UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Heterogeneous target speech separation

Efthymios Tzinis[1,2,*], Gordon Wichern[1], Aswin Subramanian[1], Paris Smaragdis[2] and Jonathan Le Roux[1]

[1]Mitsubishi Electric Research Laboratories (MERL)
[2]University of Illinois at Urbana-Champaign

*Work done during an internship at MERL.

INTERSPEECH 2022
September 18 - 22 • Incheon Korea

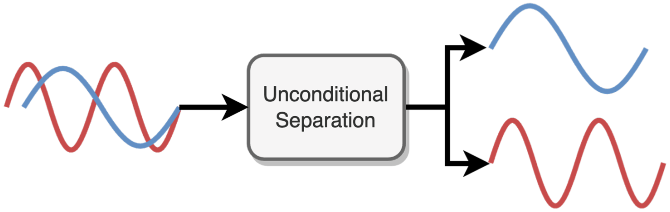
MITSUBISHI ELECTRIC
Changes for the Better

Efthymios Tzinis
etzinis2@illinois.edu
etzinis.com

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

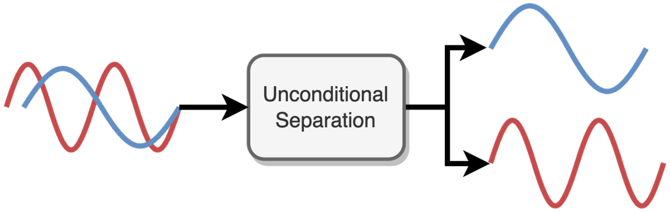
Introduction

- Audio source separation
  - Co-occurence of multiple sounds
  - Extract independent sound sources
    - **All sources:** Unconditional source separation
    - **Specify sources:** Conditional / Target source separation

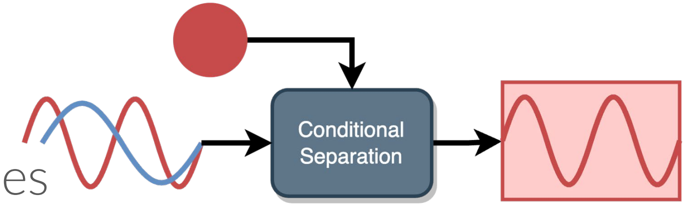


Introduction

- Audio source separation
  - Co-occurence of multiple sounds
  - Extract independent sound sources
    - **All sources:** Unconditional source separation
    - **Specify sources:** Conditional / Target source separation



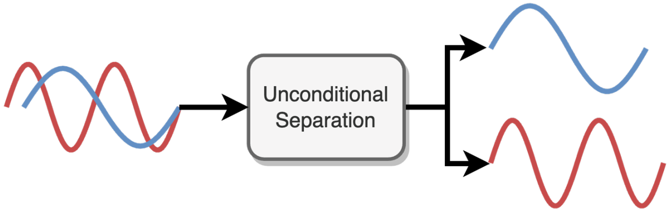
- Target speech separation
  - Solves the disambiguation of the sources
  - Solves the alignment of the estimated sources



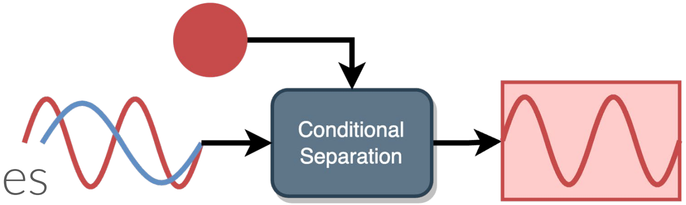
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Introduction

- Audio source separation
  - Co-occurence of multiple sounds
  - Extract independent sound sources
    - **All sources:** Unconditional source separation
    - **Specify sources:** Conditional / Target source separation



- Target speech separation
  - Solves the disambiguation of the sources
  - Solves the alignment of the estimated sources



- What kind of conditional targets can we use?

