Combining Fully Convolutional and Recurrent Neural Networks for Single Channel Audio Source Separation

Emad M. Grais and Mark D. Plumbley

{grais, m.plumbley}@surrey.ac.uk

Centre for Vision, Speech and Signal Processing, University of Surrey Guildford, UK.

Overview

Combining different models is a common strategy to build a good single channel audio source separation (SCSS) system. We combine fully convolutional neural networks (FCNs) and recurrent neural networks, specifically, bidirectional long short-term memory recurrent neural networks (BLSTMs). FCNs are good at extracting useful features from the audio data and BLSTMs are good at modeling the temporal structure of the audio signals. Our experimental results show that combining FCNs and BLSTMs achieves better separation performance than using each model individually.

Problem formulation of SCSS

Given a mixture of I audio sources as $y(t) = \Sigma_{i=1}^{I} s_i(t)$, the aim of the SCSS is to find estimates $\hat{s}_i(t)$ for the sources $s_i(t)$, $\forall i$ from the mixed signal $y(t)$. This can be formulated in the short time Fourier transform (STFT) domain as $Y(n,f) = \Sigma_{i=1}^{I} S_i(n,f)$, where $S_i(n,f)$ is the unknown STFT of source $s_i(t)$, $Y(n,f)$ is the STFT of the observed mixed signal $y(t)$, n, and f are the time and frequency indices respectively.

FCN for source separation

The fully convolutional neural network (FCN) consists of an encoder part and a decoder part. The encoder part is composed of repetitions of a convolutional layer and an activation layer. Each convolutional layer consists of filters that extract features from its input layer, the activation layer imposes nonlinearity to the extracted features. The decoder part consists of repetitions of deconvolutional (transposed convolution) layer and an activation layer.

