TasNet: time-domain audio separation network for real-time, single-channel speech separation

COLUMBIA | ENGINEERING The Fu Foundation School of Engineering and Applied Science

Yi Luo, Nima Mesgarani — Department of Electrical Engineering, Columbia University

Single-channel speech separation

- Deep learning systems have significantly advanced the state of the problem [1, 2, 3, 4].
- Time-frequency mask estimation, which relies on Short-time Fourier transform (STFT), remains the mainstream method.
- Most of the systems are noncausal that cannot be implemented in applications or devices that require real-time processing.

Drawbacks of STFT

- It is unclear if spectrogram is the optimal feature for separation.
- Phase information is often lost, theoretical performance upper-bound exists.
- Trade-off between latency and frequency resolution needs to be considered.
- STFT and its inverse lead to higher system latency.

Time-domain modeling for separation

Targets:

- Replace STFT, learn a better front-end specialized for separation.
- Enables real-time, low-latency processing.

Ideas:

- 1-D convolution and deconvolution autoencoder as an adaptive front-end.
- Nonnegativity constraint on encoder output.
- Separation as mask estimation on the learnt front-end.
- Learnable, frequency selective filters as decoder.

Problem description

Mixture waveform as the summation of sources:

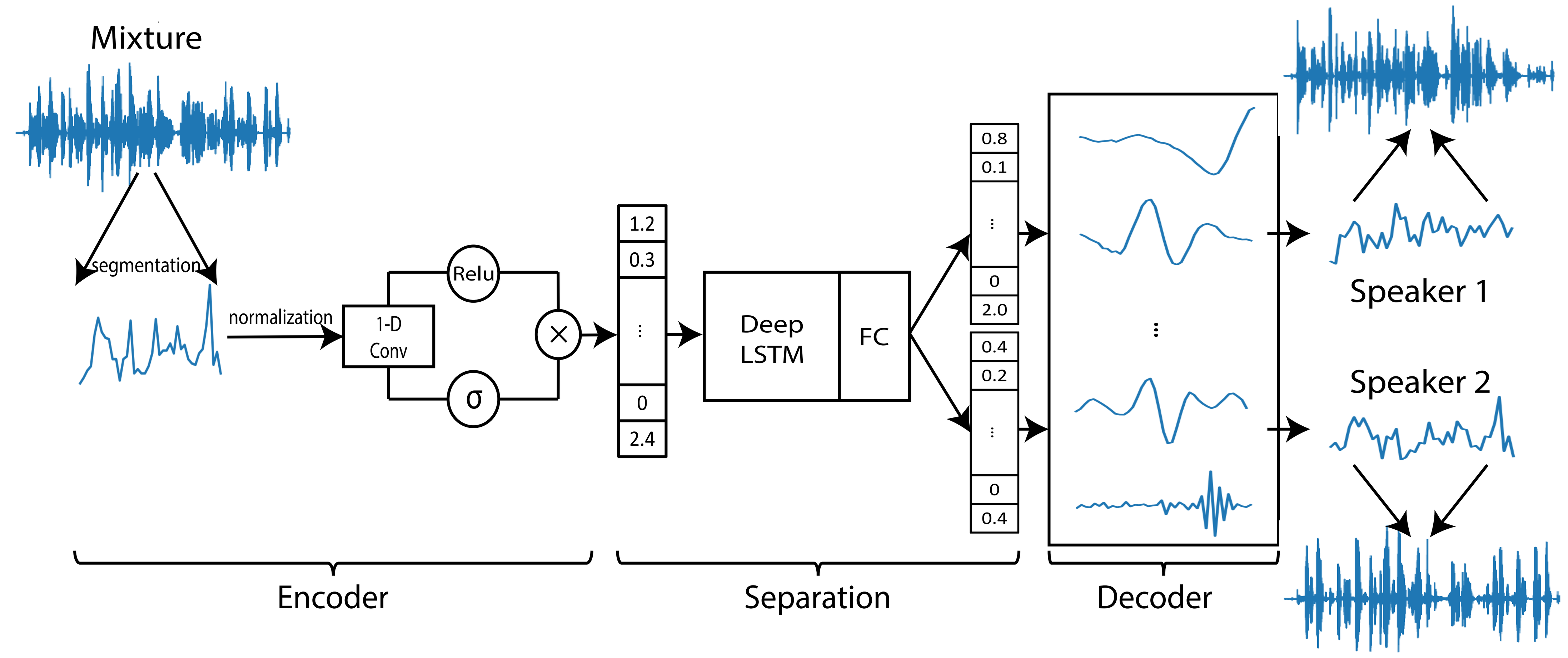
$$x(t) = \sum_{i=1}^{C} s_i(t)$$

Split signals into segments:

