Two-Step Sound Source Separation: Training on Learned Latent Targets

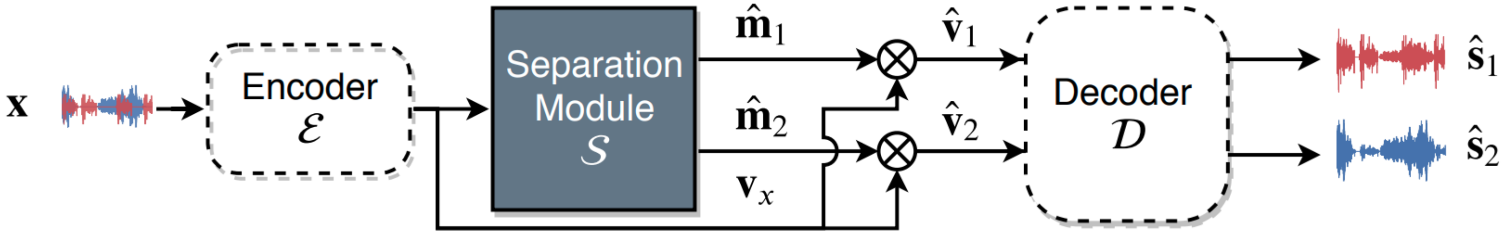
Efthymios Tzinis[1], Shrikant Venkataramani[1], Zhepei Wang[1], Cem Subakan[2], Paris Smaragdis[1,3]

[1]University of Illinois at Urbana-Champaign
[2]Mila–Quebec Artificial Intelligence Institute
[3]Adobe Research

International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2020

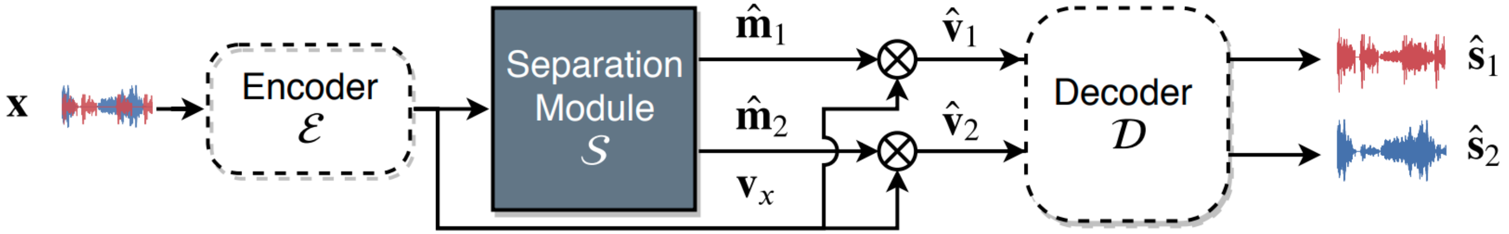
End-to-end source separation

- Time-domain audio source separation
  - Directly optimizing **all parts jointly** using a time-domain loss
    - Scale-Invariant Signal to Distortion Ratio (SI-SDR)



End-to-end source separation

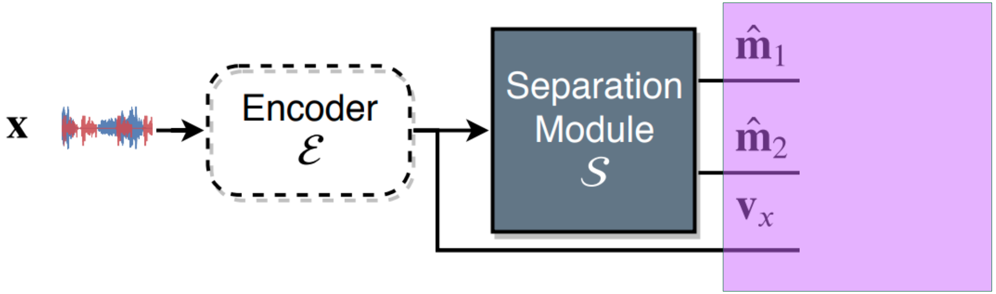
- Time-domain audio source separation
  - Directly optimizing **all parts jointly** using a time-domain loss
    - Scale-Invariant Signal to Distortion Ratio (SI-SDR)