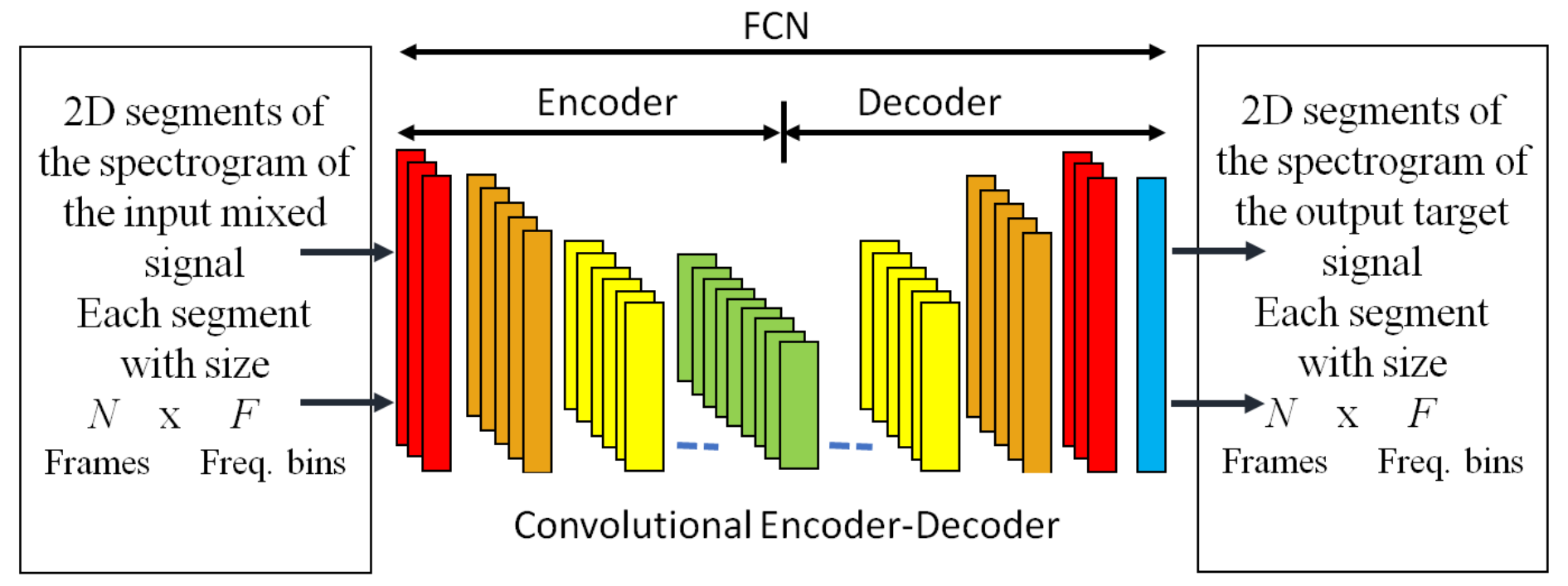

FCN
Encoder
Decoder
2D segments of the spectrogram of the input mixed signal
Each segment with size
N x F
Frames Freq. bins
2D segments of the spectrogram of the output target signal
Each segment with size
N x F
Frames Freq. bins
Convolutional Encoder-Decoder

Figure: The overview of the structure of a FCN that separates one target source from the mixed signal.

The FCN is used to map the magnitude spectrogram of the input mixture into the magnitude spectrogram of the target source. The FCN in this work is a fully 2D convolutional deep neural network. The input and output data for the FCN are 2D signals (magnitude spectrograms) and the filtering is a 2D operator.

BLSTMs and LSTMs for source separation

The Bidirectional Long Short-Term Memory (BLSTM) is a Long Short-Term Memory (LSTM) recurrent neural network that uses contextual information from the past and future of its input/output sequences. Fig. 2 shows the recurrent neural network structure that we use in this work. The hidden layers are BLSTM layers and the output layer is an LSTM layer. The input of the BLSTM are sequences of N consecutive frames from the spectrograms of the mixed signal. The output of the LSTM layer is a spectral mask corresponding to the N consecutive input frames.

The mask represents the contribution of the target source in the input mixture. The spectral mask scales the mixed signal according to the contribution of the target source in the mixed signal as follows:

$$\hat{\mathbf{S}}_i(n,f) = \mathbf{M}_i(n,f) \times \mathbf{Y}(n,f) \quad (1)$$

where $\hat{\mathbf{S}}_i(n,f)$ is the estimate of the magnitude spectrogram of the target source i, $\mathbf{Y}(n,f)$ is the magnitude spectrogram of the mixed signal, and $\mathbf{M}_i(n,f)$ is the output spectral mask from the LSTM layer.

