the definition of $\mathbf{S}$ in Eq. (15), we have, for every $1 \leq j \leq J$,

$$U_j^{k-1/2} \leftarrow \bar{\mathbf{U}}^{k-1} + \sigma(2\bar{\mathbf{X}}^{k-1/2} - \bar{\mathbf{X}}^{k-1}) - (1/J)\sigma\mathbf{m}_t. \tag{17}$$

Furthermore, considering both Eqs. (11) and (17), we can conclude that at any iteration step k (or $k-1/2$), all the elements of $\mathbf{U}^k$ (or $\mathbf{U}^{k-1/2}$) are equal,

$$U_j^k = \mathbf{u}^k, \quad U_j^{k-1/2} = \mathbf{u}^{k-1/2}, \qquad (1 \leq j \leq J). \tag{18}$$

where the elements of $\mathbf{U}^k$ (or $\mathbf{U}^{k-1/2}$) are assumed to be $\mathbf{u}^k$ (or $\mathbf{u}^{k-1/2}$). Based on Eq. (18), Eq. (17) can be simplified as

$$\mathbf{u}^{k-1/2} \leftarrow \mathbf{u}^{k-1} + \sigma(2\bar{\mathbf{X}}^{k-1/2} - \bar{\mathbf{X}}^{k-1}) - (1/J)\sigma\mathbf{m}_t. \tag{19}$$

Based on Eqs. (12) and (19), the iteration of Eqs. (8)-(11) becomes (the index t is omitted for simplicity.)

$$\begin{aligned} \mathbf{x}_j^{k-1/2} &\leftarrow \text{Prox}_{\tau\phi_j}(\mathbf{x}_j^{k-1} - \tau\mathbf{u}^{k-1}), && (1 \leq j \leq J) \\ \mathbf{x}_j^{k} &\leftarrow \mathbf{x}_j^{k-1} + \rho_k(\mathbf{x}_j^{k-1/2} - \mathbf{x}_j^{k-1}), && (1 \leq j \leq J) \\ \mathbf{u}^{k} &\leftarrow \mathbf{u}^{k-1} + \rho_k\left(\sigma(2\bar{\mathbf{X}}^{k-1/2} - \bar{\mathbf{X}}^{k-1}) - (1/N)\sigma\mathbf{m}_t\right). \end{aligned}$$

Results

Figure 2 presents the vocal separating performance of both P-DRNN and S-RNN for various depths L with respect to $T = 10$. When the number of layers is increased to more than 3, the S-RNN experienced a rapid performance decrease. For P-DRNN, we can see that its performances of GNSDR and GSAR improve stably with deeper layers.

