Aiming for Generalization, Efficiency and Interpretability in Machine Learning for Speech and Audio

Cem Subakan

April 21, 2023

Plan

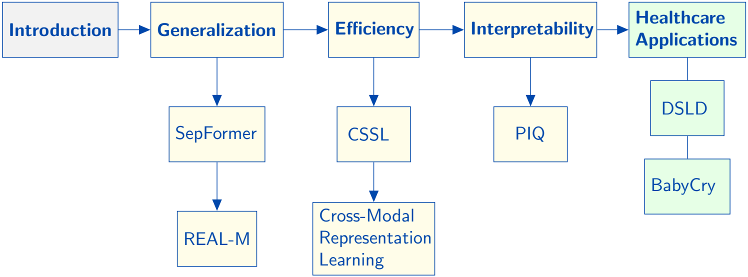
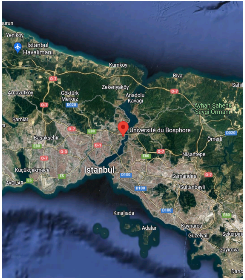

Introduction
Generalization
Efficiency
Interpretability
Healthcare Applications
SepFormer
REAL-M
CSSL
Cross-Modal Representation Learning
PIQ
DSLD
BabyCry

Table of Contents

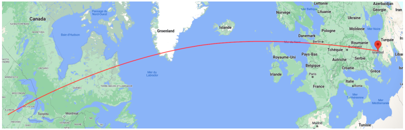
Table of Contents

Short Biography

BSc+MSc, EE - Signal Processing, Bogazici U.



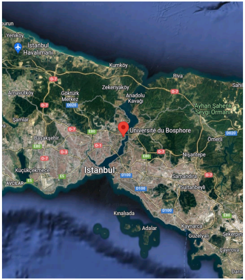
Short Biography

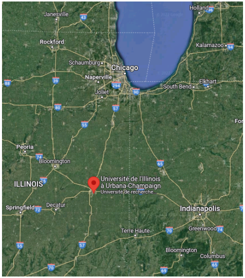


BSc+MSc, EE - Signal Processing, Bogazici U.

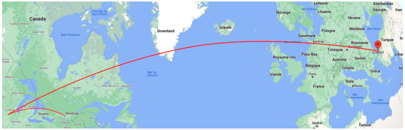
↓

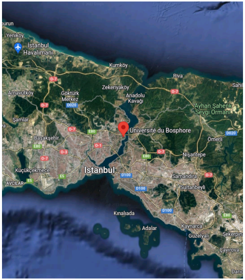
PhD, CS - Machine Learning, UofI UrbanaChampaign





Short Biography





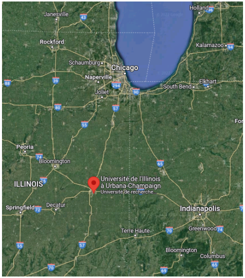
BSc+MSc, EE - Signal Processing, Bogazici U.

↓

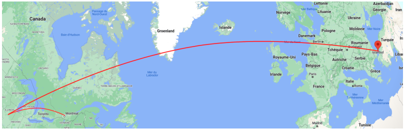
PhD, CS - Machine Learning, UofI UrbanaChampaign

↓

Postdoc, Mila (Quebec AI Institute)



Short Biography

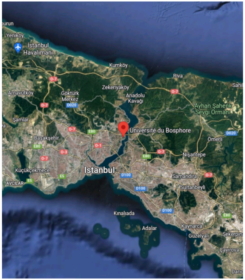


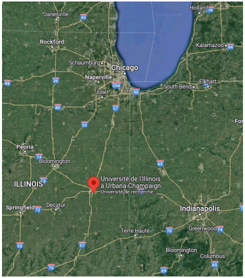
BSc+MSc, EE - Signal Processing, Bogazici U.

PhD, CS - Machine Learning, UofI UrbanaChampaign

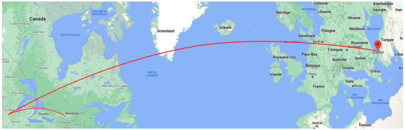
Postdoc, Mila (Quebec AI Institute)

Applied Research Team-Mila





Short Biography



BSc+MSc, EE - Signal Processing, Bogazici U.

↓

PhD, CS - Machine Learning, UofI UrbanaChampaign

↓

Postdoc, Mila (Quebec AI Institute)

↓

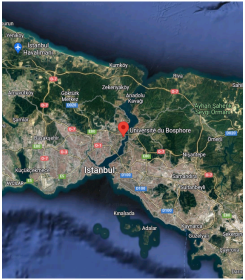
Applied Research Team-Mila

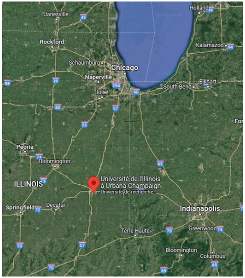
↓

Postdoc, USherbrooke+Mila

↓

Asst. Prof. ULaval+Adjunct Prof. at Concordia U.



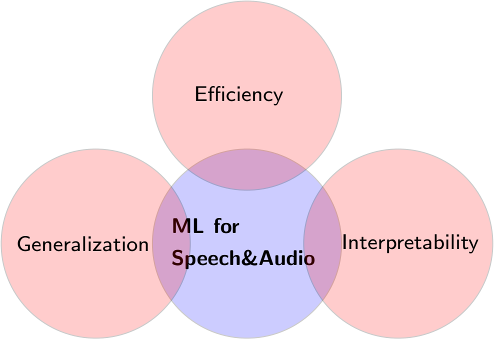


My Research

- I work on developing machine learning methods for speech and audio.

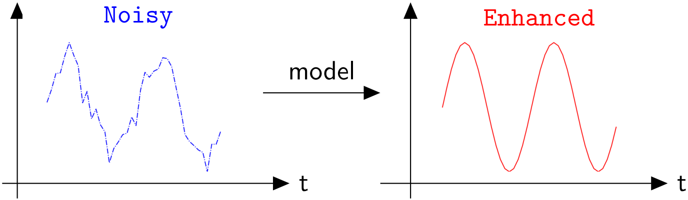
My Research

- I work on developing machine learning methods for speech and audio.
- My current research goals revolve around,
  - Generalization under real-life settings
  - Efficiency (e.g. Continual Learning)
  - Interpretability, Explainability

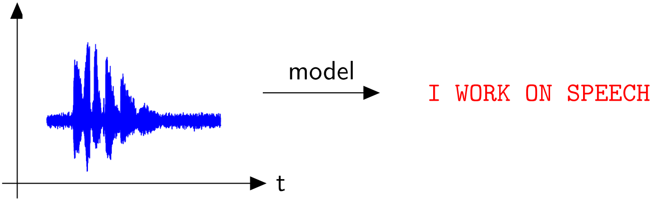


Speech and Audio Modeling

- Speech Enhancement

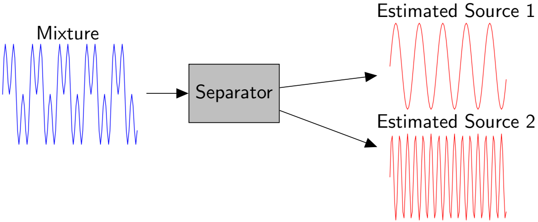


- Speech Recognition

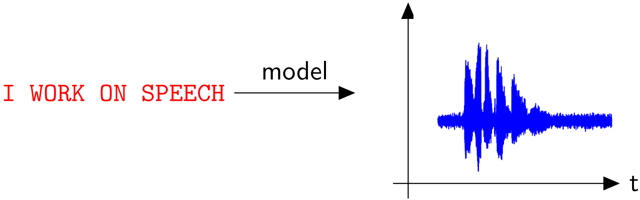


Speech and Audio Modeling

- Speech Separation

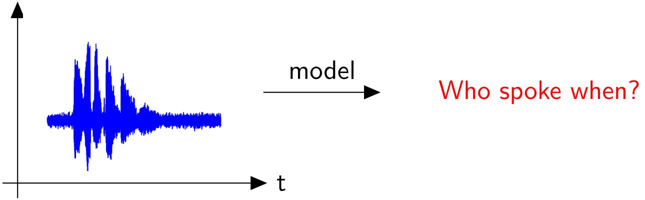


- Text-to-Speech

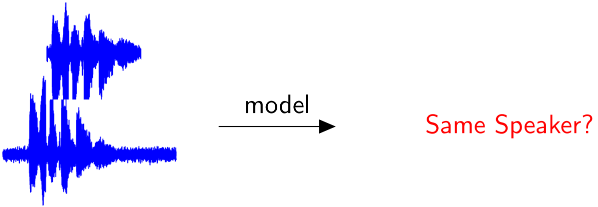


Speech and Audio Modeling

- Speaker Diarization



- Speaker Verification



- Other problems: Spoof Detection, Music Source Separation, Music Transcription, Sound Event Detection/Classification...