$$\begin{cases} \mathbf{x}_k = x(t) \\ \mathbf{s}_{i,k} = s_i(t) \end{cases} \quad t \in [kL, (k+1)L),\ k = 1, 2, \ldots, K$$

Represent signals by ***nonnegative*** weighted sum of a set of basis signals (a nonnegative autoencoder):

$$\begin{cases} \mathbf{x} = \mathbf{w}\mathbf{B} \\ \mathbf{s}_i = \mathbf{d}_i\mathbf{B} \end{cases} \quad \text{s.t.}\ \ \mathbf{w} = \sum_{i=1}^{C} \mathbf{d}_i$$



Source weight matrices can be treated as masks applied on the mixture weight matrix (separation module):

$$\mathbf{w} = \sum_{i=1}^{C} \mathbf{w} \odot (\mathbf{d}_i \oslash \mathbf{w}) := \mathbf{w} \odot \sum_{i=1}^{C} \mathbf{m}_i$$

$$\mathbf{d}_i = \mathbf{m}_i \odot \mathbf{w}$$

Relation with traditional methods

- The autoencoder is similar to independent component analysis (ICA) [5] with nonnegative mixing matrix and semi-nonnegative matrix factorization (semi-NMF) [6].
- Unlike those methods, the weights and basis signals are fitted in a nonnegative convolutional autoencoder framework, which is jointly trained with the separation module.

Model design

Encoder: Gated 1-D convolution

$$\mathbf{w}_k = ReLU(\mathbf{x}_k \circledast \mathbf{U}) \odot \sigma(\mathbf{x}_k \circledast \mathbf{V}), \quad k = 1, 2, \ldots, K$$

Separator: Deep LSTM + dense layer with Softmax activation for mask estimation

Decoder: Linear 1-D deconvolutional layer

Objective function: Scale-invariant SNR (SI-SNR)

$$\mathbf{s}_{target} = \frac{\langle \hat{\mathbf{s}}, \mathbf{s} \rangle \mathbf{s}}{\|\mathbf{s}\|^2}$$

$$\mathbf{e}_{noise} = \hat{\mathbf{s}} - \mathbf{s}_{target}$$

$$\text{SI-SNR} := 10 \log_{10} \frac{\|\mathbf{s}_{target}\|^2}{\|\mathbf{e}_{noise}\|^2}$$

Experiment results

Data:
- WSJ0-2mix dataset, 30 hours of training data/10 hours of validation data/5 hours of test data
- Downsample to 8k Hz sample rate

Network:
- 5 ms long (40 samples) 1-D filters in encoder and decoder
- 500 filters (channels)
- 500/1000 hidden units in LSTM layers with noncausal/causal settings
- 1000 hidden units in dense layer

Training:
- Batch size: 128
- Learning rate: 1e-3, halve after no new best model in validation set is found in 3 consecutive epochs
- Curriculum training: First train on 0.5s long segments, then continue training on 4s long segments
- Optimizer: Adam

