Fully Complex Deep Neural Network for Phase-Incorporating Monaural Source Separation

{ Yuan-Shan Lee[1], Chien-Yao Wang[1], Shu-Fan Wang[1], Jia-Ching Wang[1], and Chung-Hsien Wu[2] }

[1]Department of Computer Science & Information Engineering, National Central University, Taiwan

[2]Department of Computer Science & Information Engineering, National Cheng Kung University, Taiwan

Main Contributions

1. Unlike conventional DNN-based methods, the developed **fully complex-valued DNN (FCDNN)** directly learns the nonlinear relationship between input mixture and target sources in a fully complex domain.
2. In addition, to reinforce the sparsity of the estimated spectra, a **sparse penalty term** is incorporated into the objective function of the FCDNN. The advantage is that the number of **free parameters** of the FCDNN is reduced, ensuring that the model does not find a poor local minimum during the learning.

Fully Complex-valued DNN

Without loss of generality, a **two-layer FCDNN** is considered, as shown in Fig. 1. The objective function of the FCDNN can be defined as follows,

$$\sum_{n=1}^{N} E_n = \sum_{n=1}^{N} (\mathbf{d}(n) - \mathbf{y}(n))\,(\mathbf{d}(n) - \mathbf{y}(n))^H \in \mathbb{R} \tag{1}$$

where $\mathbf{y}(n) \in \mathbb{C}^{KP}$ is the output, E_n is the n-th partial error term, $\mathbf{d}(n) = (\mathbf{d}_1(n), \mathbf{d}_2(n), ..., \mathbf{d}_P(n)) \in \mathbb{C}^{KP}$ is the spectra of the P sources. Omitting the frame index n, the j-th element of $\mathbf{y}(n)$ can be represented as

$$y_j = x_j^{(2)} = f(\underbrace{\sum_{k=1}^{N_1} w_{jk}^{(2)} \cdot f\left(a_k^{(1)}\right) + b_j^{(2)}}_{a_j^{(2)}}) \in \mathbb{C} \tag{2}$$

where $a_k^{(1)} = \sum_{m=1}^{N_0} w_{km}^{(1)} x_m^{(0)} + b_k^{(1)}$; $f : \mathbb{C} \to \mathbb{C}$ is a nonlinear activation function in the complex domain. Notably, $x_k^{(0)}$,$x_k^{(1)}$, $w_{jk}^{(l)}$, and $b_j^{(l)}$ are **complex-valued**.

Introduction

DNN have become a popular means of separating a target source from a mixed signal. Most of DNN-based methods modify only the magnitude spectrum of the mixture. **The phase spectrum is left unchanged**, which is inherent in the STFT coefficients of the input signal. However, recent studies have revealed that **incorporating phase information** can improve the quality of separated sources. To estimate simultaneously the magnitude and the phase of STFT coefficients, this work paper developed a FCDNN that learns the nonlinear mapping from complex-valued STFT coefficients of a mixture to sources.

Sparse Model Training

This work considers prior knowledge of the **inherent sparse structure** of speech signals in the time-frequency domain. A **sparse constraint** is further imposed on the objective function of the FCDNN.

$$E_n^{\text{sparse}} = E_n + \beta \cdot \sum_{j=1}^{M} D_{\text{KL}}\left(\rho \parallel \hat{\rho}_{nj}\right) \tag{3}$$

where $\hat{\rho}_j = \frac{1}{m}\sum_{i=1}^{m}\left|f(a_j^{(l)})\right|$ denotes the **mean activation** of the j-th hidden unit; M represents the number of neurons in the l-layer, and ρ is the predefined sparse parameter. To train the FCDNN, the **stochastic gradient decent** (SGD) is adopted in our work. $\mathbb{CR}$-calculus [2] is utilized to calculate the partial derivative of E_n^{sparse} with respect to complex-valued parameters. For example, the partial derivative of E_n^{sparse} with respect to $w_{jk}^{(2)}$ can be calculated by,

$$\frac{\partial E_n^{\text{sparse}}}{\partial (w_{jk}^{(2)})} = \frac{\partial E_n}{\partial (w_{jk}^{(2)})^{\Re}} + i \cdot \frac{\partial E_n}{\partial (w_{jk}^{(2)})^{\Im}} + \beta \cdot \left(-\frac{\rho}{\hat{\rho}_{nk}} + \frac{1-\rho}{1-\hat{\rho}_{nk}}\right) \cdot x_k^{*(1)} \tag{4}$$