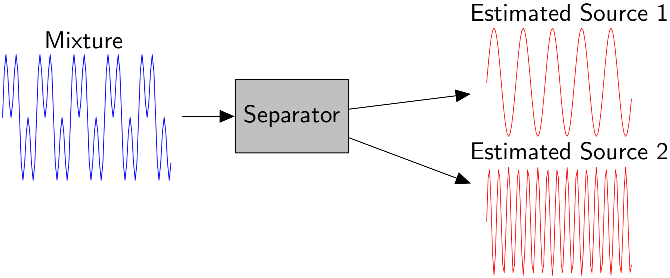
Speech and Audio Modeling

- Field with huge economic value & job opportunities,
  - Speech Recognition (e.g. Siri)
  - Speech Enhancement (e.g. Google meet, Zoom)
  - Text-to-Speech
  - Speaker Verification, Spoof Detection(Banks)
  - Speaker Diarization for Meeting Analysis (Nuance, Microsoft)
  - Source Separation (e.g. Beatles Rock Band, Meeting Analysis)

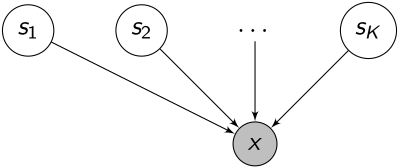
Table of Contents

Table of Contents

Source Separation



Source Separation

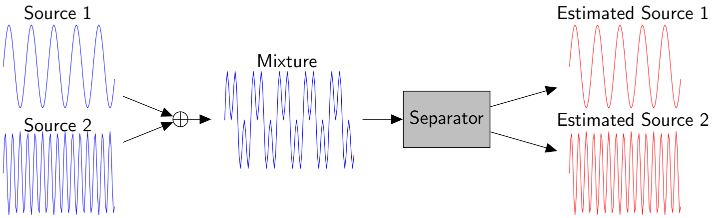


- The observation x is dependent on latent factors $s_1, s_2, \dots, s_K$.
  - Technical definition:

$$s_1 \sim p(s_1) \;\dots\; s_K \sim p(s_K)$$
$$x \sim p(x|s_1, \dots, s_K)$$

  - **Goal:** Obtain $p(s_1|x)$, $p(s_2|x), \dots, p(s_K|x)$

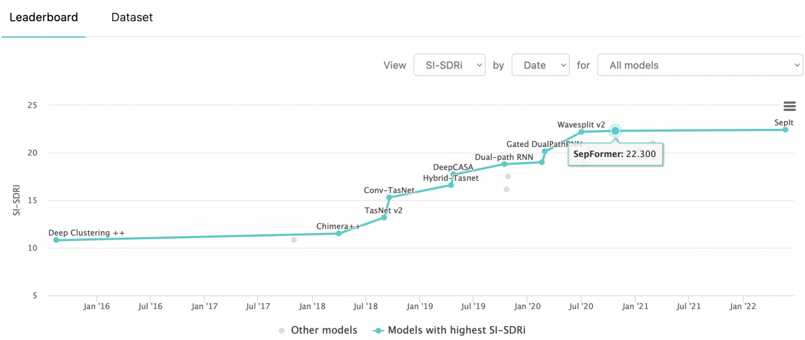
Single-Microphone Source Separation Problem



- **Goal:** To recover the original sources from the observed mixture
- **Applications:** Music production, hearing devices, meeting analysis, editing software, and more...
- **Some of my contributions**
  - Hierarchical tensor factorizations
  - Globally optimal unsupervised source separation with FHMM.
  - Neural network analogs to matrix factorization (best paper award)
  - GANs in source separation
  - **SepFormer**, a self-attention based source separation architecture and obtain state-of-the-art results on multiple datasets.
  - **REAL-M** dataset and evaluation framework

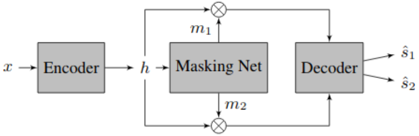
Table of Contents

WSJ0-2Mix Leaderboard

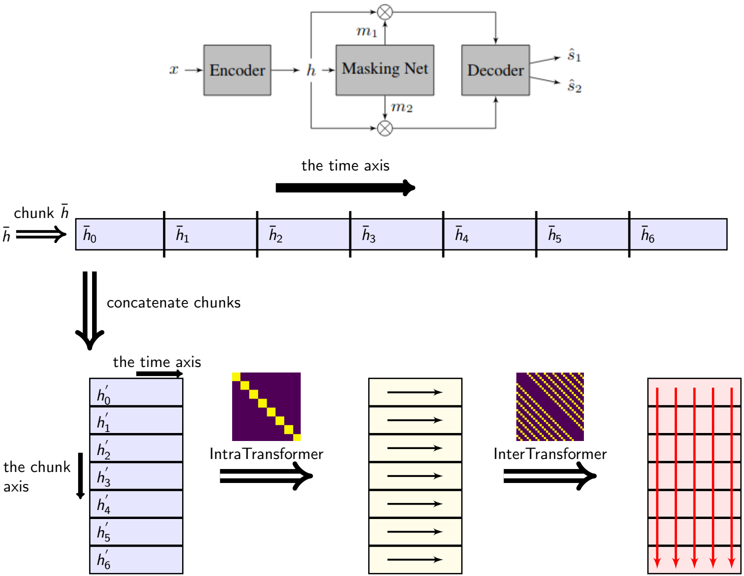


- Taken from `https://paperswithcode.com/sota/speech-separation-on-wsj0-2mix` on October 2022. SepFormer stayed state of the art on WSJ0-2Mix from October 2020-September 2022.
- ICASSP2021 + IEEE TASL. Currently has ~~125~~ >200 citations according to google scholar. ~1000 Monthly downloads.

SepFormer Architecture



SepFormer Architecture



Best Results on Mixtures of 2 speakers (WSJ0-2Mix)

Model	SI-SNRi	SDRi	# Param	Stride
Tasnet	10.8	11.1	n.a	20
SignPredictionNet	15.3	15.6	55.2M	8
ConvTasnet	15.3	15.6	5.1M	10
Two-Step CTN	16.1	n.a.	8.6M	10
DeepCASA	17.7	18.0	12.8M	1
FurcaNeXt	n.a.	18.4	51.4M	n.a.
DualPathRNN	18.8	19.0	2.6M	1
sudo rm -rf	18.9	n.a.	2.6M	10
VSUNOS	20.1	20.4	7.5M	2
DPTNet	20.2	20.6	2.6M	1
Wavesplit	22.2	22.3	29M	1
SepFormer	**22.3**	**22.4**	26M	8

$$SNR \propto 10 \log \left(\frac{\text{Ener. Signal}}{\text{Ener. Noise}} \right)$$

Best Results on Mixtures of 3 Speakers (WSJ0-3Mix)

Model	SI-SNRi	SDRi	# Param
ConvTasnet	12.7	13.1	5.1M
DualPathRNN	14.7	n.a	2.6M
VSUNOS	16.9	n.a	7.5M
Wavesplit	17.8	18.1	29M
Sepformer	**19.5**	**19.7**	26M

Example Results on Test Set:

`Click for Mixture`

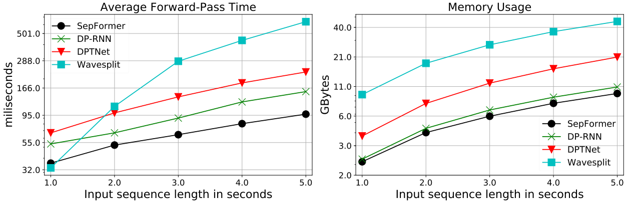
`Click for Estimated Source1`

`Click for Estimated Source2`

`Click for Estimated Source3`

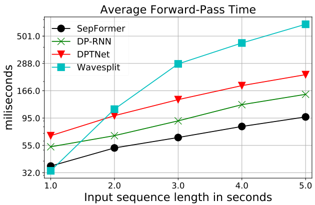
Speed/Memory Comparison with Other Methods

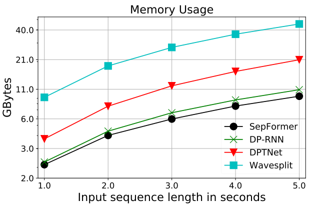
Speed and Memory Comparison on Forward Pass:



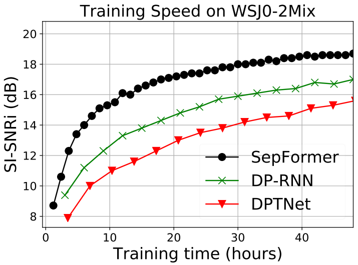
Speed/Memory Comparison with Other Methods

Speed and Memory Comparison on Forward Pass:





Training Curve Comparison:



Environmental Corruption

We try our model with environmental noise / reverberation.

Best results on the WHAM dataset (noise).

Model	SI-SNRi	SDRi
ConvTasnet	12.7	-
Learnable fbank	12.9	-
Wavesplit	16.0	16.5
Sepformer	**16.4**	**16.7**

Best results on the WHAMR (noise + reverb) dataset.

Model	SI-SNRi	SDRi
ConvTasnet	8.3	-
BiLSTM Tasnet	9.2	-
Wavesplit	13.2	12.2
Sepformer	**14.0**	**13.0**

Cross-Dataset Experiment

We test our model trained on WSJ0-2Mix on LibriMix.

Model	**SI-SNRi**	**SDRi**
ConvTasnet	14.7	-
Sepformer trained on WSJ0-2Mix	17.0	17.5
Wavesplit	20.5	20.7
Sepformer	20.2	20.5
Sepformer + FT	**20.6**	**20.8**

We test our model trained on WSJ0-3Mix on LibriMix.

Model	**SI-SNRi**	**SDRi**
ConvTasnet	10.4	-
Sepformer trained on WSJ0-3Mix	15.0	15.6
Wavesplit	17.5	18.0
Sepformer	18.2	18.6
Sepformer + FT	**18.7**	**19.0**

Note: We release our pretrained models, training scripts on SpeechBrain!