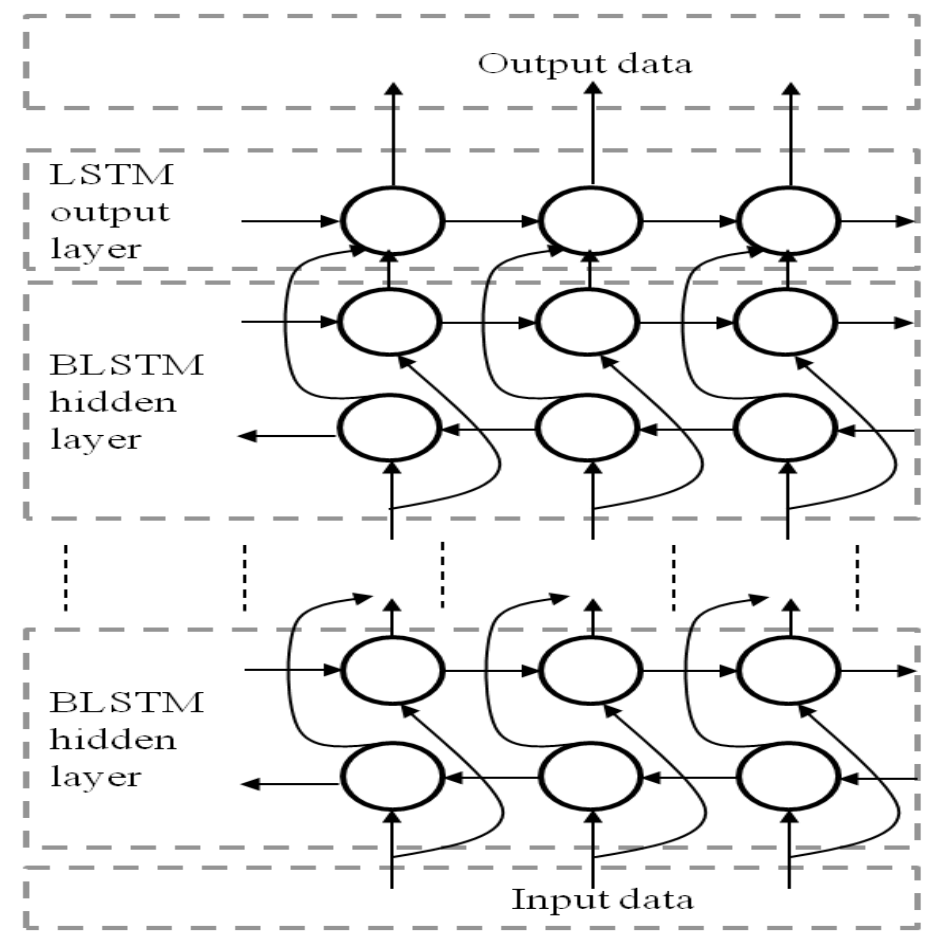


Figure: The unfolded in time of the recurrent neural network that we use in this work.

Training the FCN and BLSTM models

The FCN that separates source i from the mixture is trained to minimize the following cost function:

$$C_i = \underset{n,f}{\Sigma} \left(\mathbf{Z}_i(n,f) - \mathbf{S}_{\mathrm{tr}_i}(n,f)\right)^2 \quad (2)$$

where $\mathbf{Z}_i$ is the actual output of the last layer of the FCN of source i and $\mathbf{S}_{\mathrm{tr}_i}$ is the reference output signal for source i.

The BLSTM that separates source i from the mixture is trained to minimize the following cost function:

$$D_i = \underset{n,f}{\Sigma} \left(\mathbf{Q}_i(n,f) - \mathbf{M}_{\mathrm{tr}_i}(n,f)\right)^2 \quad (3)$$

where $\mathbf{Q}_i$ is the actual output of the last layer of the BLSTM (the LSTM layer) of source i and $\mathbf{M}_{\mathrm{tr}_i}(n,f)$ is the reference spectral mask for source i. The reference spectral mask for source i is computed from the training data as follows:

$$\mathbf{M}_{\mathrm{tr}_i}(n,f) = \frac{\mathbf{S}_{\mathrm{tr}_i}(n,f)}{\Sigma_j^I \mathbf{S}_{\mathrm{tr}_j}(n,f)}, \quad \forall i. \quad (4)$$

The inputs of the FCN and BLSTM are 2D-segments from the magnitude spectrogram $\mathbf{Y}_{\mathrm{tr}}$ of the mixed signal

Combining FCNs and BLSTMs for source separation

The aim of the FCN-BLSTM combination is to build a model that captures the spectro-temporal characteristics of the audio data better than each model (FCN or BLSTM) individually. The FCN is used first to extract an initial estimate of the magnitude spectrogram of the target source from the input sequence. The initial estimate is then passed to the BLSTM network to enhance the output sequence of the FCN. The FCN is good at extracting useful features from the input signals and the BLSTM is good at modeling the temporal structure of the input sequence.