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

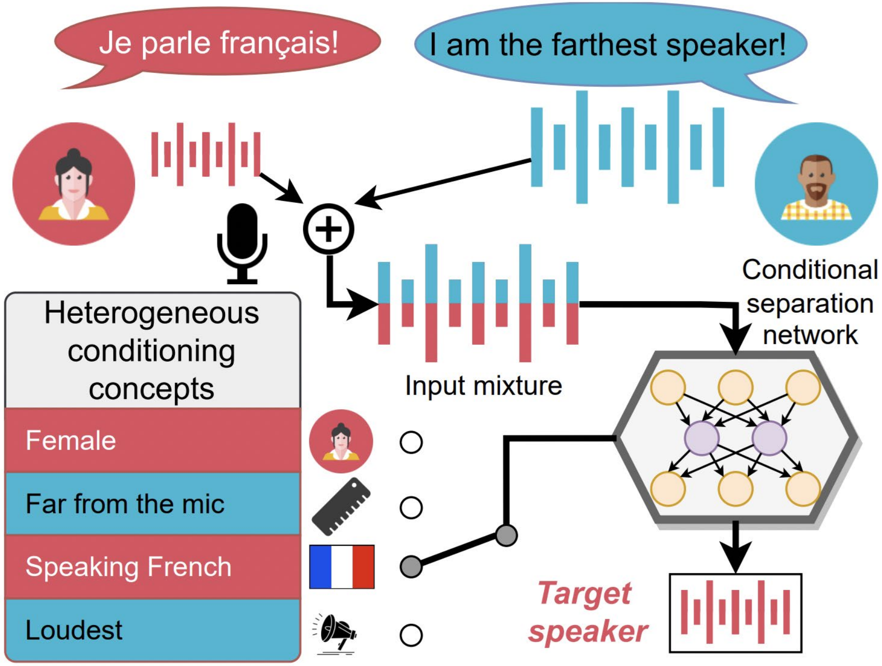
Heterogeneous target separation

- Slicing an acoustic scene has multiple solutions
  - Based on user's intention
  - Multiple ways to describe the same target source

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

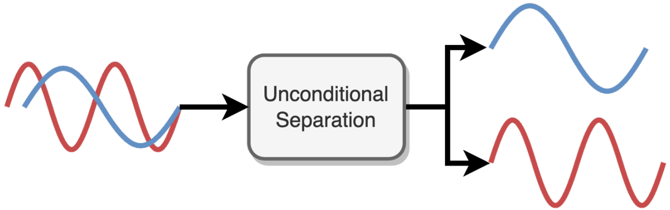
Heterogeneous target separation

- Slicing an acoustic scene has multiple solutions
  - Based on user’s intention
  - Multiple ways to describe the same target source
- Isolate a speaker based on different semantic concepts
  - Gender
  - Distance from the microphone
    - Far/Near microphone
  - Language spoken
    - French, English, etc.
  - Energy of the speaker
    - Loudest / Less energetic



Heterogeneous training

- Permutation invariant training (Oracle)
  - Backpropagate the minimum loss under all permutations of the estimated speakers

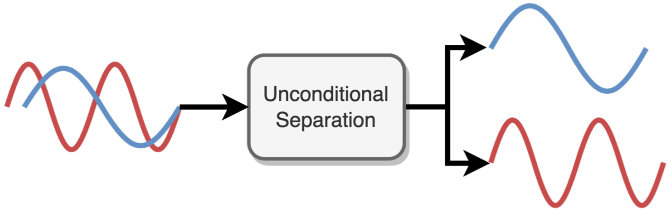


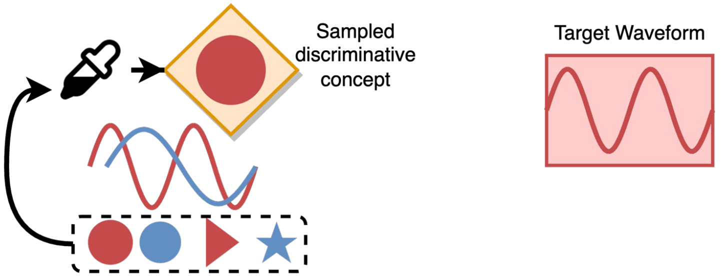
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Heterogeneous training