Synthetic vs Real Life Mixture

Synthetic: WSJ0-2Mix test set

Click for Mixture

Click for Estimated Source1

Click for Estimated Source2

Real-life: One mic, two people speaking, reverberant environment

Click for Mixture

Click for Estimated Source1

Click for Estimated Source2

Click for Mixture

Click Estimated Source 1

Click Estimated Source 2

Table of Contents

Closing the reality gap

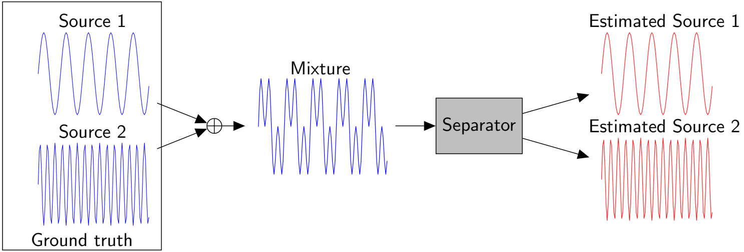
- We need evaluation sets that represents the challanges of real-life so that researchers can more meaningfully benchmark their performance.
- We can then design data augmentations, and models to improve performance on real-life data.

Closing the reality gap

- We need evaluation sets that represents the challanges of real-life so that researchers can more meaningfully benchmark their performance.
- We can then design data augmentations, and models to improve performance on real-life data.
- An important hurdle: Ground truth data.

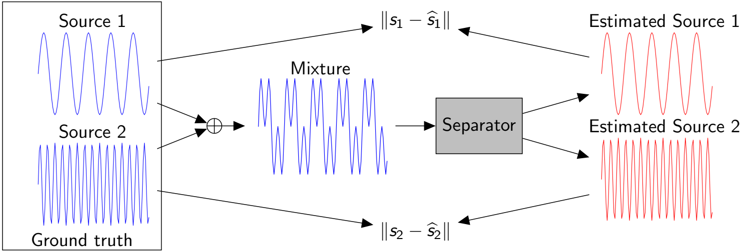
Lack of ground truth in real-life separation

- Synthetic source separation datasets



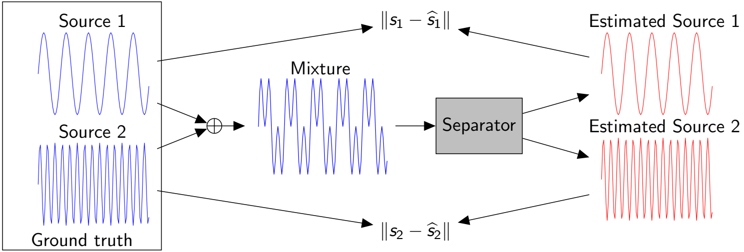
Lack of ground truth in real-life separation

- Synthetic source separation datasets

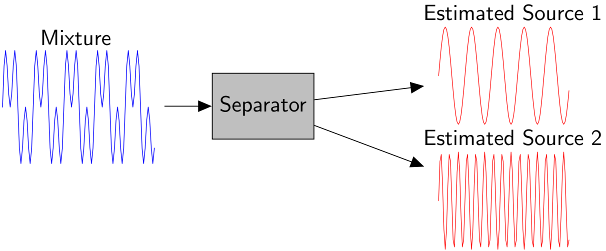


Lack of ground truth in real-life separation

- Synthetic source separation datasets

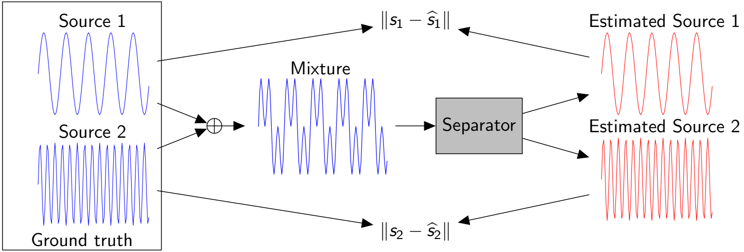


- Real-life source separation

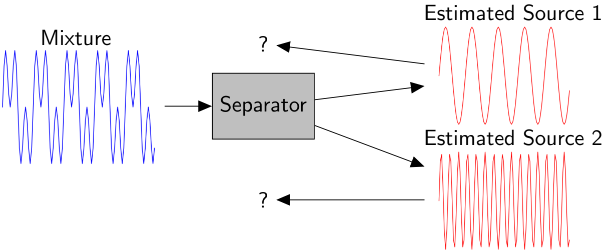


Lack of ground truth in real-life separation

- Synthetic source separation datasets



- Real-life source separation



Tackling the lack of ground truth

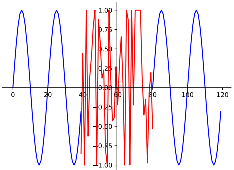
- The lack of ground truth prevents evaluating estimation quality on real-life data.

Tackling the lack of ground truth

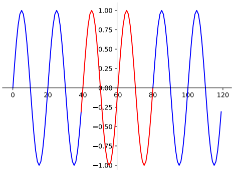
- The lack of ground truth prevents evaluating estimation quality on real-life data.
- We can however estimate the performance!
- We can train a model to estimate the performance.

Tackling the lack of ground truth

- The lack of ground truth prevents evaluating estimation quality on real-life data.
- We can however estimate the performance!
- We can train a model to estimate the performance.



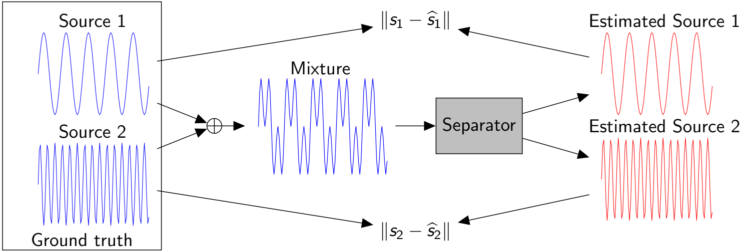
BAD



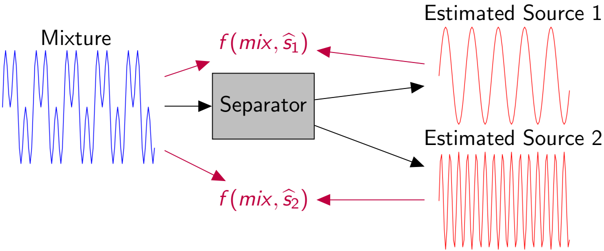
GOOD

Tackling the lack of ground truth

- Synthetic source separation datasets



- Real-life source separation



REAL-M: Towards Speech Separation on Real Mixtures

- **Goal: Systematic Evaluation of Speech Separation Models on Real-Life Speech Mixtures**.
- **Contributions:**
  - We propose a dataset for **real-life speech separation**. The dataset is **crowdsourced**, hence **scalable and diverse** in acoustic conditions, recording hardware, speakers.
  - We show that **blind SI-SNR estimation** is a feasible way to evaluate real-life speech separation.
  - Therefore, this opens up a scalable methodology for large-scale real-life source separation evaluation.
  - The 5th most viewed poster in ICASSP 2022! (out of 1900 posters)

Table of Contents

Table of Contents

Continual Learning

Batch Learning

$$\mathcal{D} \rightarrow f_\theta(.)$$

Continual Learning

$$\begin{gathered}\mathcal{D}_1 \rightarrow f_\theta(.) \\ \mathcal{D}_2 \rightarrow f_\theta(.) \\ \vdots \\ \mathcal{D}_T \rightarrow f_\theta(.)\end{gathered}$$

Continual Learning

Batch Learning

$$\mathcal{D} \to f_\theta(.)$$

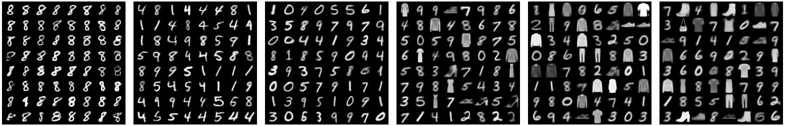
Continual Learning

$$\mathcal{D}_1 \to f_\theta(.)$$

$$\mathcal{D}_2 \to f_\theta(.)$$

$$\vdots$$

$$\mathcal{D}_T \to f_\theta(.)$$



Continual Supervised Learning vs Continual Self-Supervised Learning

- Continual Supervised Learning (CSUP)

$$\begin{array}{cccc} \mathcal{D}_1 & \mathcal{D}_2 & \cdots & \mathcal{D}_T \\ \downarrow & \downarrow & \downarrow & \downarrow \\ g_\theta & g_\theta & \cdots & g_\theta \end{array}$$

Continual Supervised Learning vs Continual Self-Supervised Learning

- Continual Supervised Learning (CSUP)

$$\begin{tikzcd}
\mathcal{D}_1 \arrow[d] & \mathcal{D}_2 \arrow[d] & \cdots \arrow[d] & \mathcal{D}_T \arrow[d] \\
g_\theta & g_\theta & \cdots & g_\theta
\end{tikzcd}$$

- Continual Representation Learning

$$\begin{tikzcd}
\mathcal{D}_1 \arrow[d] & \mathcal{D}_2 \arrow[d] & \cdots \arrow[d] & \mathcal{D}_T \arrow[d] \\
f_\theta \arrow[d] & f_\theta \arrow[d] & \cdots \arrow[d] & f_\theta \arrow[d] \\
h_\theta & h_\theta & \cdots & h_\theta
\end{tikzcd}$$

Continual Supervised Learning vs Continual Self-Supervised Learning

- Continual Supervised Learning (CSUP)

$$\begin{tikzcd}
\mathcal{D}_1 \arrow[d] & \mathcal{D}_2 \arrow[d] & \cdots \arrow[d] & \mathcal{D}_T \arrow[d] \\
g_\theta & g_\theta & \cdots & g_\theta
\end{tikzcd}$$

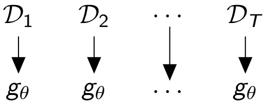
- Continual Representation Learning

$$\begin{tikzcd}
\mathcal{D}_1 \arrow[d] \arrow[ddr, red] \arrow[ddrr, red] \arrow[ddrrr, red] & \mathcal{D}_2 \arrow[d] \arrow[ddr, red] \arrow[ddrr, red] & \cdots \arrow[d] \arrow[ddr, red] & \mathcal{D}_T \arrow[d] \\
f_\theta \arrow[d] & f_\theta \arrow[d] & \cdots \arrow[d] & f_\theta \arrow[d] \\
h_\theta & h_\theta & \cdots & h_\theta
\end{tikzcd}$$

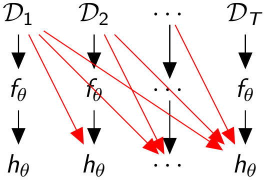
- **Potential Advantages of CRL:** Flexible, Empirically less prone to forgetting, Computational Savings

Continual Supervised Learning vs Continual Self-Supervised Learning

- Continual Supervised Learning (CSUP)



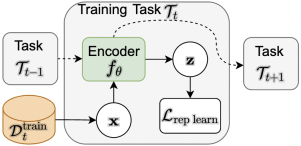
- Continual Representation Learning



- **Potential Advantages of CRL:** Flexible, Empirically less prone to forgetting, Computational Savings
- Handles the realistic case where only a subset of labels is available.

