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,

$$E_n = \sum_{n=1}^{N} (d(n) - y(n))^2 \in \mathbb{R}$$

where $y(n) \in \mathbb{C}^{K \times P}$ is the output, $d(n)$ is the $n$-th partial error term, $d(n) = (d_1(n), d_2(n), ..., d_P(n)) \in \mathbb{C}^{K \times P}$ is the spectra of the $P$ sources. Omitting the frame index $n$, the $j$-th element of $y(n)$ can be represented as

$$y_j = x_j^{(2)} = f_j \left( \sum_{k=1}^{N_j} w_{jk}^{(2)} \cdot f_k \left( x_k^{(1)} + b_k^{(2)} \right) \right) \in \mathbb{C}$$

where $x_k^{(1)} = \sum_{m=0}^{N_m} w_{km}^{(1)} x_m + b_k^{(1)} \in \mathbb{C}$ is a non-linear activation function in the complex domain. Notably, $x_k^{(0)}, x_{jk}^{(1)}, w_{jk}^{(2)}$, and $b_k^{(2)}$ are complex-valued.

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.

$$E_{\text{sparse}} = E_n + \lambda \cdot \sum_{j=1}^{M} D_{KL}(\rho || \hat{\rho}_j)$$

where $\hat{\rho}_j = \frac{1}{N} \sum_n f(a_j^{(0)})$ denotes the mean activation of the $j$-th hidden unit; $M$ represents the number of neurons in the $l$-layer, and $\rho$ is the pre-defined sparse parameter. To train the FCDNN, the stochastic gradient descent (SGD) is adopted in our work. SGD utilizes the partial derivative of $E_{\text{sparse}}$ with respect to complex-valued parameters. For example, the partial derivative of $E_{\text{sparse}}$ with respect to $w_{jk}^{(2)}$ can be calculated by,

$$\frac{\partial E_{\text{sparse}}}{\partial w_{jk}^{(2)}} = \frac{\partial E_n}{\partial x_k^{(2)}} \cdot \frac{\partial x_k^{(2)}}{\partial w_{jk}^{(2)}} + \lambda \cdot \frac{\partial D_{KL}(\rho || \hat{\rho}_j)}{\partial w_{jk}^{(2)}}$$

where $\frac{\partial E_n}{\partial x_k^{(2)}}$, $\frac{\partial x_k^{(2)}}{\partial w_{jk}^{(2)}}$, and $\frac{\partial D_{KL}(\rho || \hat{\rho}_j)}{\partial w_{jk}^{(2)}}$ are calculated by the backpropagation formula.

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.

**Experimental Results**

Table 1: Performance of Speech Quality Measures

<table>
<thead>
<tr>
<th>Methods</th>
<th>SNR_{ref}</th>
<th>PESQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixture</td>
<td>-0.89±1.29</td>
<td>1.22±0.43</td>
</tr>
<tr>
<td>IRM</td>
<td>5.36±1.37</td>
<td>1.99±0.41</td>
</tr>
<tr>
<td>DNN-M</td>
<td>0.56±1.66</td>
<td>1.45±0.37</td>
</tr>
<tr>
<td>DNN-RI</td>
<td>1.65±2.00</td>
<td>1.53±0.33</td>
</tr>
<tr>
<td>FCDNN</td>
<td>1.50±1.90</td>
<td>1.50±0.34</td>
</tr>
<tr>
<td>FCDNN-S</td>
<td>1.83±2.02</td>
<td>1.59±0.33</td>
</tr>
</tbody>
</table>

• 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.

**Conclusion**

The ReLU_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.

**References**