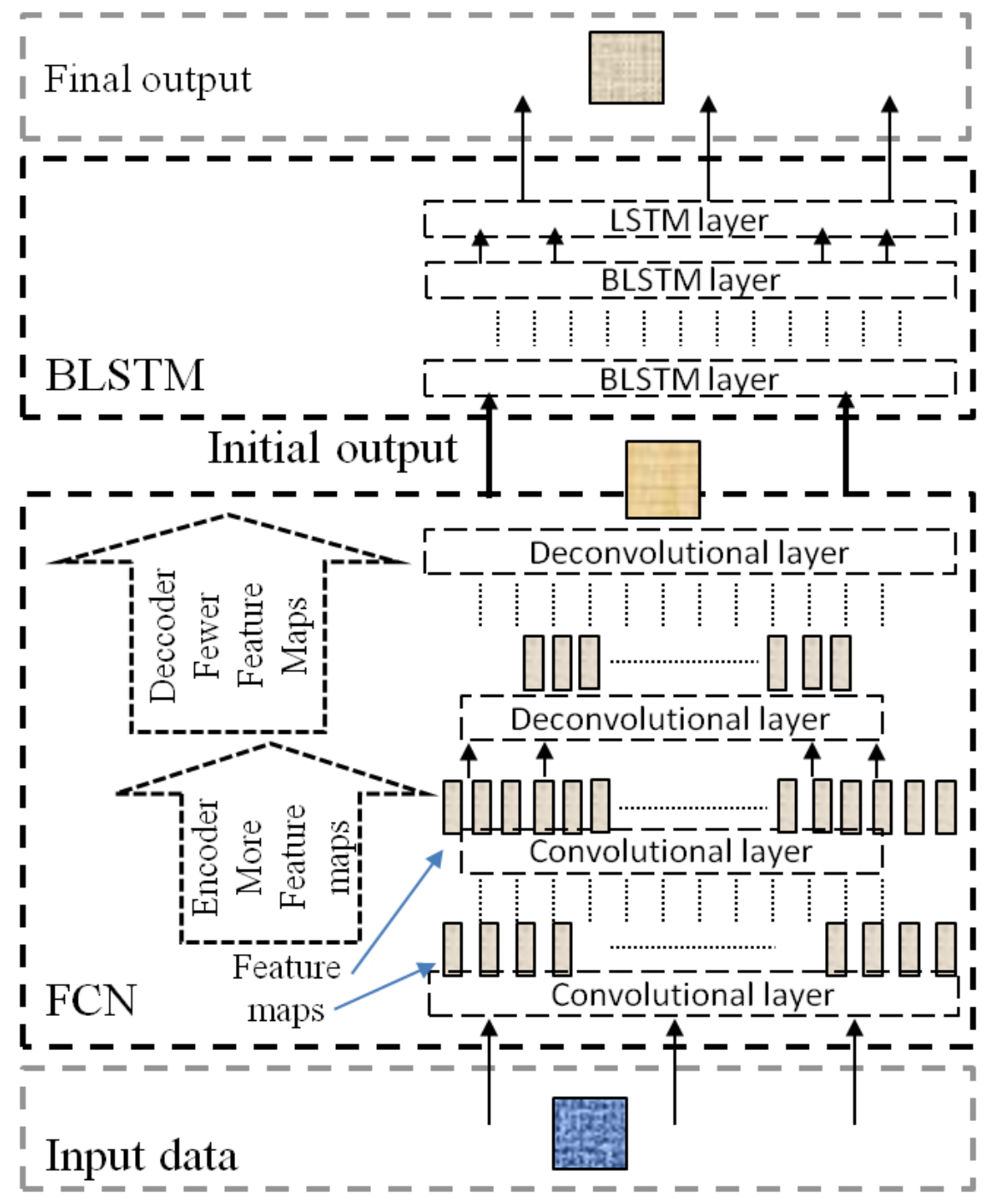


Figure: The proposed combination of FCN and BLSTM models for SCSS.

Joint training for the combined models

The trained layers of the FCN and BLSTM models for source i are stacked to form the combination of the FCN and BLSTM models: FCN-BLSTM. A joint training is then run over the combined model (FCN-BLSTM) to refine the parameters of the trained models to fit the training data well. The input of the combined FCN-BLSTM model during training is the magnitude spectrogram $\mathbf{Y}_{\mathrm{tr}}$ of the mixed signal and the reference output is the reference spectral mask $\mathbf{M}_{\mathrm{tr}_i}(n,f)$ computed from Eq. (4). The training of the FCN-BLSTM model is done by minimizing the cost function in Eq. (3).