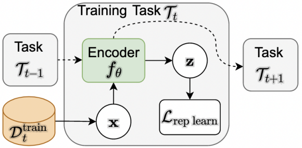
Continual Self-Supervised Learning

- Published in SPL 2023, Will be presented in ICASSP 2023.
- Training the encoder

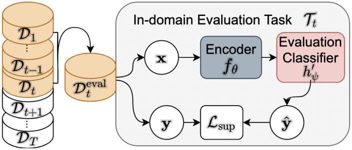


Continual Self-Supervised Learning

- Published in SPL 2023, Will be presented in ICASSP 2023.
- Training the encoder

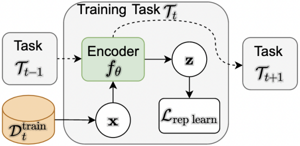


- Training the output head

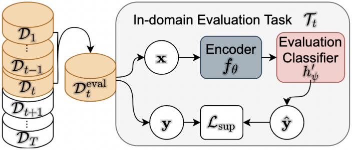


Continual Self-Supervised Learning

- Published in SPL 2023, Will be presented in ICASSP 2023.
- Training the encoder



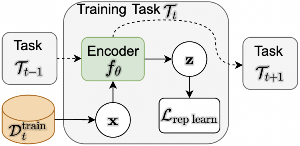
- Training the output head



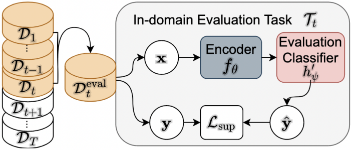
- **Linear Evaluation Protocol (LEP):** A linear layer is trained on top of the pretrained encoder using data from current and previous tasks $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_t$.

Continual Self-Supervised Learning

- Published in SPL 2023, Will be presented in ICASSP 2023.
- Training the encoder



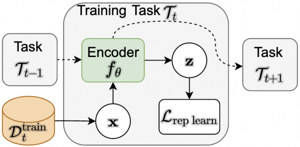
- Training the output head



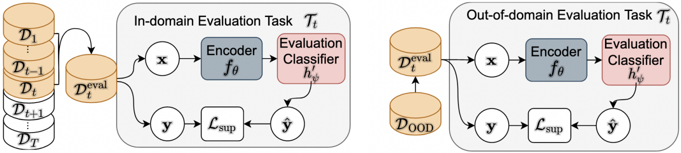
- **Linear Evaluation Protocol (LEP):** A linear layer is trained on top of the pretrained encoder using data from current and previous tasks $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_t$.
- **Subset Linear Evaluation Protocol (SLEP):** Only a percentage of data is labeled. This simulates the real-life use cases.

Continual Self-Supervised Learning

- Published in SPL 2023, Will be presented in ICASSP 2023.
- Training the encoder



- Training the output head



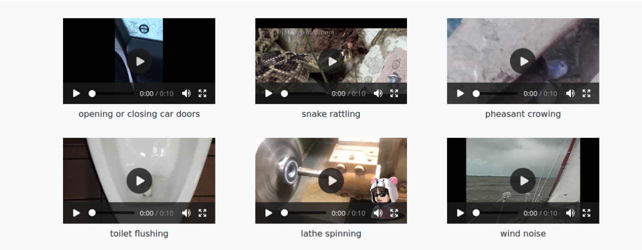
- **Linear Evaluation Protocol (LEP):** A linear layer is trained on top of the pretrained encoder using data from current and previous tasks $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_t$.
- **Subset Linear Evaluation Protocol (SLEP):** Only a percentage of data is labeled. This simulates the real-life use cases.

Datasets

- In Domain Evaluation
  - UrbanSound8k, 10 possible urban sound classes (car horn, street music, air conditioner, ...), Each class has 1000, 4second long recordings. In total 8.75 hours. (smallish)
  - DCASE, TAU19, 10 possible urban sound classes, 40 hours of audio.

Datasets

- In Domain Evaluation
  - UrbanSound8k, 10 possible urban sound classes (car horn, street music, air conditioner, ...), Each class has 1000, 4second long recordings. In total 8.75 hours. (smallish)
  - DCASE, TAU19, 10 possible urban sound classes, 40 hours of audio.
- Out-of-domain Evaluation
  - VGG Sound, 560 hours of audio/visual data scraped from youtube. 300 sound classes such as instruments, horns, city sounds. Labels are not reliable, so we use it for unsupervised learning.



Empirical Findings on In Domain Data

- CSSL is more robust to forgetting than Supervised Representation Learning (CSUP).

Method	USound8K		DCASE	
	A (↑)	F (↓)	A (↑)	F (↓)
CSUP	65.6	19.6	48.2	27.6
CSSL-SimCLR	**70.3**	15.3	**59.7**	**17.8**
CSSL-Barlow Twins	68.5	**14.1**	55.9	19.0
CSSL-MoCo	68.4	15.6	49.5	20.2

Empirical Findings on In Domain Data

- CSSL is more robust to forgetting than Supervised Representation Learning (CSUP).
- CSSL (even without an explicit mechanism against forgetting) is robust against forgetting (comparable perf. with distillation).

Method	USound8K		DCASE	
	A (↑)	F (↓)	A (↑)	F (↓)
No distillation				
CSUP	65.6	19.6	48.2	27.6
SimCLR	**70.3**	15.3	**59.7**	**17.8**
Barlow Twins	68.5	**14.1**	55.9	19.0
MoCo	68.4	15.6	49.5	20.2
With distillation				
CSUP + $\mathcal{L}_{\text{MSE}}$	58.6	27.0	49.1	26.1
CSUP + $\mathcal{L}_{\text{sim}}$	70.6	**13.8**	56.2	19.7
CSUP + $\mathcal{L}_{\text{KLD}}$	69.8	15.9	55.7	19.4
SimCLR + $\mathcal{L}_{\text{MSE}}$	**70.9**	14.6	56.2	19.6
SimCLR + $\mathcal{L}_{\text{sim}}$	70.6	14.0	**60.0**	**17.6**

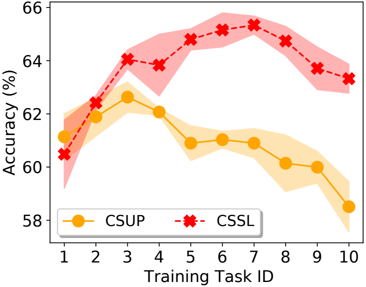
Empirical Findings on In Domain Data

- CSSL is more robust to forgetting than Supervised Representation Learning (CSUP).
- CSSL (even without an explicit mechanism against forgetting) is robust against forgetting (comparable perf. with distillation).
- **Practical use case:** CSSL is robust against forgetting in the scarce label case as well.

Method	UrbanSound8K				DCASE TAU19			
	LEP		SLEP		LEP		SLEP	
	A (↑)	F (↓)	A (↑)	F (↓)	A (↑)	F (↓)	A (↑)	F (↓)
No distillation								
CSUP	65.6	19.6	48.4	33.1	48.2	27.6	32.5	38.4
SimCLR	**70.3**	15.3	**50.3**	26.6	**59.7**	**17.8**	**42.1**	27.9
Barlow Twins	68.5	**14.1**	49.7	**20.5**	55.9	19.0	41.0	**23.4**
MoCo	68.4	15.6	**50.3**	25.8	49.5	20.2	34.8	25.8
With distillation								
CSUP + $\mathcal{L}_{\text{MSE}}$	58.6	27.0	43.8	39.2	49.1	26.1	35.1	35.8
CSUP + $\mathcal{L}_{\text{sim}}$	70.6	**13.8**	54.9	27.1	56.2	19.7	42.1	32.4
CSUP + $\mathcal{L}_{\text{KLD}}$	69.8	15.9	**55.4**	27.4	55.7	19.4	42.6	30.3
SimCLR + $\mathcal{L}_{\text{MSE}}$	**70.9**	14.6	50.6	25.1	56.2	19.6	42.0	26.5
SimCLR + $\mathcal{L}_{\text{sim}}$	70.6	14.0	51.1	**25.0**	**60.0**	**17.6**	**42.8**	**25.9**

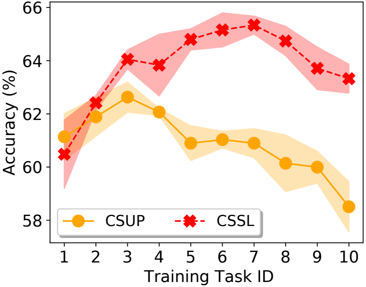
Out-of-domain evaluation

- We train the encoder on a stream of unlabeled data. We test on a fixed, out-of-domain test set.