- Mask-based architecture

End-to-end source separation

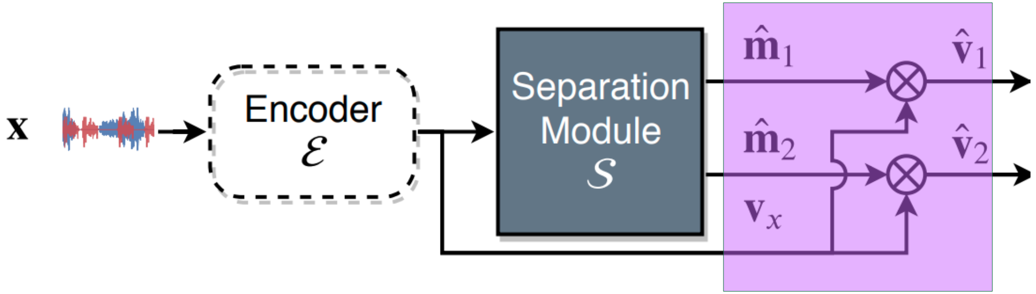
- Time-domain audio source separation
  - Directly optimizing **all parts jointly** using a time-domain loss
    - Scale-Invariant Signal to Distortion Ratio (SI-SDR)



- Mask-based architecture
  - Estimate masks in the latent space

End-to-end source separation

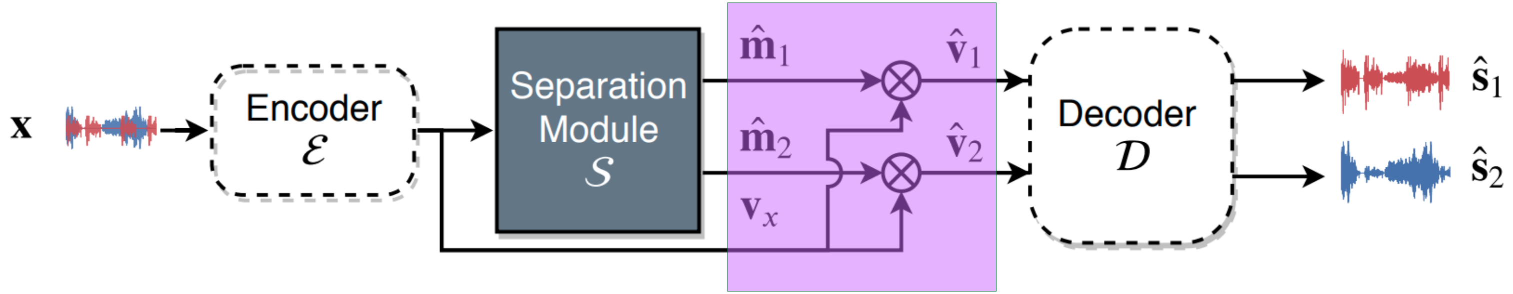
- Time-domain audio source separation
  - Directly optimizing **all parts jointly** using a time-domain loss
    - Scale-Invariant Signal to Distortion Ratio (SI-SDR)



- Mask-based architecture
  - Estimate masks in the latent space
  - Apply masks on the latent representation of the mixture

End-to-end source separation

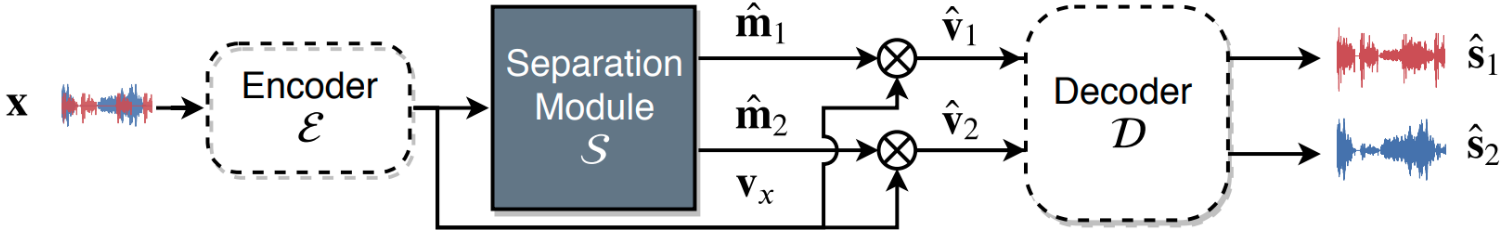
- Time-domain audio source separation
  - Directly optimizing **all parts jointly** using a time-domain loss
    - Scale-Invariant Signal to Distortion Ratio (SI-SDR)