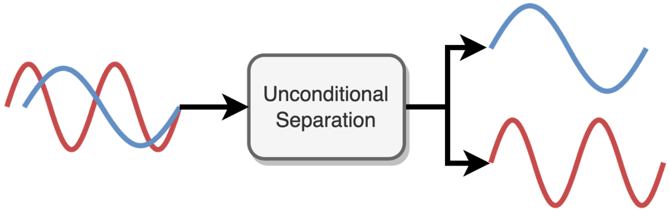
- Permutation invariant training (Oracle)
  - Backpropagate the minimum loss under all permutations of the estimated speakers
- Heterogeneous
  - Generate a mixture from a set of sources
  - Sample a discriminative concept to create the target waveform
    - Could contain more than one sources

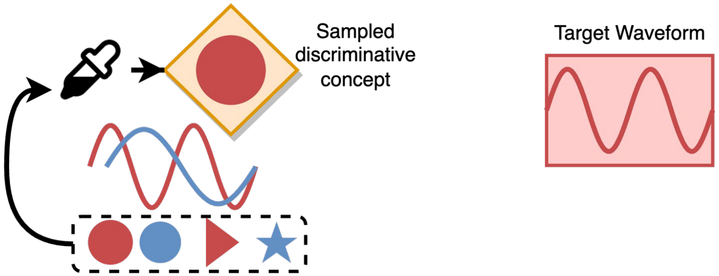


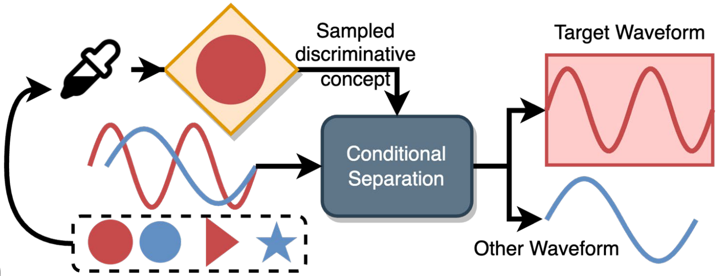


Heterogeneous training

- Permutation invariant training (Oracle)
  - Backpropagate the minimum loss under all permutations of the estimated speakers
- Heterogeneous
  - Generate a mixture from a set of sources
  - Sample a discriminative concept to create the target waveform
    - Could contain more than one sources
  - Train the model under a targeted L1 loss
  - Example conditions and their **discriminative concepts**:
    - Distance from the microphone: (**Far** or **Near**)
    - Language spoken: (**French**, **English**, etc.)







Introduced datasets

- Generated three different datasets
  - Wall Street Journal (WSJ - anechoic)
    - Energy (**E**), gender (**G**)
  - Spatial LibriSpeech (SLIB - reverberant)
    - **E**, **G**, spatial location (**S**)
  - Spatial VoxForge (SVOX - multi-lingual and reverberant):
    - **E**, **S**, language (**L**)