Out-of-domain evaluation

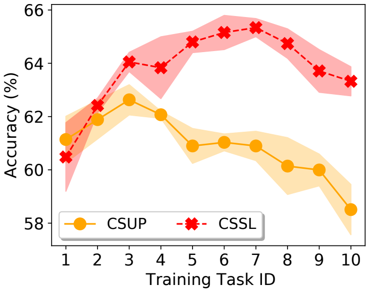
- We train the encoder on a stream of unlabeled data. We test on a fixed, out-of-domain test set.



- We observe that CSSL results in better OOD generalization than continual supervised representation learning!

Out-of-domain evaluation

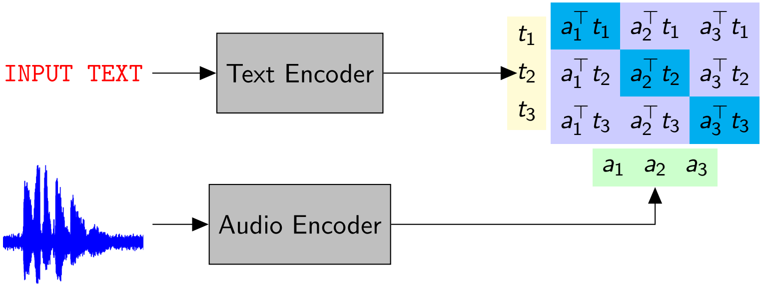
- We train the encoder on a stream of unlabeled data. We test on a fixed, out-of-domain test set.



- We observe that CSSL results in better OOD generalization than continual supervised representation learning!
- **Current Objectives:**
  - Long term goal is to have domain generalization under the continual learning setting.
  - Accelerating learning. (Similar to how humans learn)

Table of Contents

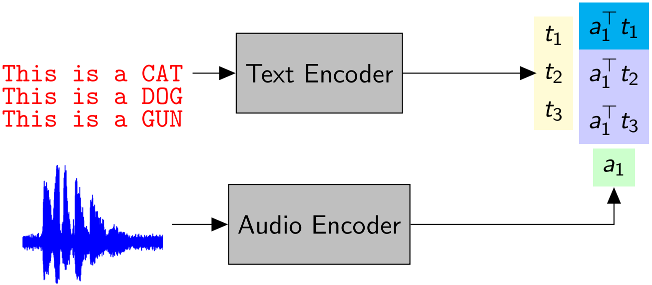
Cross-Modal Representation Learning



- CLAP
  - We maximize $a_i^\top t_j$ for $i = j$, and minimize for $i \neq j$.
  - This enables text-based audio retrieval, zero-shot classification.

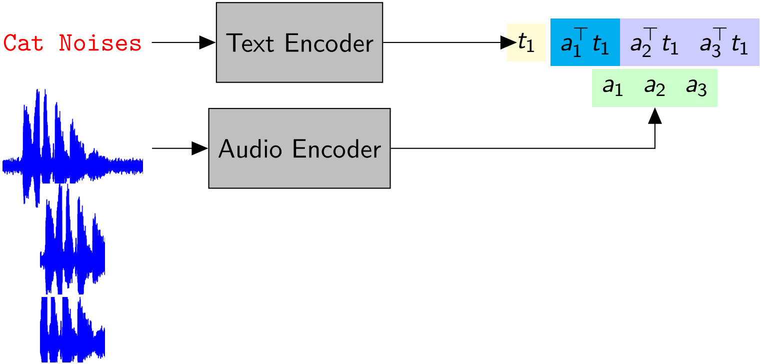
Cross-Modal Representation Learning

- Zero-shot evaluation

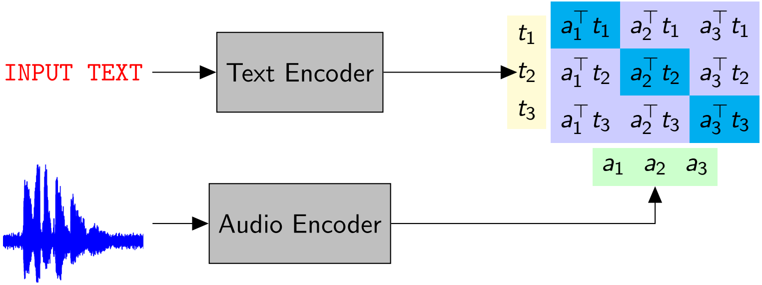


Cross-Modal Representation Learning

- Audio Retreival

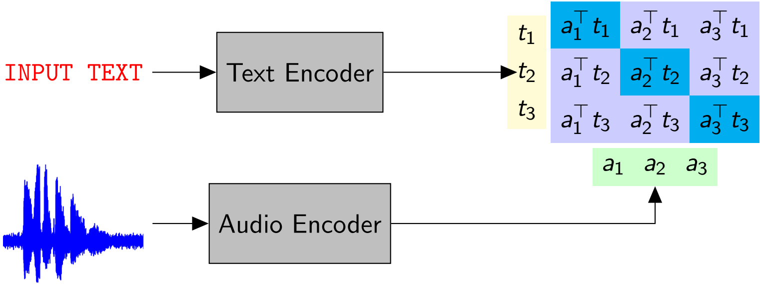


Cross-Modal Representation Learning



- CLAP
  - Training this model requires large number of paired data.

Cross-Modal Representation Learning



- CLAP
  - Training this model requires large number of paired data.
  - **Ongoing work:** We are working on a method where we improve the model performance using unpaired text and audio.

Table of Contents

Table of Contents

Neural Network Explanation

- *Why does this particular input lead to that particular output?*

Neural Network Explanation

- *Why does this particular input lead to that particular output?*

Neural Network Explanation

- *Why does this particular input lead to that particular output?*

Recording, Classified as DOG

Neural Network Explanation

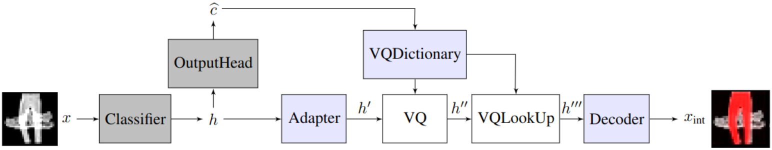
- *Why does this particular input lead to that particular output?*

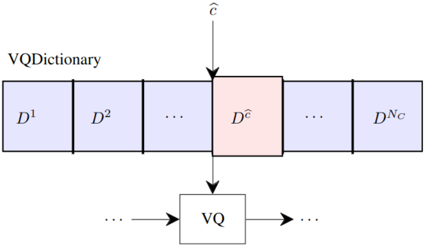
Recording, Classified as DOG
Interpretation

Posthoc Interpretation via Quantization

- We have developed a method that learns "high-level" concepts for each class in form of latent VQ dictionary, and then reconstructs the input using this VQ dictionary conditioned on the class information.

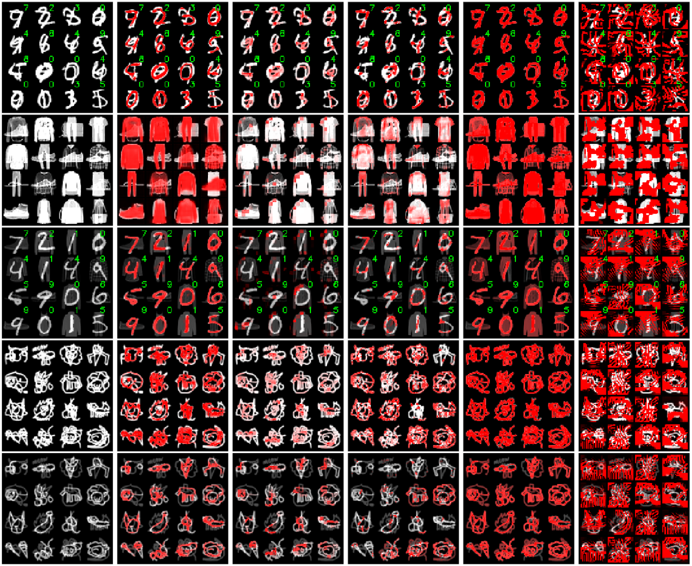
Posthoc Interpretation via Quantization





Above shows the inference time. In training, we only use images with single classes. NOT mixtures.

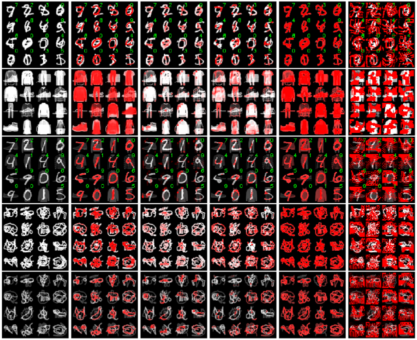
Qualitative Results on Images



Left-to-Right: Input, PIQ (ours), VIBI, L2I, LIME, FLINT

Mean-Opinion-Scores on Images

Dataset	Method	MOS (↑)
MNIST B1 (Case 1)	PIQ (ours)	**4.04 ± 0.48**
	VIBI	1.77 ± 0.68
	L2I	2.4 ± 0.66
	FLINT	1 ± 0
	LIME	2 ± 1.34
MNIST B2 (Case 1)	PIQ (ours)	**3.95 ± 0.72**
	VIBI	1.86 ± 0.71
	L2I	1.86 ± 0.56
	FLINT	1.04 ± 0.21
	LIME	2.13 ± 1.21
FMNIST Mix (Case 2)	PIQ (ours)	**4.87 ± 0.50**
	VIBI	1.37 ± 0.50
	L2I	3.18± 0.91
	FLINT	1.12± 0.50
	LIME	1.37 ± 0.89
MNIST+FMN (Case 3)	PIQ (ours)	**4.78 ± 0.43**
	VIBI	1.14 ± 0.47
	L2I	2.18 ± 0.96
	FLINT	1.09 ± 0.47
	LIME	3.23 ± 0.72
QuickDraw1 (Case4-i)	PIQ (ours)	**2.6 ± 1.67**
	LIME	2.35 ± 1.46
QuickDraw2 (Case4-ii)	PIQ (ours)	**3.55 ± 1.0**
	LIME	3 ± 1,38