Note that, when the BLSTM model was trained individually, the input of the BLSTM was the magnitude spectrogram $\mathbf{Y}_{\mathrm{tr}}$ of the input mixed signal, but when the BLSTM is trained within the combined FCN-BLSTM model, the input of the BLSTM is the output of the FCN. Similarly, the updating of the FCN parameters in the combined model is based on the propagated errors between the output of the BLSTM and the reference mask $\mathbf{M}_{\mathrm{tr}_i}(n,f)$ and not based on the magnitude spectrograms as it was when it was trained individually. These differences in the training conditions for the FCN and BLSTM in the combined model makes the joint training of the combined FCN-BLSTM model necessary.

Experiments

We applied our proposed SCSS using FCN-BLSTM model to separate the singing voice from a group of songs from the SiSEC-2015-MUS-task dataset.

We compared the performance of the combined FCN-BLSTM model with using each model (FCN and BLSTM) individually and also with the feedforward neural network (FFN).

The table shows the number of layers, the type of each layer, the number of filters/units in each layer, the size of the filters, and the total number of parameters for the FCN, BLSTM, FCN-BLSTM and FFN models.

The activation function in the FNN and FCN layers is the rectified linear unit (ReLU). The activation function in the BLSTM is sigmoid in the forward direction and hard-sigmoid in the recurrent direction.

We built the combined model from the trained FCN model and only the first and last layers from the trained BLSTM model. We then retrained/fine-tuned the parameters of the combined model. By removing the middle layer of the BLSTM model in the combined model, the number of the parameters in the combined model becomes less than the number of parameters in the BLSTM model only.

The parameters for FCN, BLSTM, and FFN networks were initialized randomly. The parameters of the combined model FCN-BLSTM were initialized from their corresponding parameters from the trained FCN and BLSTM models.

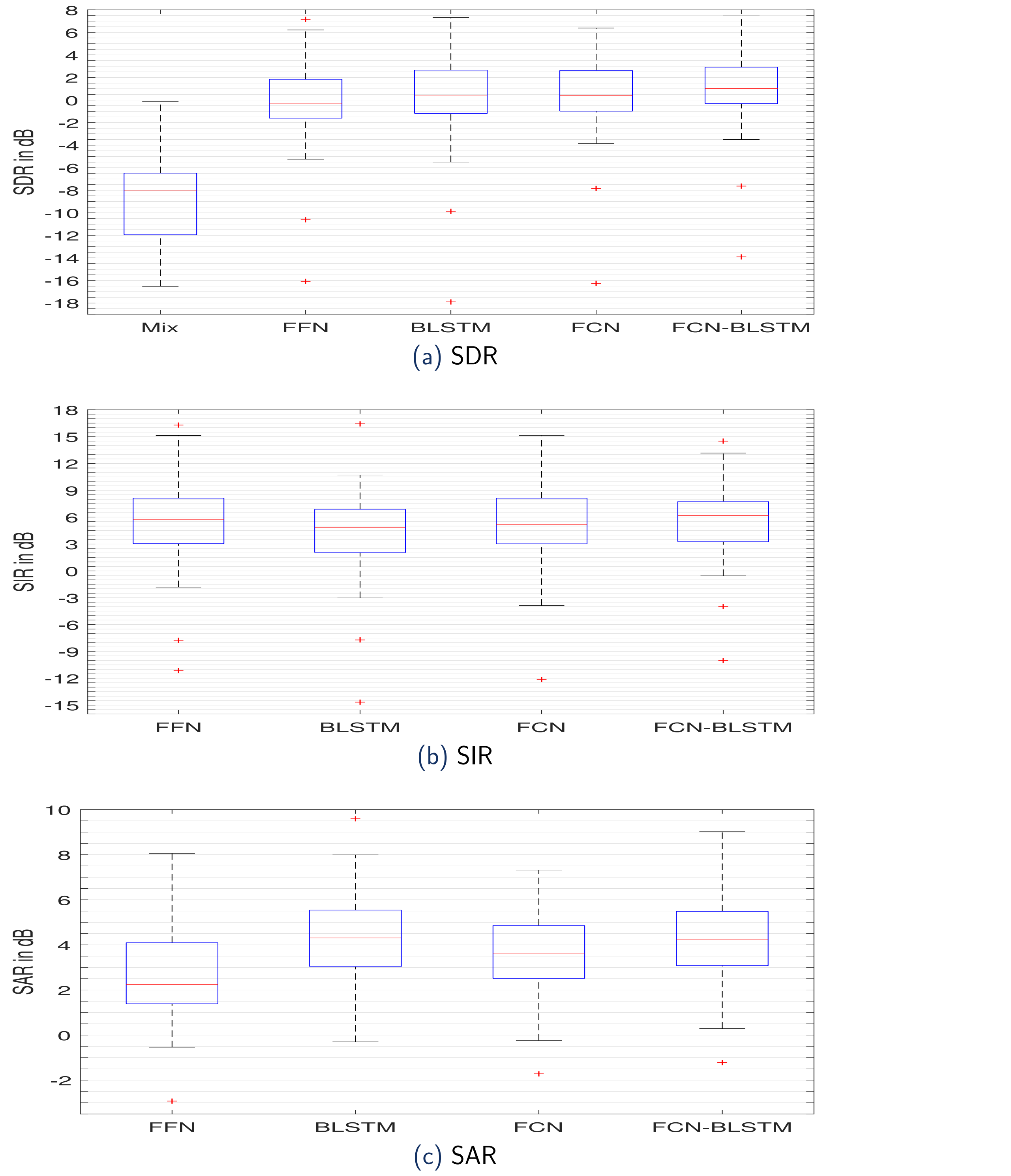
FCN, BLSTM, FCN-BLSTM, and FFN model summary				
	The input/output data with size 15 frames and 1025 frequency bins			
Layer number	FCN	BLSTM	FCN-BLSTM	FFN
1	Conv2D [12,(15,39)]	BLSTM 2050 units	Conv2D [12,(15,39)]	DENS 1025 units
2	Conv2D [22,(9,19)]	BLSTM 2050 units	Conv2D [22,(9,19)]	DENS 1025 units
3	Conv2D [32,(5,5)]	LSTM 1025 units	Conv2D [32,(5,5)]	DENS 1025 units
4	Conv2D [22,(9,19)]		Conv2D [22,(9,19)]	
5	Conv2D [12,(15,39)]		Conv2D [12,(15,39)]	
6	Conv2D [1,(15,1025)]		Conv2D [1,(15,1025)]	
7			BLSTM 2050 units	
8			LSTM 1025 units	
Total number of parameters	529,189	172,339,400	71,992,189	4,206,600