https://github.com/etzinis/heterogeneous_separation

Metadata	WSJ	SLIB	SVOX
Conditions $\mathscr{C}$	$\{\mathcal{E}, \mathcal{G}\}$	$\{\mathcal{E}, \mathcal{G}, \mathcal{S}\}$	$\{\mathcal{E}, \mathcal{L}, \mathcal{S}\}$
Room height (m)	-	$\mathcal{U}[2.6, 3.5]$	$\mathcal{U}[2.75, 3.25]$
Room length (m)	-	$\mathcal{U}[9.0, 11.0]$	$\mathcal{U}[8.0, 10.0]$
Room width (m)	-	$\mathcal{U}[9.0, 11.0]$	$\mathcal{U}[8.0, 10.0]$
RT 60 (sec)	-	$\mathcal{U}[0.3, 0.6]$	$\mathcal{U}[0.4, 0.6]$
Microphone location	-	Center	Center
Source height (m)	-	$\mathcal{U}[1.5, 2.0]$	$\mathcal{U}[1.6, 1.9]$
Far field distance (m)	-	$\mathcal{U}[1.7, 3.0]$	$\mathcal{U}[1.5, 2.5]$
Near field distance (m)	-	$\mathcal{U}[0.2, 0.6]$	$\mathcal{U}[0.3, 0.5]$
Number of test recordings	1,770	2,620	11,083
Number of test speakers	18	40	294
Number of train recordings	8,769	132,553	124,937
Number of train speakers	101	1172	2347
Number of val recordings	3,557	2,703	10,244
Number of val speakers	101	40	279

Conditional separation network

- Conditional sudo rm -rf
    - One-hot conditioning vector based on all semantic concepts

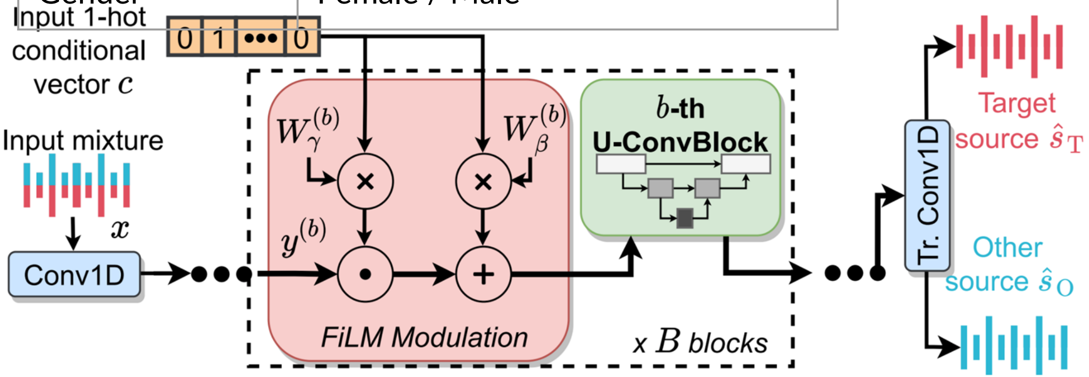
Condition	Discriminative concept values
Energy	Loudest / Most silent
Spatial Location	Far / Near field
Language	English / French / German / Spanish
Gender	Female / Male

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Conditional separation network

- Conditional sudo rm -rf
  - One-hot conditioning vector based on all semantic concepts
  - FiLM modulation in the input of all B=16 U-ConvBlocks
  - Always estimate the target and the non-target estimate

Condition	Discriminative concept values
Energy	Loudest / Most silent
Spatial Location	Far / Near field
Language	English / French / German / Spanish
Gender	Female / Male

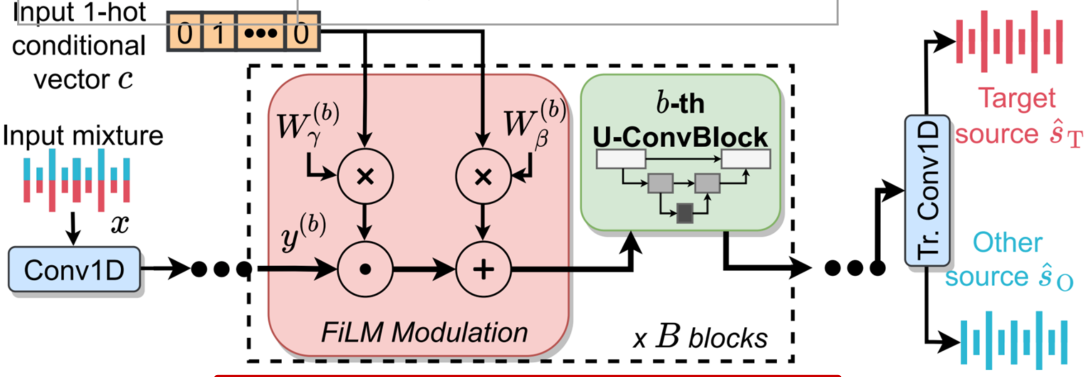


UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Conditional separation network