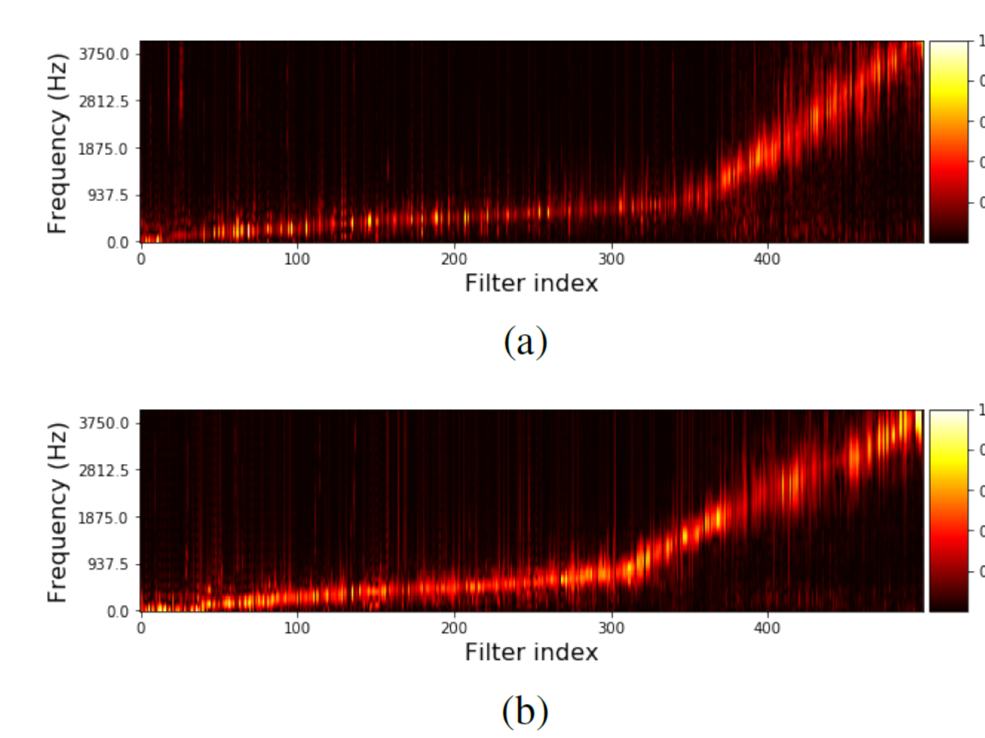


Fig. 2. Frequency response of basis signals in (a) causal and (b) noncausal networks.

Table 1. SI-SNR (dB) and SDR (dB) for different methods on WSJ0-2mix dataset.

Method	Causal	SI-SNRi	SDRi
uPIT-LSTM [4]	✓	–	7.0
TasNet-LSTM	✓	7.7	**8.0**
DPCL++ [3]	×	**10.8**	–
DANet [5]	×	10.5	–
uPIT-BLSTM-ST [4]	×	–	10.0
TasNet-BLSTM	×	**10.8**	**11.1**

Table 2. Minimum latency (ms) of causal methods.

Method	T_i	T_p	T_{tot}
uPIT-LSTM [4]	32	–	>32
TasNet-LSTM	5	0.23	**5.23**

Conclusion

- Experiments show that TasNet has advantage on both separation performance and system latency.
- The 1-D convolutional autoencoder can be an adaptive frontend specified for the task.
- The same procedure can be applied to various of other tasks in audio processing.

Future works

- Further improve the performance of TasNet.
- Investigate the choice of number/length/overlap in the convolutional autoencoder.
- Look into the learnt representation and compare it with STFT.
- Test this system in other audio processing tasks.

References

[1] Xiao-Lei Zhang and DeLiang Wang, "A deep ensemble learning method for monaural speech separation," IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), vol. 24, no. 5, pp. 967–977, 2016.

[2] Yusuf Isik, Jonathan Le Roux, Zhuo Chen, Shinji Watanabe, and John R Hershey, "Single-channel multi-speaker separation using deep clustering," Interspeech 2016, pp. 545–549, 2016.

[3] Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 10, pp. 1901–1913, 2017.

[4] Yi Luo, Zhuo Chen, and Nima Mesgarani, "Speakerindependent speech separation with deep attractor network," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 4, pp. 787–796, 2018.

[5] Fa-Yu Wang, Chong-Yung Chi, Tsung-Han Chan, and Yue Wang, "Nonnegative least-correlated component analysis for separation of dependent sources by volume maximization," IEEE transactions on pattern analysis and machine intelligence, vol. 32, no. 5, pp. 875–888, 2010.

[6] Chris HQ Ding, Tao Li, and Michael I Jordan, "Convex and semi-nonnegative matrix factorizations," IEEE transactions on pattern analysis and machine intelligence, vol. 32, no. 1, pp. 45–55, 2010.