Table: The detail information about the structures of the FCN, BLSTM, FCN-BLSTM, and FFN neural networks. For example "Conv2D[12,(15,39)]" denotes 2D convolutional layer with 12 filters and the size of each filter is 15×39 where 15 is the size of the filter in the time-frame direction and 39 in the frequency direction of the spectrogram.

All networks were trained using back-propagation with gradient descent optimization using Adam with parameters: $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, batch size 100, and a learning rate 0.0001. The maximum number of epochs was 25.

The quality of the separated vocal source was measured using the signal to distortion ratio (SDR), signal to interference ratio (SIR), and signal to artifact ratio (SAR).

In Figs. 4a to 4c, the differences between each pair of models for SDR are statistically significant except the difference between the BLSTM and FCN models. The differences between each pair of models for SIR are not statistically significant except the differences between the BLSTM model and all other models. The differences between each pair of models for the SAR are statistically significant except the difference between the BLSTM and FCN-BLSTM models.



(a) SDR

(b) SIR

(c) SAR

Figure: (a) SDR, (b) SIR, and (c) SAR (values in dB) for the separated vocal signals of using: deep fully connected feedforward neural network (FFN), BLSTM, fully convolutional neural networks (FCN), and the proposed combination of FCN and BLSTM (FCN-BLSTM). "Mix" denotes the input mixed signal.

The results indicate that combining the FCN and BLSTM models achieves the best performance of the FCN in SIR (more separation) and the best performance of the BLSTM in SAR (less artifacts). The proposed method of using FCN followed by BLSTM (FCN-BLSTM) works better than BLSTM, even with fewer parameters in the FCN-BLSTM than the BLSTM.

Acknowledgements

This work is supported by grant EP/L027119/2 from the UK Engineering and Physical Sciences Research Council (EPSRC).