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 is based on the convolutional stacked RNN (S-RNN) by introducing a new 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 exemplar 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 MSS: (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-S-RNN (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 MSVS 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:}

minimum \( \phi_1(x_1) + \phi_2(x_2) + \cdots + \phi_j(x_j) \)
subject to \( \sum_{j=1}^{J} x_j = m_k \) \( \forall j \) (1)

where the variable \( x_j \) corresponds to the \( j \)-th estimated source from the mixture \( m_k \). The goal of Eq. (1) is to decompose each frequency feature vector \( m_k \) into \( x_j \). The proposed method is derived from this optimization problem.

The primal-dual method gives the following iteration:

\[
X^{k+1} = X_k - \frac{1}{\gamma_k} (1 + \rho_k) \nabla F_k(X_k)
\]

where \( \gamma_k \) and \( \rho_k \) are the conjugate of \( \gamma \) and \( \rho \), respectively. The parameters \( \rho_k \) and \( \gamma_k \) are positive. Since Eq. (4) suggests that \( f \) is separable, according to Proposition 2.1.3 in [1], \( \phi_k \) in Eq. (8) can be broken into \( N \) smaller functions that can be carried out independently in parallel.

\[
\phi_k(\bar{y}) = \sum_{j \in \mathcal{J}} \phi_k(\bar{y}_j) \iff \bar{y} = \{ \bar{y}_j \} \in \mathbb{R}^{N \times J}\] (9)

The \( \phi_k \) can be evaluated analytically. In fact, the proximal operator of an indicator function is a projection operator [1].

\[
\text{Prox}_{\sigma}(y) = \begin{cases} 
0, & \text{if } y \leq 0; \\
\frac{y}{\sigma}, & \text{if } 0 < y \leq \sigma; \\
\sigma, & \text{if } y > \sigma.
\end{cases}
\]

The \( \text{Prox}_{\sigma} \) can be simplified as follows,

\[
\text{Prox}_{\sigma}(\bar{y}) = \left[ \begin{array}{c} \text{Prox}_{\sigma}(\bar{y}_1) \\
\vdots \\
\text{Prox}_{\sigma}(\bar{y}_J) \end{array} \right].
\]

Thus, Eq. (9) becomes

\[
\text{Prox}_{\sigma}(\bar{y}) = \left[ \begin{array}{c} \text{Prox}_{\sigma}(\bar{y}_1) \\
\vdots \\
\text{Prox}_{\sigma}(\bar{y}_J) \end{array} \right] = \left[ \begin{array}{c} \text{Prox}_{\sigma}(\bar{y}) \\
\vdots \\
\text{Prox}_{\sigma}(\bar{y}) \end{array} \right]
\]

which implies that all elements of \( \bar{y} \) are \( \leq \sigma \).

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