- Mask-based architecture
  - Estimate masks in the latent space
  - Apply masks on the latent representation of the mixture
  - Reconstruct sources from the latent representation using the decoder

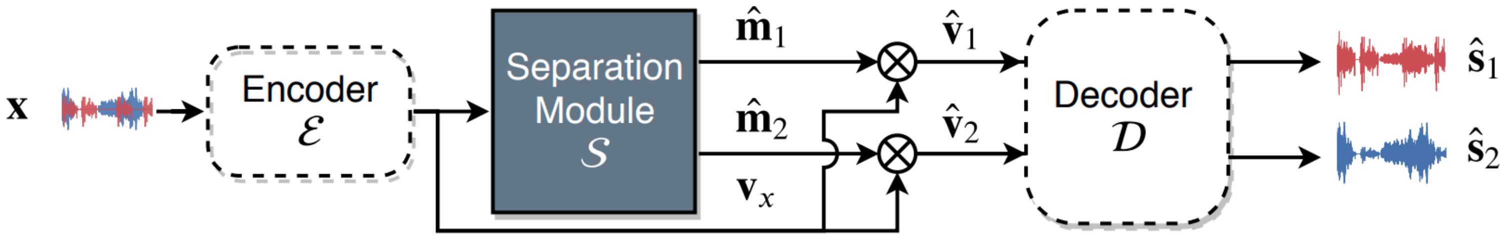
Motivation

- Challenges with the joint training approach:
  - Jointly optimizing encoder/decoder and separator can be suboptimal



Motivation

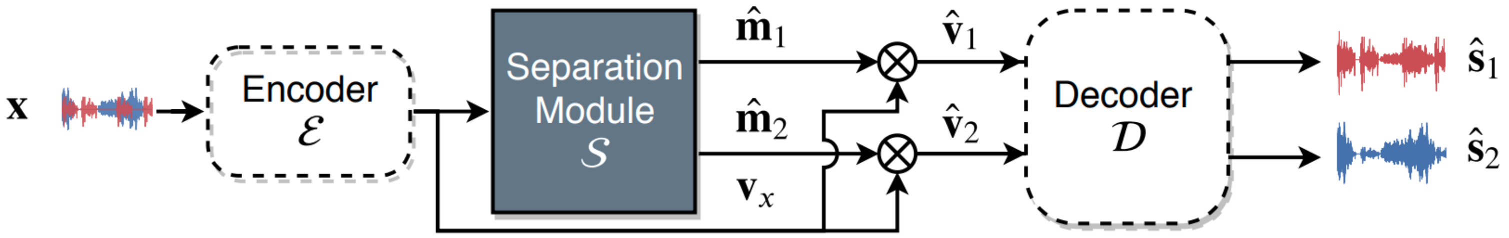
- Challenges with the joint training approach:
  - Jointly optimizing encoder/decoder and separator can be suboptimal



- Putting an **end** to the **end-to-end** optimization?
  - Independently learn a latent representation that facilitates separation

Motivation

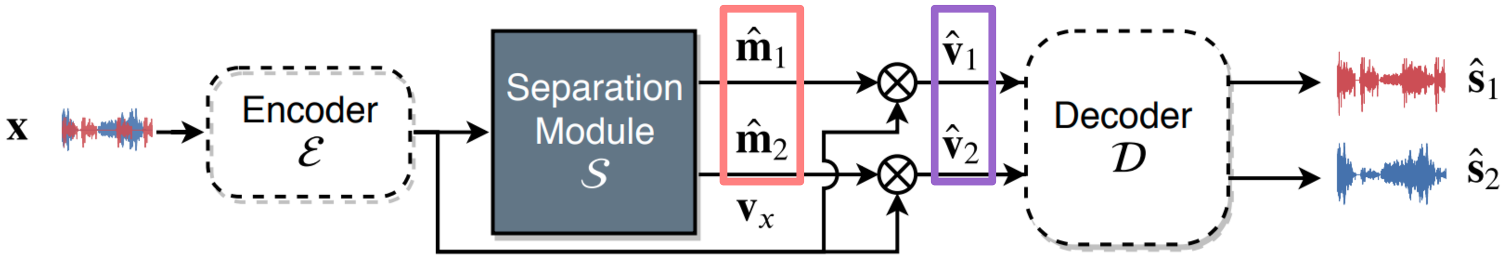
- Challenges with the joint training approach:
  - Jointly optimizing encoder/decoder and separator can be suboptimal