Quantitative Results on Images

Dataset	MNIST			FMNIST		
Metric	Fidelity-In (↑)	Faithfulness (↑)	FID (↓)	Fidelity-In (↑)	Faithfulness (↑)	FID (↓)
PIQ (ours)	**98.03 ± 0.05**	**0.588 ± 0.00021**	**0.029 ± 0.0004**	**81.3 ± 0.2**	**0.773 ± 0.004**	**0.030 ± 0.0004**
VIBI	73.90 ± 16.08	0.369 ± 0.002	0.710 ± 0.962	42.4 ± 17.8	0.578 ± 0.073	0.395 ± 0.104
L2I	96.56 ± 2.66	0.453 ± 0.002	0.160 ± 0.010	68.3 ± 1.5	0.343 ± 0.011	0.188 ± 0.011
FLINT	10.9	0.361	0.677	15.37	-0.097	0.482

Dataset	Quickdraw		
Metric	Fidelity-In (↑)	Faithfulness (↑)	FID (↓)
PIQ (ours)	**60.89 ± 0.60**	**0.675 ± 0.005**	**0.034 ± 0.0001**
VIBI	26.36 ± 3.01	0.341 ± 0.031	0.388 ± 0.032
L2I	25.97 ± 0.82	0.340 ± 0.031	0.397 ± 0.020
FLINT	15.62	-0.057	0.672

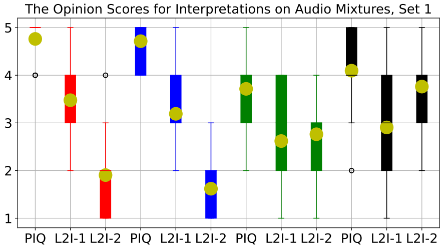
- Input Fidelity

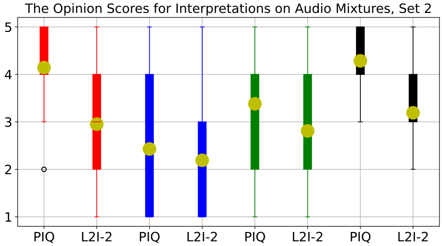
$$\text{FID-I} = \frac{1}{N}\sum_{n=1}^{N}\left[\arg\max_c f_c(x_n) = \arg\max_c f_c(x_{\text{int},n})\right],$$

- Faithfulness

$$\text{Faithfulness} = f_{\hat{c}}(x) - f_{\hat{c}}(x - x_{\text{int}}),$$

Mean-Opinion Scores on Audio





Click for More Example Results

Conclusions on Interpretability

- I am trying to build a research axis on Interpretability / Explanations. One incoming PhD student.
- Several interesting applications on audio domain.
- Working on generalizing our approach to more complex audio / images.

Table of Contents

Table of Contents

ML for Speech and Language Disorders

- We have just submitted a grant application from CIHR-SickKids foundation with Prof. Selcuk Guven from UdeM Audiologie department.
- The goal is to use ML to Diagnose and Understand Speech and Language Disorders.
  - **Example 1**
  - **Example 2**

```
Model1: CHUCK SEEMS THIRSTY AFTER THE RACE TRIFLING
        THIRSTY OVER THE LADS
Model2: CHUCK SEEMS THIRSTY AFTER THE RACE CHRISHENG
        THIRSTY OVER THE WAYS
Model3: CHUCK SEENS BURSTY AFTER THE RACE CHICKING
        THIRSDAY OVER THE WAGE
```

ML for Speech and Language Disorders

- We have just submitted a grant application from CIHR-SickKids foundation with Prof. Selcuk Guven from UdeM Audiologie department.
- The goal is to use ML to Diagnose and Understand Speech and Language Disorders.
  - **Example 1**
  - **Example 2**

```
Model1:  CHUCK SEEMS THIRSTY AFTER THE RACE TRIFLING
                THIRSTY OVER THE LADS
Model2:  CHUCK SEEMS THIRSTY AFTER THE RACE CHRISHENG
                THIRSTY OVER THE WAYS
Model3:  CHUCK SEENS BURSTY AFTER THE RACE CHICKING
                THIRSDAY OVER THE WAGE
```

- Subgoals include,
  - Data collection under clinical setting
  - ML for diagnosis models + Active learning for noisy label re-labeling
  - Robust phoneme based ASR to interpret the diagnosis.
  - Applying and developing neural network interpretation methods.
- This project is included in an accepted Compute Canada RRG application.

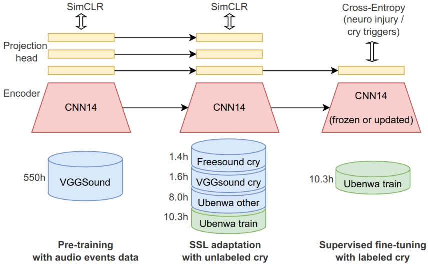
Table of Contents

ML for Infant Cry Analysis

- Collaboration with UbenwaAI, a Mila based startup.
- The goal is to develop machine learning methods for Infant Cry Analysis.
- Recent accepted paper:
  Self-Supervised Learning for Cry Analysis,
  ICASSP 2023 Workshop on Self-Supervision in Audio, Speech and Beyond (gets into conference proceedings).

SELF-SUPERVISED LEARNING FOR INFANT CRY ANALYSIS

Arsenii Gorin[⋆], Cem Subakan[♮♯♭], Sajjad Abdoli[⋆], Junhao Wang[⋆], Samantha Latremouille[⋆], Charles Onu[⋆♭]



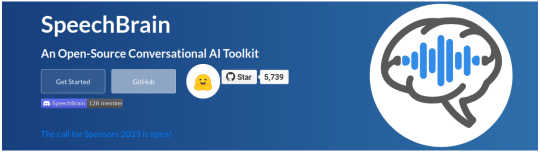
Baby Identification Challenge: CryCeleb

The CryCeleb2023 Challenge!

Ubenwa Collaboration

We will have one PhD position in collaboration with Ubenwa on **Interpretability, Continual Learning, Self-Supervised Learning**. Contact me if you are interested!

SpeechBrain



On August 28th, We will have the first annual SpeechBrain summit (with Interspeech endorsement). There will be talks from industry, academia, and a panel discussion with creators of torchaudio, Kaldi, Librosa, ESPNet, NeMO on open source software for speech.

Conclusions

- I tried to summarize my recent work and goals concerning generalization, efficiency and interpretability, centered around speech and audio applications.

Conclusions

- I tried to summarize my recent work and goals concerning generalization, efficiency and interpretability, centered around speech and audio applications.
- Today I talked about:

Obtaining SOTA results **(SepFormer)**

Conclusions

- I tried to summarize my recent work and goals concerning generalization, efficiency and interpretability, centered around speech and audio applications.
- Today I talked about:

Obtaining SOTA results **(SepFormer)**

$\downarrow$

Systematic Evaluation of Real-data Perf. **(REAL-M)**

Conclusions

- I tried to summarize my recent work and goals concerning generalization, efficiency and interpretability, centered around speech and audio applications.
- Today I talked about:

Obtaining SOTA results **(SepFormer)**

↓

Systematic Evaluation of Real-data Perf. **(REAL-M)**

↓

Efficient representation learning **(CSSL+CrossModal)**

Conclusions

- I tried to summarize my recent work and goals concerning generalization, efficiency and interpretability, centered around speech and audio applications.
- Today I talked about:

Obtaining SOTA results **(SepFormer)**

↓

Systematic Evaluation of Real-data Perf. **(REAL-M)**

↓

Efficient representation learning **(CSSL+CrossModal)**

↓

Interpretability **(PIQ)**

Conclusions

- I tried to summarize my recent work and goals concerning generalization, efficiency and interpretability, centered around speech and audio applications.
- Today I talked about:

  Obtaining SOTA results **(SepFormer)**

  ↓

  Systematic Evaluation of Real-data Perf. **(REAL-M)**

  ↓

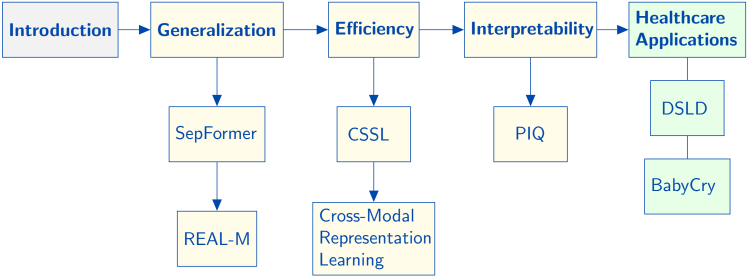
  Efficient representation learning **(CSSL+CrossModal)**

  ↓

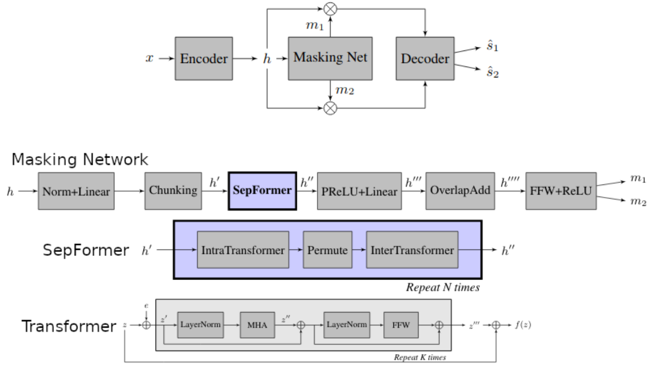
  Interpretability **(PIQ)**

- Machine Learning for Signal Processing Class in fall!
- Would love to chat if anything picks your attention!

