Sandglasset: A Light Multi-Granularity Self-attentive Network For Time-Domain Speech Separation

Tencent AI Lab

Max W. Y. Lam* Jun Wang* Dan Su* Dong Yu†

* Tencent AI Lab, Shenzhen, China
† Tencent AI Lab, Bellevue WA, USA

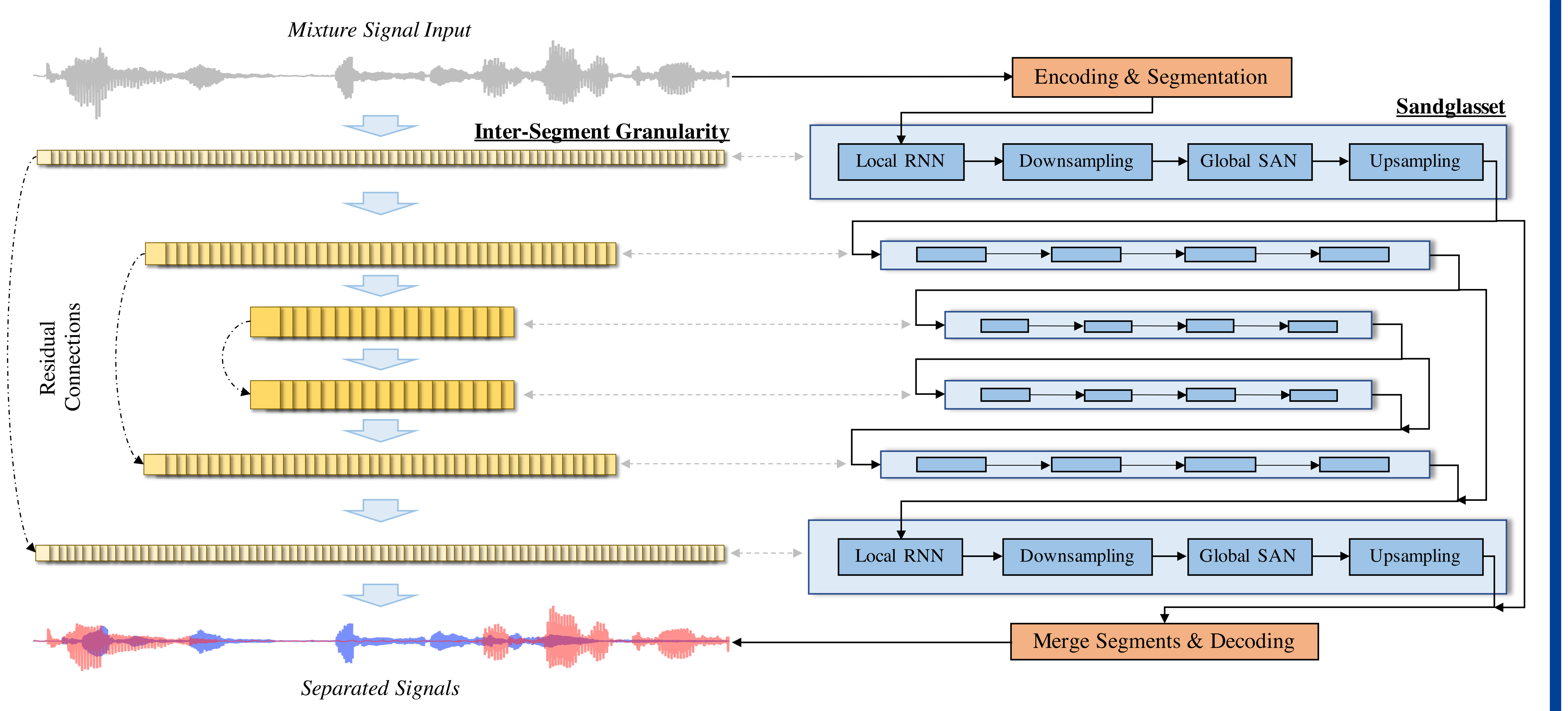
Introduction

- 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).
- 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.

Multi-Granularity

- **Existing Method**: Use a fixed segment size unchanged throughout all layers; The global contexts being learnt in the inter-segment sequence is limited to one manually defined scale.
- **Fact**: Speech signals contain different level of contexts, for example, the phoneme scale, the syllable scale, and the word scale, what we refer to as the multi-granularity.
- **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.

Sandglasset



Sandglasset block

- **Sandglasset block**: Each consists of these modules: 1) A local RNN; 2) A global SAN; 3) A downsampling and an upsampling modules 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.
- Mathematically, a Sandglasset block computes the following:

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

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

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

- The downsampling and upsampling operations are defined as

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

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

where

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

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

Results

1) 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**

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**

3) 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%)**

Conclusions

- 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 and is also low-cost in terms of memory and computations.