- Putting an **end** to the **end-to-end** optimization?
  - Independently learn a latent representation that facilitates separation
  - Learn to separate using this **pre-trained** transformation

Motivation

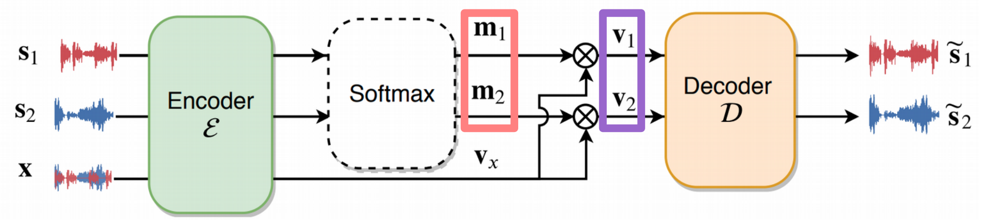
- Challenges with the joint training approach:
  - Jointly optimizing encoder/decoder and separator can be suboptimal



- Putting an **end** to the **end-to-end** optimization?
  - Independently learn a latent representation that facilitates separation
  - Learn to separate using this **pre-trained** transformation
    - Use the **“ideal” targets** of this latent space and train the **separator**
    - Reconstruct the **targets** or the **masks** (just as STFT ideal masks)

Two-step source separation

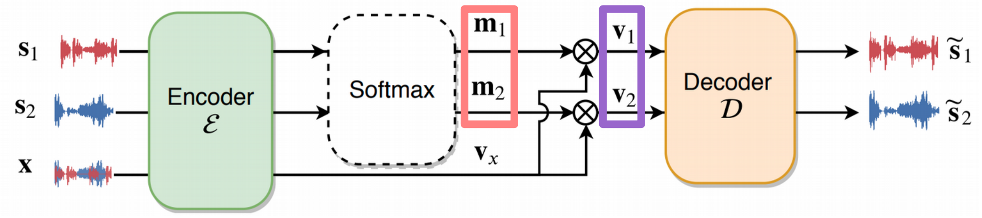
- Step 1: Learning the latent targets
  - Use the clean sources $\mathbf{s}_i, \ \forall i \in \{1, \cdots, N\}$
  - Train **only** the encoder and decoder
  - Get ideal latent targets $\boxed{\mathbf{v}}$
  - Get the corresponding masks $\boxed{\mathbf{m}}$



$$\mathcal{L}_1 = -\text{SI-SDR}(\mathbf{s}^*, \widetilde{\mathbf{s}})$$

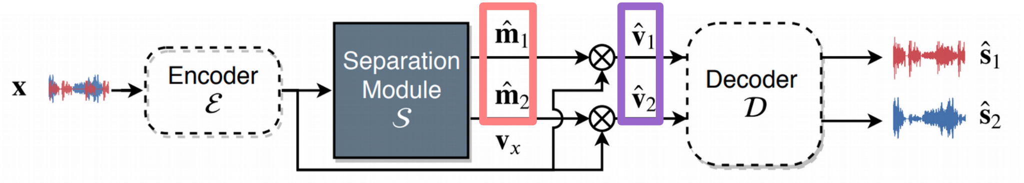
Two-step source separation

- Step 1: Learning the latent targets
  - Use the clean sources $\mathbf{s}_i, \ \forall i \in \{1, \cdots, N\}$
  - Train **only** the encoder and decoder
  - Get ideal latent targets $\boxed{\mathbf{v}}$
  - Get the corresponding masks $\boxed{\mathbf{m}}$



$$\mathcal{L}_1 = -\text{SI-SDR}(\mathbf{s}^*, \widetilde{\mathbf{s}})$$

- Step 2: Training the separation module
  - Use the **pre-trained** encoder and decoder
    - Regress on the ideal **latent targets**
    - Regress on the corresponding **masks**



$$\mathcal{L}_2 = -\text{SI-SDR}(\mathbf{v}^*, \hat{\mathbf{v}})$$

$$\mathcal{L}_2 = -\text{SI-SDR}(\mathbf{m}^*, \hat{\mathbf{m}})$$

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Why to optimize on the latent space?