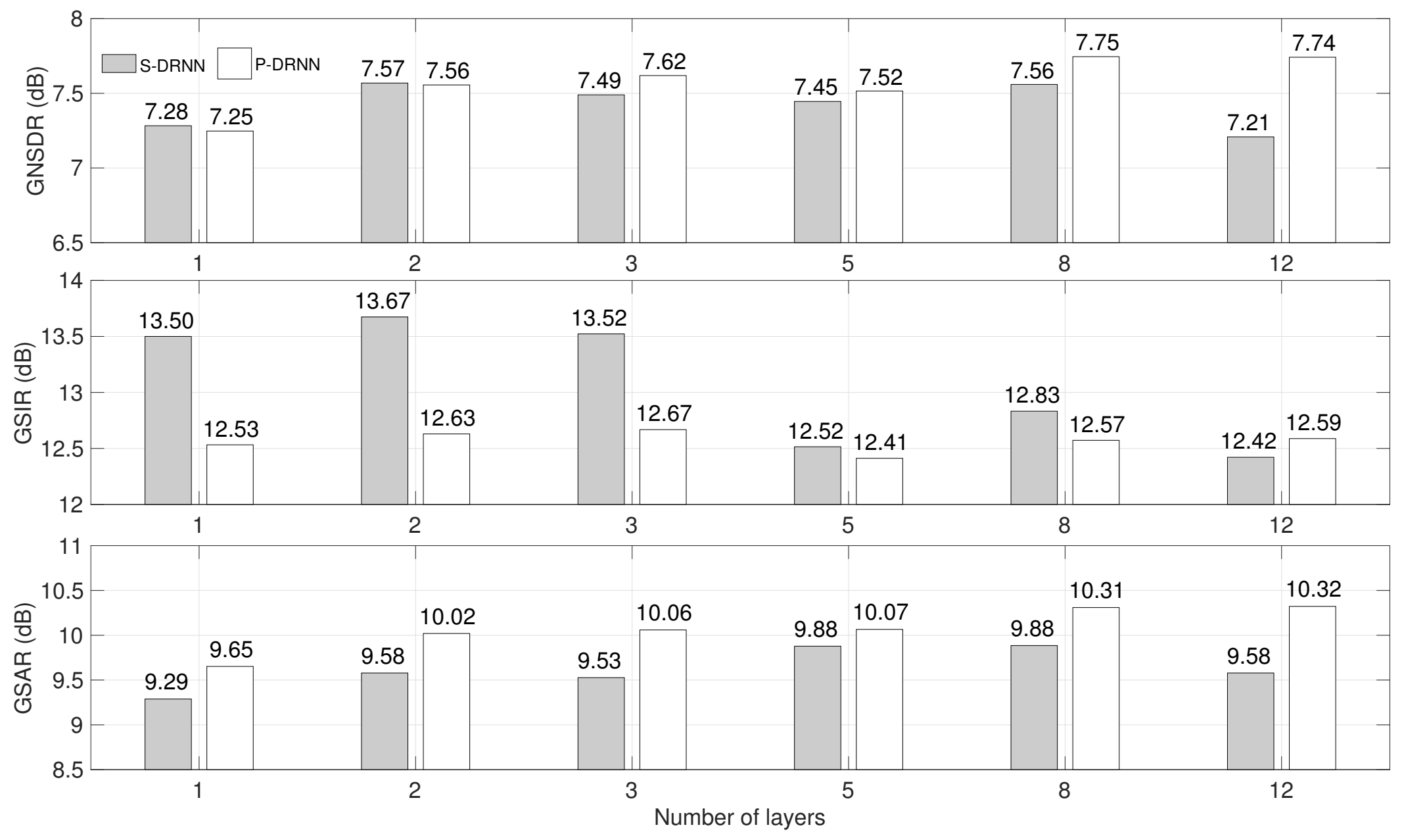

Figure 2: The separation performances of S-RNN and P-DRNN: $T = 10$

We compared our results with other previous works. Table 1 shows the results with unsupervised and supervised settings. For the loss L_2, our model of $T = 10$ obtained 0.47 dB GNSDR gain, 0.61 dB GSIR gain, and 0.33 dB GSAR gain, compared to the best results ('DRNN-2').

Unsupervised			
Model	GNSDR (dB)	GSIR (dB)	GSAR(dB)
RPCA	3.15	4.43	11.09
RPCAh	3.25	4.52	11.10
RPCAh + FASST	3.84	6.22	9.19
Supervised			
Model	GNSDR (dB)	GSIR (dB)	GSAR (dB)
MLRR	3.85	5.63	10.70
RNMF	4.97	7.66	10.03
DRNN-2 (L_2)	7.27	11.98	9.99
P-DRNN ($L_2, T = 4$)	7.36	12.31	9.91
P-DRNN ($L_2, T = 10$)	7.74	12.59	10.32

Table 1: Comparisons of the separation results (in dB) between the proposed method (12-layer) and previous approaches.

Conclusions

We have introduced a new method to deepen RNNs, i.e., Proximal DRNN, to improve separation performance in MSVS. Our design was derived from the primal-dual method, which offered a proximal interlayer structure that induced more effective information transfer between different layers. In numerical tests, the P-DRNN outperformed many previous approaches on the MSVS problem.

References

[1] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces, 2nd ed.* Springer, 2017.