Sandglasset: A Light Multi-Granularity Self-Attentive Network for Time-Domain Speech Separation

Max W. Y. Lam, Jun Wang, Dan Su, Dong Yu

Tencent AI Lab

ICASSP 2021

Outline of This Presentation

Next Section...

Single-Channel Speech Separation

- Speech separation is a fundamental component for many downstream speech processing tasks.
- Single-channel speech separation has recently been advanced by the time-domain audio separation networks (TasNets) (Luo and Mesgarani, 2018).
- Recent development of TasNets:
  1. Bi-LSTM TasNet (Luo and Mesgarani, 2018)
  2. Conv-TasNet (Luo and Mesgarani, 2019)
  3. DPRNN (Luo, Z. Chen, and Yoshioka, 2019)
  4. DPTNet (J. Chen, Mao, and Liu, 2020)
  5. Gated DPRNN (Nachmani, Adi, and Wolf, 2020)
  6. Wavesplit (Zeghidour and Grangier, 2020)
  7. GALR (Lam et al., 2021)

State-of-the-art Speech Separation Models

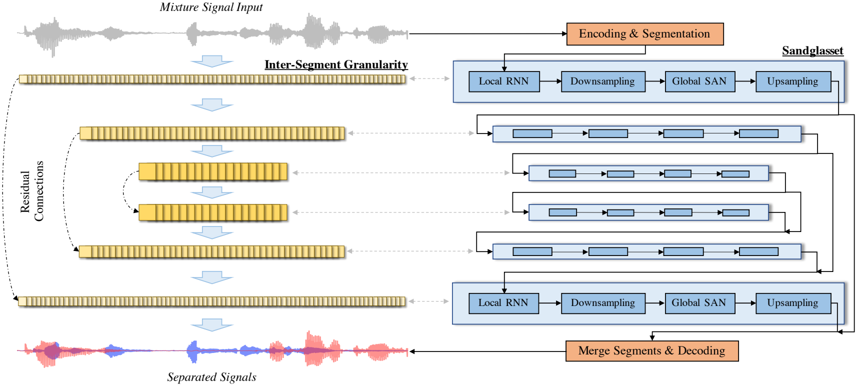
- State-of-the-art (SOTA) models both employ a dual-path technique, which is to process the segment sequence in an intra-segment (local) direction and an inter-segment (global) direction alternatively.
- In our previous work of GALR, we found that self-attentive networks (SANs) are superior over RNNs in modeling the inter-segment sequence.
- SAN can connect every element to another element with a direct path (i.e., in $\mathcal{O}(1)$ time), in contrast to $\mathcal{O}(N)$ time in RNNs.

Essence of Multi-Granularity

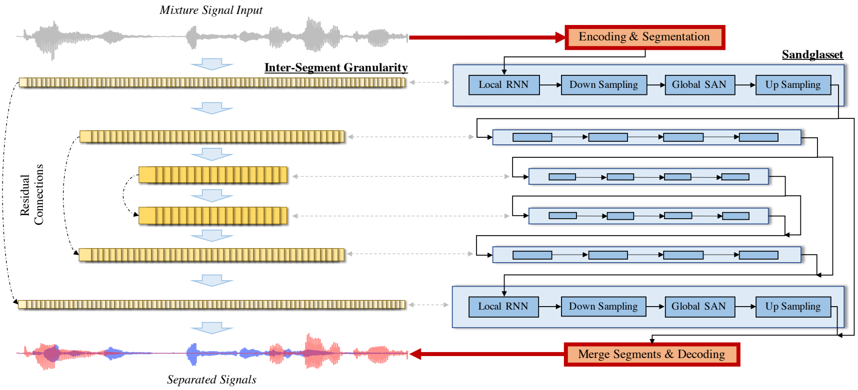
- **Existing Method**: Use a fixed segment size unchanged throughout all layers.
- **Fact**: Speech signals contain different level of contexts, e.g., phonemes, syllables, or words, at different time scales.
- **Our Observation**: SANs have superior capabilities in modeling sequences of high-level contexts, as examined in LM and in NLP.
- **Our Idea:** Design a novel network that allows SANs to capture multi-granularity information for enhancing contextual modeling and computational efficiency.
- **Our Proposed:** Sandglasset, named for its sandglass shape and its modest model size and complexity.