- Separation objective function (maximization):
  - Time domain: $\boxed{\text{SI-SDR}(\mathbf{s}^*, \widetilde{\mathbf{s}})}$ **Latent space:** $\boxed{\text{SI-SDR}(\mathbf{v}^*, \hat{\mathbf{v}})}$

Why to optimize on the latent space?

- Separation objective function (maximization):
  - Time domain: $\boxed{\text{SI-SDR}(\mathbf{s}^*, \widetilde{\mathbf{s}})}$ **Latent space:** $\boxed{\text{SI-SDR}(\mathbf{v}^*, \hat{\mathbf{v}})}$
- Convolutional decoder (**Latent space** → **Time domain**)
  - Expressed as a matrix multiplication $\hat{\mathbf{s}}_i = \mathcal{D}\left(\hat{\mathbf{v}}_i\right) = \mathbf{P}\hat{\mathbf{v}}_i, \;\; \mathbf{s}_i = \mathbf{P}\mathbf{v}_i, \;\; \forall i \in \{1, \cdots, N\}$

Why to optimize on the latent space?

- Separation objective function (maximization):
  - Time domain: $\boxed{\text{SI-SDR}(\mathbf{s}^*, \widetilde{\mathbf{s}})}$ Latent space: $\boxed{\text{SI-SDR}(\mathbf{v}^*, \hat{\mathbf{v}})}$
- Convolutional decoder (**Latent space** → **Time domain**)
  - Expressed as a matrix multiplication $\hat{\mathbf{s}}_i = \mathcal{D}\left(\hat{\mathbf{v}}_i\right) = \mathbf{P}\hat{\mathbf{v}}_i, \;\; \mathbf{s}_i = \mathbf{P}\mathbf{v}_i, \;\; \forall i \in \{1, \cdots, N\}$
- Relationship between SI-SDR in time-domain and latent space:

Why to optimize on the latent space?

- Separation objective function (maximization):
  - Time domain: $\boxed{\text{SI-SDR}(\mathbf{s}^*, \widetilde{\mathbf{s}})}$ Latent space: $\boxed{\text{SI-SDR}(\mathbf{v}^*, \hat{\mathbf{v}})}$
- Convolutional decoder (Latent space → Time domain)
  - Expressed as a matrix multiplication $\hat{\mathbf{s}}_i = \mathcal{D}(\hat{\mathbf{v}}_i) = \mathbf{P}\hat{\mathbf{v}}_i, \;\; \mathbf{s}_i = \mathbf{P}\mathbf{v}_i, \;\; \forall i \in \{1, \cdots, N\}$
- Relationship between SI-SDR in time-domain and latent space:
  - Equivalent SI-SDR objective **P 1.** *Maximizing SI-SDR$(\boldsymbol{y}, \hat{\boldsymbol{y}})$ w.r.t. $\hat{\boldsymbol{y}}$ is equivalent to maximizing* $\left(\hat{\boldsymbol{y}}^\top \boldsymbol{y}\right)^2$.

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Why to optimize on the latent space?

- Separation objective function (maximization):
  - Time domain: $\boxed{\mathrm{SI\text{-}SDR}(\mathbf{s}^*, \widetilde{\mathbf{s}})}$ Latent space: $\boxed{\mathrm{SI\text{-}SDR}(\mathbf{v}^*, \hat{\mathbf{v}})}$
- Convolutional decoder (Latent space → Time domain)
  - Expressed as a matrix multiplication
- Relationship between SI-SDR in time-domain and latent space:
  - Equivalent SI-SDR objective
  - Lower bound

P 1. *Maximizing SI-SDR*$(\boldsymbol{y}, \hat{\boldsymbol{y}})$ *w.r.t.* $\hat{\boldsymbol{y}}$ *is equivalent to maximizing* $\left(\hat{\boldsymbol{y}}^\top \boldsymbol{y}\right)^2$.