- Conditional sudo rm -rf
  - One-hot conditioning vector based on all semantic concepts
  - FiLM modulation in the input of all B=16 U-ConvBlocks
  - Always estimate the target and the non-target estimate
  - **Low overhead** conditioning mechanism

Condition	Discriminative concept values
Energy	Loudest / Most silent
Spatial Location	Far / Near field
Language	English / French / German / Spanish
Gender	Female / Male



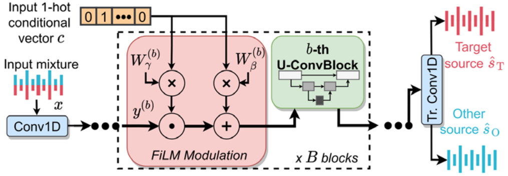
Parameters: 9.66 millions -> **9.84** millions

Training and evaluation details

- Training
  - Sample a discriminative concept given a pre-defined prior
  - L1 norm for both "target" and "other" estimated sources
    - We train for 120 epochs
      - 20,000 8kHz mixtures
      - Uniform [75-100]% overlap

Condition	WSJ	SVOX	SLIB
Input-SNR	Uniform [-5,5]	Uniform [-2.5, 2.5]	
Conditions	Energy, Gender	Energy, Gender, Spatial Loc.	Energy, Language, Spatial Loc.

$$L_{\boldsymbol{\theta}} = |\widehat{\mathbf{s}}_{\mathrm{T}} - \mathbf{s}_{\mathrm{T}}| + |\widehat{\mathbf{s}}_{\mathrm{O}} - \mathbf{s}_{\mathrm{O}}| \quad \widehat{\mathbf{s}}_{\mathrm{T}}, \widehat{\mathbf{s}}_{\mathrm{O}} = f(\mathbf{x}, \mathbf{c}; \boldsymbol{\theta})$$



Training and evaluation details

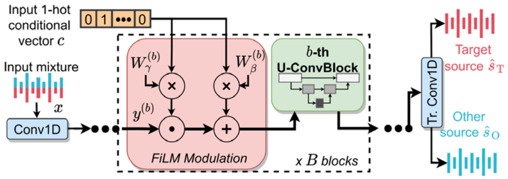
- Training
  - Sample a discriminative concept given a pre-defined prior
  - L1 norm for both "target" and "other" estimated sources
    - We train for 120 epochs
      - 20,000 8kHz mixtures
      - Uniform [75-100]% overlap

Condition	WSJ	SVOX	SLIB
Input-SNR	Uniform [-5,5]	Uniform [-2.5, 2.5]	
Conditions	Energy, Gender	Energy, Gender, Spatial Loc.	Energy, Language, Spatial Loc.

$$L_{\boldsymbol{\theta}} = |\widehat{\mathbf{s}}_{\mathrm{T}} - \mathbf{s}_{\mathrm{T}}| + |\widehat{\mathbf{s}}_{\mathrm{O}} - \mathbf{s}_{\mathrm{O}}| \quad \widehat{\mathbf{s}}_{\mathrm{T}}, \widehat{\mathbf{s}}_{\mathrm{O}} = f(\mathbf{x}, \mathbf{c}; \boldsymbol{\theta})$$