Next Section...

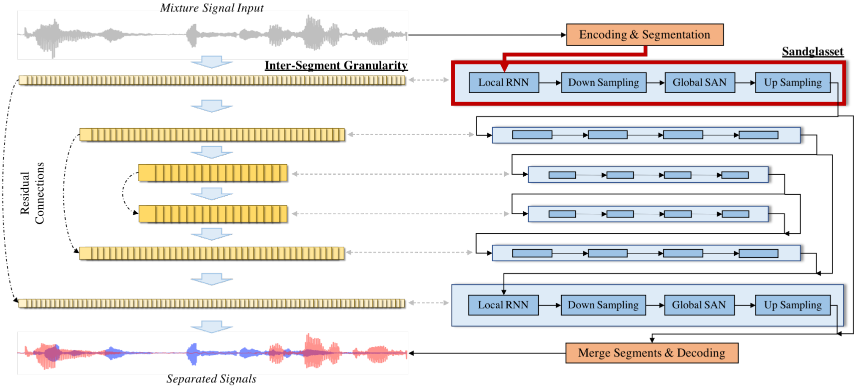
Overall Architecture



Encoding, Segmentation, Overlapadd & Decoding



Sandglasset Block



Modules within a Sandglasset block

- Each Sandglasset block consists of these modules:
  1. A local RNN for processing the intra-segment sequence and modeling locality;
  2. A global SAN for processing the inter-segment sequence and to capture the global dependencies.
  3. A downsampling and an upsampling modules surrounding the global SAN for changing the context granularity.
- For the b-th block, $\mathcal{X}_b \in \mathbb{R}^{D\times K\times S}$ is the block input enclosing S segments each containing K-length D-dimensional features.

Computations within a Sandglasset block

Mathematically, a Sandglasset block computes the following:

$$\mathcal{Y}_b^{LR} = \left[\mathrm{Linear}\left(\mathrm{RNN}_b\left(\mathcal{X}_b[:,:,s]\right)\right), s = 1, \ldots, S\right], \tag{1}$$

$$\mathcal{Y}_b^{GA} = \mathrm{US}_b\left(\mathrm{SAN}_b\left(\mathrm{DS}_b\left(\mathrm{LN}\left(\mathcal{Y}_b^{LR}\right) + \mathcal{X}_b\right)\right)\right), \tag{2}$$

where

$$\mathrm{SAN}(\mathcal{X}) = \left[\mathrm{SelfAttn}\left(\mathrm{LN}\left(\mathcal{X}[:,k,:]\right) + \mathrm{P}\right), k = 1, \ldots, K\right], \tag{3}$$

Downsampling and Upsampling Operations

The downsampling and upsampling operations are defined as

$$\mathrm{DS}_b(\mathcal{X}) = \begin{cases} \mathrm{Conv1D}_K\left(\mathcal{X}; 4^b\right) & \text{if} \quad b \leq N/2; \\ \mathrm{Conv1D}_K\left(\mathcal{X}; 4^{N-b-1}\right) & \text{if} \quad b > N/2. \end{cases} \tag{4}$$