P 2. *Let* $\boldsymbol{y}, \hat{\boldsymbol{y}} \in \mathbb{R}^d$ *and their corresponding projections through* $\boldsymbol{P} \in \mathbb{R}^{n \times d}$ *to* $\mathbb{R}^n$ *defined as* $\boldsymbol{P}\boldsymbol{y}$ *and* $\boldsymbol{P}\hat{\boldsymbol{y}}$*, respectively. If* $\|\boldsymbol{y}\| = \|\hat{\boldsymbol{y}}\| = 1$ *then the absolute value of their inner product on the projection space* $\mathbb{R}^n$ *is bounded above from the absolute value of their inner product in* $\mathbb{R}^d$*, namely:* $\left(\hat{\boldsymbol{y}}^\top \boldsymbol{P}^\top \boldsymbol{P}\boldsymbol{y}\right)^2 \leq g\left(\boldsymbol{P}\right) + \left(\hat{\boldsymbol{y}}^\top \boldsymbol{y}\right)^2$*, where* $g\left(\boldsymbol{P}\right) \geq 0$ *and depends only on the values of* $\boldsymbol{P}$.

Why to optimize on the latent space?

- Separation objective function (maximization):
  - Time domain: $\boxed{\text{SI-SDR}(\mathbf{s}^*, \widetilde{\mathbf{s}})}$ Latent space: $\boxed{\text{SI-SDR}(\mathbf{v}^*, \hat{\mathbf{v}})}$
- Convolutional decoder (Latent space → Time domain)
  - Expressed as a matrix multiplication $\hat{\mathbf{s}}_i = \mathcal{D}\left(\hat{\mathbf{v}}_i\right) = \mathbf{P}\hat{\mathbf{v}}_i, \;\; \mathbf{s}_i = \mathbf{P}\mathbf{v}_i, \;\; \forall i \in \{1, \cdots, N\}$
- Relationship between SI-SDR in time-domain and latent space:
  - Equivalent SI-SDR objective
  - Lower bound
  - Derive relationship

P 1. *Maximizing SI-SDR*$(\boldsymbol{y}, \hat{\boldsymbol{y}})$ *w.r.t.* $\hat{\boldsymbol{y}}$ *is equivalent to maximizing* $\left(\hat{\boldsymbol{y}}^\top \boldsymbol{y}\right)^2$.

P 2. *Let* $\boldsymbol{y}, \hat{\boldsymbol{y}} \in \mathbb{R}^d$ *and their corresponding projections through* $\boldsymbol{P} \in \mathbb{R}^{n \times d}$ *to* $\mathbb{R}^n$ *defined as* $\boldsymbol{P}\boldsymbol{y}$ *and* $\boldsymbol{P}\hat{\boldsymbol{y}}$*, respectively. If* $\|\boldsymbol{y}\| = \|\hat{\boldsymbol{y}}\| = 1$ *then the absolute value of their inner product on the projection space* $\mathbb{R}^n$ *is bounded above from the absolute value of their inner product in* $\mathbb{R}^d$*, namely:* $\left(\hat{\boldsymbol{y}}^\top \boldsymbol{P}^\top \boldsymbol{P}\boldsymbol{y}\right)^2 \leq g\left(\boldsymbol{P}\right) + \left(\hat{\boldsymbol{y}}^\top \boldsymbol{y}\right)^2$, *where* $g\left(\boldsymbol{P}\right) \geq 0$ *and depends only on the values of* $\boldsymbol{P}$.

$$\boxed{\text{SI-SDR}(\mathbf{v}^*, \hat{\mathbf{v}})} \quad \boxed{\left(\hat{\mathbf{v}}_i^\top \mathbf{v}_i\right)^2} = \left[\hat{\mathbf{s}}_i^\top \left(\mathbf{P}^\dagger\right)^\top \mathbf{P}^\dagger \mathbf{s}_i\right]^2 \leq g\left(\mathbf{P}^\dagger\right) + \boxed{\left(\hat{\mathbf{s}}_i^\top \mathbf{s}_i\right)^2} \quad \boxed{\text{SI-SDR}(\mathbf{s}^*, \widetilde{\mathbf{s}})}$$

Overall process

- Training procedure
  - Step1: Train the encoder and decoder **only**
    - Extract "ideal" latent targets $\mathbf{v}$
  - Step2: Train the separation module only
    - Regress over the "ideal" latent targets $\mathcal{L}_2 = -\text{SI-SDR}(\mathbf{v}^*, \hat{\mathbf{v}})$

Overall process

- Training procedure
  - Step1: Train the encoder and decoder **only**
    - Extract "ideal" latent targets
  - Step2: Train the separation module only
    - Regress over the "ideal" latent targets $\mathcal{L}_2 = -\text{SI-SDR}(\mathbf{v}^*, \hat{\mathbf{v}})$