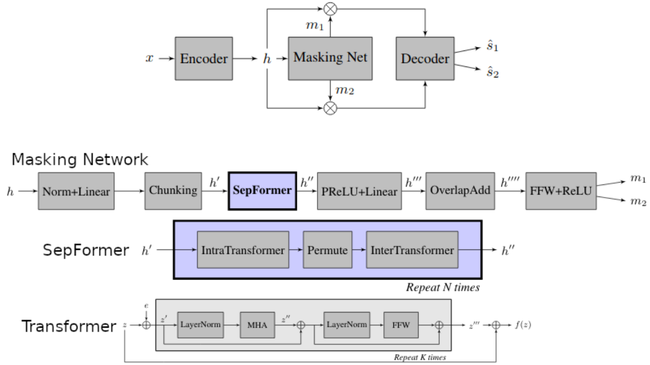
Thanks!


Introduction
Generalization
Efficiency
Interpretability
Healthcare Applications
SepFormer
REAL-M
CSSL
Cross-Modal Representation Learning
PIQ
DSLD
BabyCry

Appendix 1 - The SepFormer Masking Architecture



Appendix 1 - The SepFormer Masking Architecture



We train this architecture with permutation invariant SI-SNR.

$$s_{\text{target}} := \frac{\widehat{s}^\top s}{\|s\|^2} s, \quad e_{\text{noise}} := \widehat{s} - s_{\text{target}}, \quad \text{SI-SNR} := 10 \log_{10} \left(\frac{\|s_{\text{target}}\|^2}{\|e_{\text{noise}}\|^2} \right)$$

$$\text{PIT-SISNR} = \sum_k \min_{k' \in \mathcal{P}} 10 \log_{10} \left(\frac{\|s^k_{\text{target}}\|^2}{\|\widehat{s}^{k'} - s^k_{\text{target}}\|^2} \right)$$

Appendix2 - Tackling the lack of ground truth

- The same concept in speech separation:

Click for Mixture 1	Click for Mixture 2
Estimated Source 1	Estimated Source 1
Estimated Source 2	Estimated Source 2

Appendix2 - Tackling the lack of ground truth

- The same concept in speech separation:

  Click for Mixture 1
  Estimated Source 1
  Estimated Source 2

  Click for Mixture 2
  Estimated Source 1
  Estimated Source 2

- As humans we know if the estimation is good or not.
- Standard practice in evaluation:

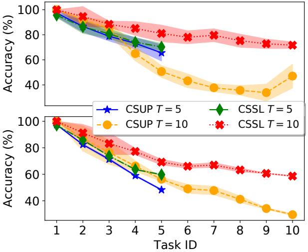
$$\text{Performance} \propto -d(\textit{estimate},\ \textit{groundtruth})$$

- However, it is not easy to have ground truth always. We can instead use a model to predict the performance as

$$\text{Performance estimate} = f(\textit{estimate},\ \textit{input})$$

  where, f is a neural network, *input* is the mixture in speech separation case.

Appendix3 - CSSL More Tasks



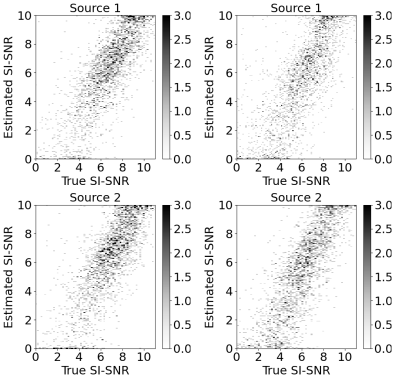
- SSL effectively combats forgetting, even without explicitly combatting forgetting!
- Offline accuracies: 80.9%, 68.2% with CSUP, 74.3%, 62.5% with CSSL.

Appendix4 - CSSL, More SSL Methods

Method	UrbanSound8K				DCASE TAU19			
	LEP		SLEP		LEP		SLEP	
	A (↑)	F (↓)	A (↑)	F (↓)	A (↑)	F (↓)	A (↑)	F (↓)
No distillation								
CSUP	65.6	19.6	48.4	33.1	48.2	27.6	32.5	38.4
SimCLR	**70.3**	15.3	**50.3**	26.6	**59.7**	**17.8**	**42.1**	27.9
Barlow Twins	68.5	**14.1**	49.7	**20.5**	55.9	19.0	41.0	**23.4**
MoCo	68.4	15.6	**50.3**	25.8	49.5	20.2	34.8	25.8
With distillation								
CSUP + $\mathcal{L}_{\text{MSE}}$	58.6	27.0	43.8	39.2	49.1	26.1	35.1	35.8
CSUP + $\mathcal{L}_{\text{sim}}$	70.6	**13.8**	54.9	27.1	56.2	19.7	42.1	32.4
CSUP + $\mathcal{L}_{\text{KLD}}$	69.8	15.9	**55.4**	27.4	55.7	19.4	42.6	30.3
SimCLR + $\mathcal{L}_{\text{MSE}}$	**70.9**	14.6	50.6	25.1	56.2	19.6	42.0	26.5
SimCLR + $\mathcal{L}_{\text{sim}}$	70.6	14.0	51.1	25.0	**60.0**	**17.6**	**42.8**	25.9
Barlow Twins + $\mathcal{L}_{\text{MSE}}$	69.5	14.0	47.3	**26.1**	56.2	19.8	41.1	25.6
Barlow Twins + $\mathcal{L}_{\text{sim}}$	70.0	13.4	49.5	23.9	55.1	19.8	41.2	**23.1**
MoCo + $\mathcal{L}_{\text{MSE}}$	67.7	14.8	51.4	25.4	49.4	21.4	34.1	27.9
MoCo + $\mathcal{L}_{\text{sim}}$	68.5	15.0	50.9	26.8	50.7	18.4	35.7	24.0

Appendix 5 - Evaluating the SI-SNR Estimator (Mismatch)

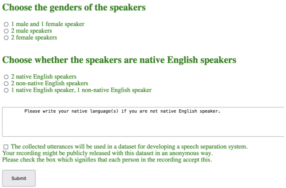
- We first evaluate the SI-SNR Estimator on synthetic data.



- Evaluating dprnn separator (left), convtasnet separator (right). The estimator is trained with SepFormer.
- Both scatter plots correspond to Pearson correlation coefficient of 0.8.

Data collection

- We crowdsource a dataset to construct an evaluation of audio mixtures in real-life. `Click to hear examples`.



Choose the genders of the speakers

- ○ 1 male and 1 female speaker
- ○ 2 male speakers
- ○ 2 female speakers

Choose whether the speakers are native English speakers

- ○ 2 native English speakers
- ○ 2 non-native English speakers
- ○ 1 native English speaker, 1 non-native English speaker

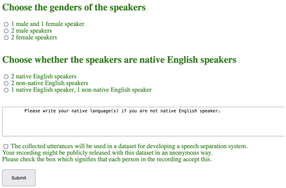
Please write your native language(s) if you are not native English speaker.

☐ The collected utterances will be used in a dataset for developing a speech separation system.
Your recording might be publicly released with this dataset in an anonymous way.
Please check the box which signifies that each person in the recording accept this.

Submit

Data collection

- We crowdsource a dataset to construct an evaluation of audio mixtures in real-life. `Click to hear examples`.



Choose the genders of the speakers

- ○ 1 male and 1 female speaker
- ○ 2 male speakers
- ○ 2 female speakers

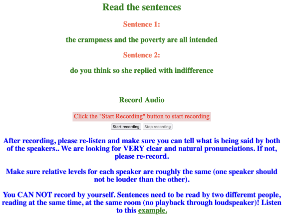
Choose whether the speakers are native English speakers

- ○ 2 native English speakers
- ○ 2 non-native English speakers
- ○ 1 native English speaker, 1 non-native English speaker

Please write your native language(s) if you are not native English speaker.

☐ The collected utterances will be used in a dataset for developing a speech separation system.
Your recording might be publicly released with this dataset in an anonymous way.
Please check the box which signifies that each person in the recording accept this.

Submit



Read the sentences

Sentence 1:

the crampness and the poverty are all intended

Sentence 2:

do you think so she replied with indifference

Record Audio

Click the "Start Recording" button to start recording

Start recording | Stop recording

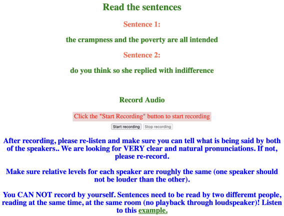
After recording, please re-listen and make sure you can tell what is being said by both of the speakers.. We are looking for VERY clear and natural pronunciations. If not, please re-record.

Make sure relative levels for each speaker are roughly the same (one speaker should not be louder than the other).

You CAN NOT record by yourself. Sentences need to be read by two differemt people, reading at the same time, at the same room (no playback through loudspeaker)! Listen to this example.

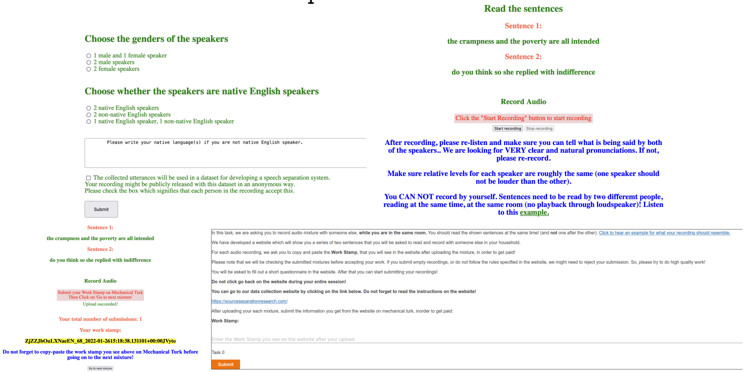
Data collection

- We crowdsource a dataset to construct an evaluation of audio mixtures in real-life. `Click to hear examples`.