Experimental Results

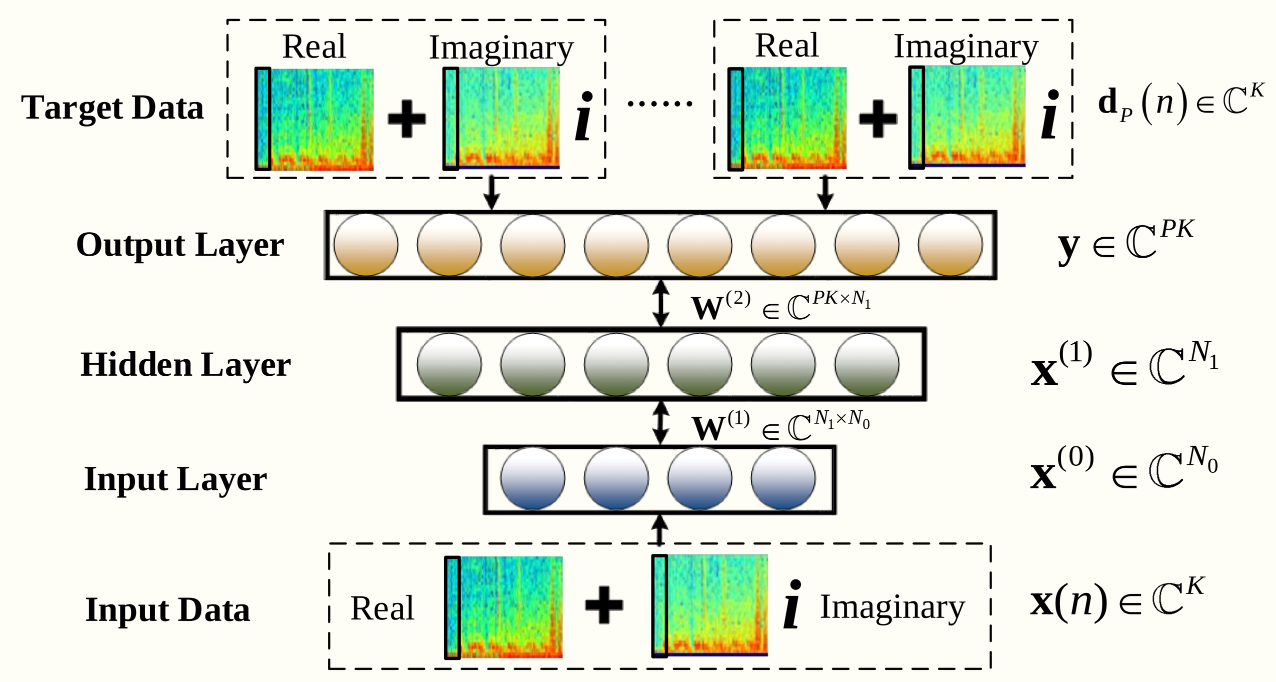

Figure 1: The architecture of FCDNN

The effectiveness of the proposed method is evaluated on the **singing source separation task.** To generate the training and development set, 175 clips of songs are selected from MIR-1K [1]. For the testing set, the remaining 825 clips of songs are used. Two sources ($P = 2$) are mixed to form the mixture. The spectrograms were generated using a 128-point STFT ($K = 65$). A standard DNN-based method: **DNN-M**, is selected as the baseline. Another method: **DNN-RI**, which jointly estimates the real and imaginary components, is also compared to the proposed FCDNN.

Fig. 2 demonstrates that the proposed method **outperformed the baseline methods in terms of SDR and SIR**. However, FCDNN achieved lower SAR compared with the baseline methods. Table 1 shows the average performance in terms of SNR_{fw} and PESQ. FCDNN had a **better PESQ** than DNN-M, but its PESQ was similar to that of DNN-RI. Comparison between FCDNN and FCDNN-S confirmed the power of the additional sparse regularization term.