- Separation procedure (Inference)
  - Estimate some latent targets $\hat{\mathbf{v}}$
    - Use the pre-trained decoder to get the time-domain reconstructions $\hat{\mathbf{s}}_i = \mathcal{D}\left(\hat{\mathbf{v}}_i\right) = \mathbf{P}\hat{\mathbf{v}}_i$

Overall process

- Training procedure
  - Step1: Train the encoder and decoder **only**
    - Extract "ideal" latent targets
  - Step2: Train the separation module only
    - Regress over the "ideal" latent targets $\mathcal{L}_2 = -\text{SI-SDR}(\mathbf{v}^*, \hat{\mathbf{v}})$

- Separation procedure (Inference)
  - Estimate some latent targets
    - Use the pre-trained decoder to get the time-domain reconstructions $\hat{\mathbf{s}}_i = \mathcal{D}\left(\hat{\mathbf{v}}_i\right) = \mathbf{P}\hat{\mathbf{v}}_i$

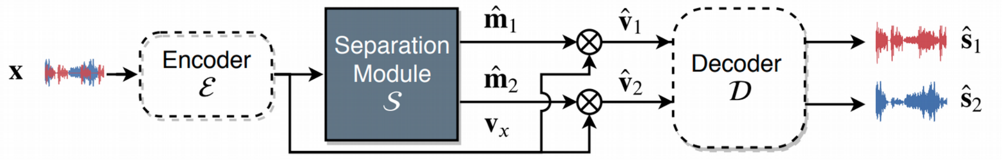
- Notable distinctions
  - Train the encoder-decoder **once** and **re-use** it!
  - **Separation on the latent space** with empirical and theoretical evidence

Sound separation tasks

- Speech separation
  - Mixing utterances from different speakers
  - Wall street journal (WSJ0)
- Non-speech separation
  - Environmental sound classification (ESC50) collection
  - 50 sound classes:
    - animal sounds, natural soundscapes, interior sounds, urban noises, etc.
- Mixed-separation
  - Mix random sources from speech and/or non-speech sounds

Separation Modules

- Time dilated convolutional network (TDCN)
  - Stacked blocks of dilated depth-wise separable convolutions
  - Similar to ConvTasNet [1]
- Residual TDCN (RTDCN)
  - Feature-wise normalization
  - Long-skip residual connections
  - Similar to TDCN++ [2]



[1] Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 8, pp. 1256–1266, 2019.

[2] Ilya Kavalerov, Scott Wisdom, Hakan Erdogan, Brian Patton, Kevin Wilson, Jonathan Le Roux, and John R Hershey, "Universal sound separation," Proc. WASPAA, 2019, pp. 175–179.

Experiments Details

- Data generation & augmentation
  - Generated mixtures: Training: 20,000, Validation: 5,000, Test: 3,000
  - Augment the data
    - Choose at random 2 source audio files
    - Choose at random 4 second source segments
    - Mix at random signal to noise ratios (SNRs)

Experiments Details

- Data generation & augmentation
  - Generated mixtures: Training: 20,000, Validation: 5,000, Test: 3,000
  - Augment the data
    - Choose at random 2 source audio files
    - Choose at random 4 second source segments
    - Mix at random signal to noise ratios (SNRs)
- End-to-end vs Two-step approach
  - Training end-to-end using the time-domain loss
  - Training using the proposed two-step approach
    - Optimizing using the “ideal” latent targets

Experiments Details

- Data generation & augmentation
  - Generated mixtures: Training: 20,000, Validation: 5,000, Test: 3,000
  - Augment the data
    - Choose at random 2 source audio files
    - Choose at random 4 second source segments
    - Mix at random signal to noise ratios (SNRs)
- End-to-end vs Two-step approach
  - Training end-to-end using the time-domain loss
  - Training using the proposed two-step approach
    - Optimizing using the “ideal” latent targets
- Evaluation
  - SI-SDR improvement (SI-SDRi) over the input mixture

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Separation performance SI-SDRi (dB)

Separation Module	Target Domain	Sound Separation Task		
		Speech	Non-speech	Mixed
TDCN	Time	15.4	7.7	11.7
	Latent (ours)	16.1	8.2	12.4
RTDCN	Time	15.6	8.3	12.0
	Latent (ours)	16.2	8.4	12.6