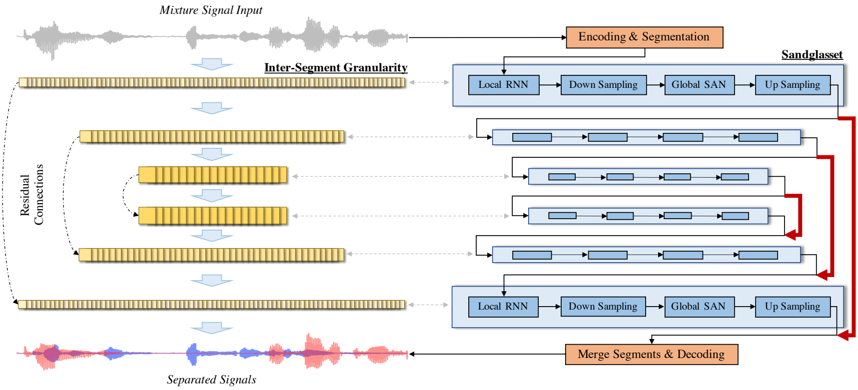
$$\mathrm{US}_b(\mathcal{X}) = \begin{cases} \mathrm{ConvTrans1D}_K\left(\mathcal{X}; 4^b\right) & \text{if} \quad b \leq N/2; \\ \mathrm{ConvTrans1D}_K\left(\mathcal{X}; 4^{N-b-1}\right) & \text{if} \quad b > N/2, \end{cases} \tag{5}$$

where

$$\mathrm{Conv1D}_K(\mathcal{X}; \tau) \in \mathbb{R}^{D \times \lceil K/\tau \rceil \times S} \tag{6}$$

$$\mathrm{ConvTrans1D}_K(\mathcal{X}; \tau) \in \mathbb{R}^{D \times K\tau \times S} \tag{7}$$

Residual Connections to Prevent Information Loss



Residual Connections to Prevent Information Loss

- One highlight of our work is to add residual connections between pairs of Sandglasset blocks that are of the same granularity.
- Skip connections are useful to prevent information loss after passing through the middle blocks, where the granularity is at the coarsest scale.
- Mathematically, we define

$$\mathcal{X}_{b+1}^{LR} = \begin{cases} \mathcal{Y}_b^{GA} & \text{if} \quad b \leq N/2; \\ \mathcal{Y}_b^{GA} + \mathcal{Y}_{N-b+1}^{GA} & \text{if} \quad b > N/2. \end{cases} \tag{8}$$

Next Section...

Performances on WSJ0-2mix

Model	Params.	SI-SNRi	SDRi
BLSTM-TasNet	23.6M	13.2	13.6
Conv-TasNet	8.8M	15.3	15.6
DPRNN	2.6M	18.8	19.1
DPTNet	2.7M	20.2	20.6
Sandglasset (w/o RES)	**2.3M**	20.1	20.3
Sandglasset (SG)	**2.3M**	20.3	20.5
Sandglasset (MG)	**2.3M**	20.8	21.0
Sandglasset (MG) + PT	**2.3M**	**21.0**	**21.2**
Gated DPRNN + Spk ID	7.5M	20.1	-
Wavesplit + Spk ID	†42.5M	**21.0**	**21.2**

Performance on WSJ0-3mix

Model	Params.	SI-SNRi	SDRi
Conv-TasNet	8.8M	12.7	13.1
DPRNN	2.6M	14.7	-
Sandglasset (MG)	**2.3M**	**17.1**	**17.4**
Gated DPRNN + Spk ID	7.5M	16.7	-
Wavesplit + Spk ID	†42.5M	**17.3**	**17.6**

Cost Analysis

Model	Params.	**Memory** (GB)	**GFLOPs** (10^9)
DPRNN	2.6M	1.97	84.7
Sandglasset	**2.3M**	**0.82 (↓58.4%)**	**28.8 (↓66.0%)**

- We measured the runtime memory and the floating-point operations (FLOPs) for processing each second of mixture input during training.
- We compared it to the SOTA model that is comparable in size – DPRNN.

Next Section...

Conclusion

- To conclude, this paper proposes a novel network named Sandglasset for time-domain speech separation.
- Sandglasset applies a downsampling-upsampling mechanism to the global SAN for modeling multi-granularity contexts.
- As the smallest TasNet in size, Sandglasset achieved the state-of-the-art results on two benchmark datasets.
- Sandglasset is also low-cost in terms of memory and computations, which suggests it a more practical model for industrial deployment.