Data collection

- We crowdsource a dataset to construct an evaluation of audio mixtures in real-life. `Click to hear examples`.



- Contributors are asked simultaneously read the shown sentences.
- This gives a way to collect real-life speech mixtures in a scalable way. We interface our platform with Mechanical Turk.
- We collected 3 hours of speech, from 50 unique speakers, with various native (e.g. US, UK) and non-native (e.g. French, Italian, Persian, Indian, African) accents, in various conditions, with various recording equipment.

SI-SNR Estimator Training Pipeline

- We construct a performance estimator $f(.)$ such that,

$$f(x, \widehat{s}) \approx \textbf{SI-SNR}(s, \widehat{s}),$$

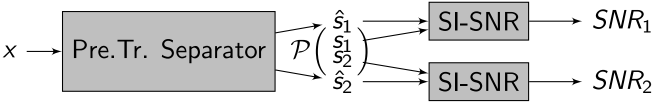
  where $f(.)$ is a neural network, x is the mixture, $\widehat{s}$ is the source estimate, s is the true source.

SI-SNR Estimator Training Pipeline

- We construct a performance estimator $f(.)$ such that,

$$f(x, \widehat{s}) \approx \textbf{SI-SNR}(s, \widehat{s}),$$

  where $f(.)$ is a neural network, x is the mixture, $\widehat{s}$ is the source estimate, s is the true source.
- We train the estimator on synthetic mixtures on which ground truth information is available. (Also a lot of data!)

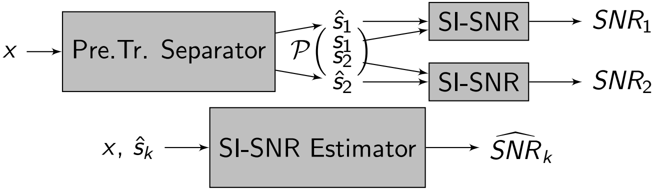


SI-SNR Estimator Training Pipeline

- We construct a performance estimator $f(.)$ such that,

$$f(x, \widehat{s}) \approx \textbf{SI-SNR}(s, \widehat{s}),$$

  where $f(.)$ is a neural network, x is the mixture, $\widehat{s}$ is the source estimate, s is the true source.
- We train the estimator on synthetic mixtures on which ground truth information is available. (Also a lot of data!)



$$\mathcal{L} \; = \; \|SNR_1 - \widehat{SNR}_1\| + \|SNR_2 - \widehat{SNR}_2\|.$$

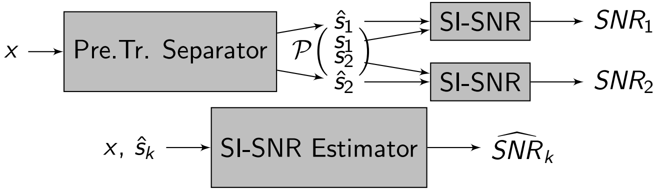
- SI-SNR Estimator is a 5-layer convolutional NNet in the time domain.

SI-SNR Estimator Training Pipeline

- We construct a performance estimator $f(.)$ such that,

$$f(x, \widehat{s}) \approx \textbf{SI-SNR}(s, \widehat{s}),$$

  where $f(.)$ is a neural network, x is the mixture, $\widehat{s}$ is the source estimate, s is the true source.
- We train the estimator on synthetic mixtures on which ground truth information is available. (Also a lot of data!)

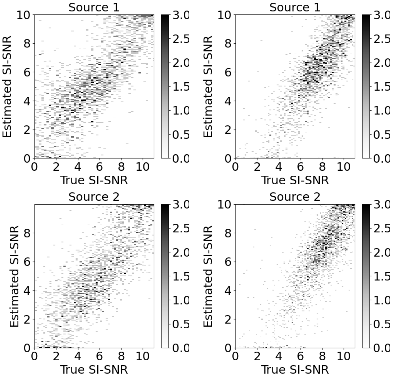


$$\mathcal{L} \ = \ \|SNR_1 - \widehat{SNR}_1\| + \|SNR_2 - \widehat{SNR}_2\|.$$

- SI-SNR Estimator is a 5-layer convolutional NNet in the time domain.
- **Important Question:** Is this estimator going to work well (generalize to) real-mixtures?

Evaluating the SI-SNR Estimator

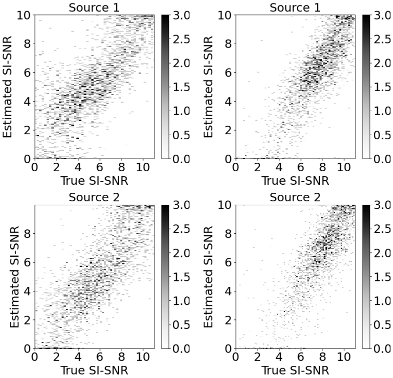
- We first evaluate the SI-SNR Estimator on synthetic data.



- Evaluating on (left) LibriMix, (right) WHAMR!
- Both scatter plots correspond to Pearson correlation coefficient of 0.8.

Evaluating the SI-SNR Estimator

- We first evaluate the SI-SNR Estimator on synthetic data.



- Evaluating on (left) LibriMix, (right) WHAMR!
- Both scatter plots correspond to Pearson correlation coefficient of 0.8.
- **Important Question:** Is this estimator going to work well (generalize to) real-mixtures?

Evaluating the SI-SNR Estimator on REAL-M

- We validate the SI-SNR estimator with a user study on real-life data.
- We presented 50 random mixtures and the separation results to 5 users.
- We asked the users to rate the presented separation result between 1-5.

Mixture

0:00 / 0:04

Estimate for Source 1

0:00 / 0:04

Estimate for Source 2

0:00 / 0:04

How good is the separation in your opinion for source 1? (Level of Separation + sound quality)

- bad
- poor
- fair
- good
- excellent

How good is the separation in your opinion for source 2? (Level of Separation + sound quality -- Note that if you hear the same source twice, this means the level of separation bad, so you should vote 'bad' in this case.)

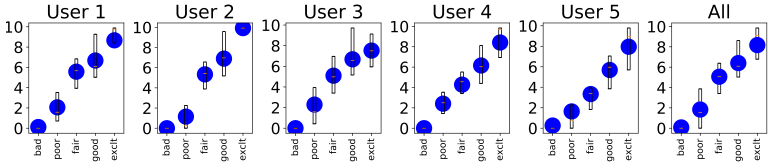
- bad
- poor
- fair
- good
- excellent

Submit

You have [7/50] submissions

Results of User Study

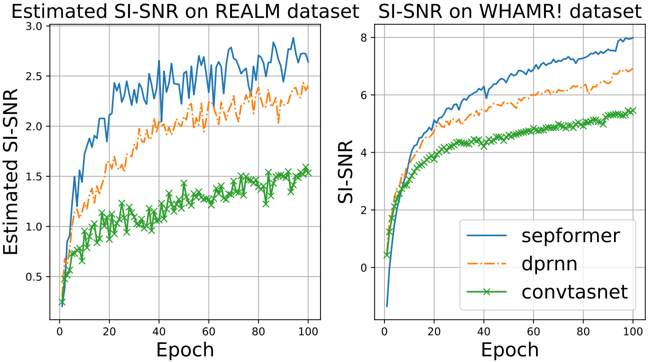
- Results of the user study suggest that on average the opinion scores correlate-well with SNR estimates.



- Y-axes show the estimated-SNR, X-axes show the user rating.

Further evaluation of SI-SNR Estimator

- The performance rankings of models on synthetic data holds true for REAL-M as well.
- We also observe that with training epochs performance on REAL-M dataset improves.



Conclusions

- With the REAL-M framework, we are moving towards working with real-mixtures. It provides,
  - Scalable data collections
  - Variability
  - Blind Performance estimation

Conclusions

- With the REAL-M framework, we are moving towards working with real-mixtures. It provides,
  - Scalable data collections
  - Variability
  - Blind Performance estimation
- The performance on real-life data is far worse than synthetic benchmarks.

Separator	$\widehat{\textbf{SNR Synth}}$	$\widehat{\textbf{SNR Real}}$
SepFormer	8.40	2.88
DPRNN	7.04	2.43
CTN	5.49	1.59

Conclusions

- With the REAL-M framework, we are moving towards working with real-mixtures. It provides,
  - Scalable data collections
  - Variability
  - Blind Performance estimation
- The performance on real-life data is far worse than synthetic benchmarks.

Separator	**$\widehat{\text{SNR}}$ Synth**	**$\widehat{\text{SNR}}$ Real**
SepFormer	8.40	2.88
DPRNN	7.04	2.43
CTN	5.49	1.59

- **Next steps:**
  - Casual talking, Meeting settings
  - Scaling up

Conclusions

- With the REAL-M framework, we are moving towards working with real-mixtures. It provides,
  - Scalable data collections
  - Variability
  - Blind Performance estimation
- The performance on real-life data is far worse than synthetic benchmarks.

Separator	**$\widehat{\text{SNR}}$ Synth**	**$\widehat{\text{SNR}}$ Real**
SepFormer	8.40	2.88
DPRNN	7.04	2.43
CTN	5.49	1.59

- **Next steps:**
  - Casual talking, Meeting settings
  - Scaling up
  - Improving the generalization (Data augmentations, Using the performance estimators, Using pretrained models, ...)