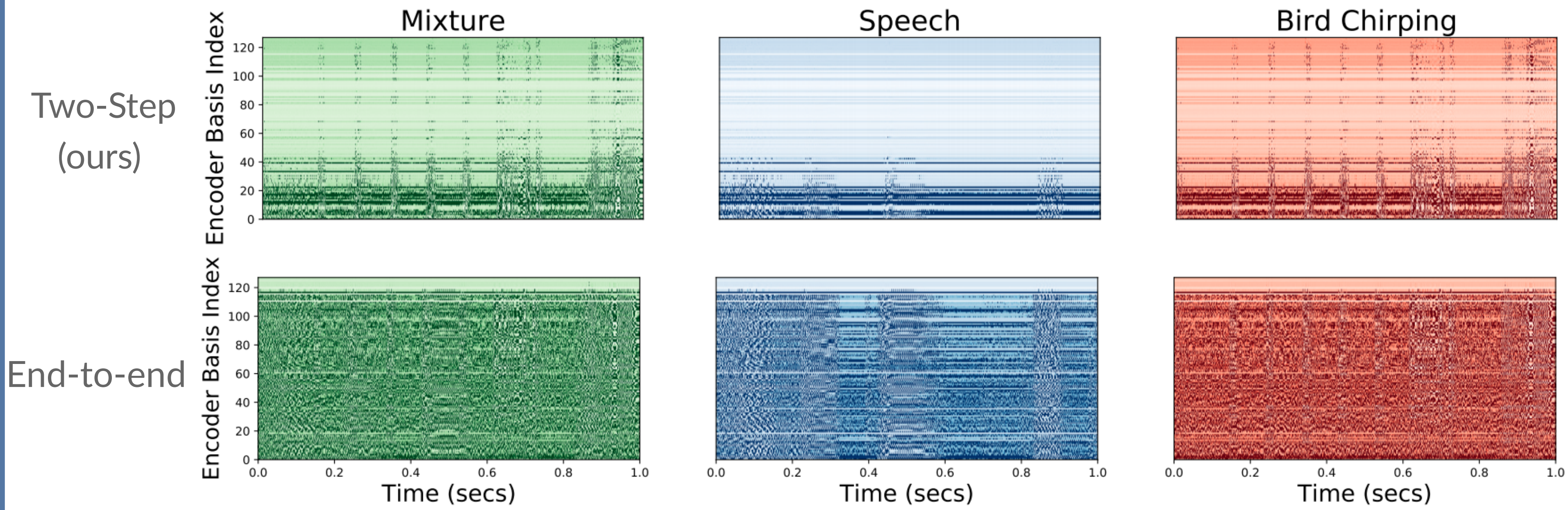
- **Time-domain end-to-end** vs **Two-step** source separation
  - Training on the latent space yields higher performance
    - Across all tasks
    - For both separation modules

Separation Oracles

Oracle Mask Domain	Sound Separation Task		
	Speech	Non-speech	Mixed
STFT	13.0	14.8	14.5
Latent (ours)	34.1	39.2	39.5

- Latent targets vs STFT ideal binary mask
  - **Significantly higher upper bound** for separation performance
  - Across all tasks

Latent targets, a closer look



- A human speaking vs a bird sound
  - We note that the Two-step source separation leads to **sparser** representations for different sounds

Conclusions

- Two-step source separation
  - **Learn a transformation** which facilitates separation
  - Optimize the separator module **using targets on the latent space**

Conclusions

- Two-step source separation
  - **Learn a transformation** which facilitates separation
  - Optimize the separator module **using targets on the latent space**

- Pre-training of the encoder and decoder
  - Consistent sound separation **performance improvement**
    - Across multiple tasks
    - Across separation modules
  - Significantly **higher upper bound of performance** for separation tasks
  - **Sparser latent representations** of sounds of different classes

Conclusions

- Two-step source separation
  - **Learn a transformation** which facilitates separation
  - Optimize the separator module **using targets on the latent space**

- Pre-training of the encoder and decoder
  - Consistent sound separation **performance improvement**
    - Across multiple tasks
    - Across separation modules
  - Significantly **higher upper bound of performance** for separation tasks
  - **Sparser latent representations** of sounds of different classes

- Further ahead
  - More complex encoder/decoder modules (reducing the number of trainable parameters)
  - Transfer learning approaches (fine-tune only the essential parts)

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Waiting to see you all in the Q&A session!

Efthymios Tzinis

etzinis2@illinois.edu
https://etzinis.com