Table 1: Performance of Speech Quality Measures

Methods	SNR_{fw}	PESQ
Mixture	-0.89±1.29	1.22±0.43
IRM	5.36±1.37	1.99±0.41
DNN-M	0.56±1.66	1.45±0.37
DNN-RI	1.65±2.00	1.53±0.33
FCDNN	1.50±1.90	1.50±0.34
FCDNN-S	1.83±2.02	1.59±0.33

Conclusion

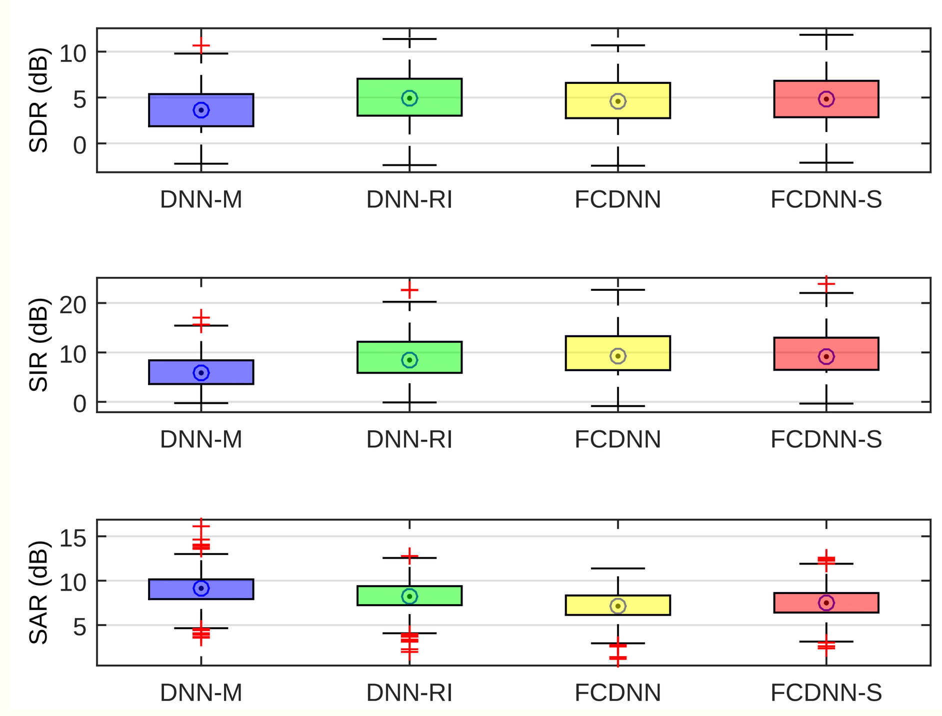

Figure 2: Results of source separation

- Unlike conventional DNN-based methods, the proposed method **operates directly in the complex domain**, and also provides an intuitive way to deal with complex-valued signals.
- Additionally, a sparsity constraint is imposed on the objective function of FCDNN, **enforcing the regularity** of the learned model.
- Experimental results indicate that the proposed method has **higher SDR and SIR** than two state-of-the-art methods.

References

[1] C. L. Hsu and J. S. Jang. On the improvement of singing voice separation for monaural recordings using the mir-1k dataset. 18(2):310–319, 2010.

[2] K. Kreutz-Delgado. The complex gradient operator and the cr-calculus. Technical report, 2009.

Complex-valued Activation

A **complex-valued ReLU** is defined as,

$$\text{ReLU}_{\mathbb{C}}(z) = \begin{cases} z & ,\phi_z \in [0, \frac{\pi}{2}] \\ 0 & ,\text{otherwise} \end{cases}$$

The $\text{ReLU}_{\mathbb{C}}$ is found herein to be **less sensitive to the initialization** of weights than other complex-valued activations, such as tanh and sigmoid, in the source separation task.

Contact Information

Name Yuan-Shan Lee
Lab http://mediasystem.csie.ncu.edu.tw/
Email kg934283@gmail.com
Phone +886 989697417