- Evaluation
  - Scale-invariant signal to noise ratio on the target source
  - 3,000 validation mixtures
  - 5,000 test mixtures

$$\alpha = \mathbf{s}_{\mathrm{T}}^{\top} \widehat{\mathbf{s}}_{\mathrm{T}} / \|\widehat{\mathbf{s}}\|^2$$



$$\text{SI-SDR}(\widehat{\mathbf{s}}_{\mathrm{T}}, \mathbf{s}_{\mathrm{T}}) = -20 \log_{10} \left(\|\alpha \mathbf{s}_{\mathrm{T}}\| / \|\alpha \mathbf{s}_{\mathrm{T}} - \widehat{\mathbf{s}}_{\mathrm{T}}\| \right)$$

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

In- and cross-domain results

- Single-conditioned models > PIT
  - Each model trained and evaluated on the corresponding condition

Training method	$\lvert\mathscr{D}\rvert$	$\lvert\mathscr{C}\rvert$	Train condition priors (%) SLIB $\mathcal{G}$	SLIB $\mathcal{S}$	SVOX $\mathcal{L}$	SVOX $\mathcal{S}$	Test conditions SLIB $\mathcal{G}$	SLIB $\mathcal{S}$	SVOX $\mathcal{L}$	SVOX $\mathcal{S}$
Conditioned*	1	1	100	100	100	100	**11.4**	**11.2**	2.5	**9.1**
PIT (Oracle)*	1	1	100	100	100	100	11.0	10.7	4.6	7.5
In-domain heterogeneous	1	2	50	50			10.9	10.7	−0.5	8.6
					50	50	−0.6	6.2	3.2	6.8
PIT (Oracle)	1	2	50	50			9.5	8.9	**5.6**	6.8
					50	50	5.2	4.5	4.6	5.6
Cross-domain heterogeneous	2	2		50	25	25	−1.4	9.2	4.3	8.2
			25	25		50	9.9	9.9	−0.7	9.0
			50			50	10.1	8.9	−0.9	9.0
				50	50		−0.5	8.4	4.3	6.8
	2	3	25	25	25	25	8.9	8.7	4.4	7.8
PIT (Oracle)	2	3	25	25	25	25	8.0	7.3	5.5	6.5

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

In- and cross-domain results

- Single-conditioned models > PIT
  - Each model trained and evaluated on the corresponding condition
- Heterogeneous training > PIT
  - For all conditions except language
  - For **in-domain data**

Training method	$\lvert\mathcal{D}\rvert$	$\lvert\mathcal{C}\rvert$	Train condition priors (%) SLIB $\mathcal{G}$	SLIB $\mathcal{S}$	SVOX $\mathcal{L}$	SVOX $\mathcal{S}$	Test conditions SLIB $\mathcal{G}$	SLIB $\mathcal{S}$	SVOX $\mathcal{L}$	SVOX $\mathcal{S}$
Conditioned*	1	1	100	100	100	100	**11.4**	**11.2**	2.5	**9.1**
PIT (Oracle)*	1	1	100	100	100	100	11.0	10.7	4.6	7.5
In-domain heterogeneous	1	2	50	50			10.9	10.7	−0.5	8.6
					50	50	−0.6	6.2	3.2	6.8
PIT (Oracle)	1	2	50	50			9.5	8.9	**5.6**	6.8
					50	50	5.2	4.5	4.6	5.6
Cross-domain heterogeneous	2	2		50	25	25	−1.4	9.2	4.3	8.2
			25	25		50	9.9	9.9	−0.7	9.0
			50			50	10.1	8.9	−0.9	9.0
				50	50		−0.5	8.4	4.3	6.8
	2	3	25	25	25	25	8.9	8.7	4.4	7.8
PIT (Oracle)	2	3	25	25	25	25	8.0	7.3	5.5	6.5

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

In- and cross-domain results

- Single-conditioned models > PIT
  - Each model trained and evaluated on the corresponding condition
- Heterogeneous training > PIT
  - For all conditions except language
  - For **in-domain data**

Training method	$\|\mathcal{D}\|$	$\|\mathcal{C}\|$	Train condition priors (%) SLIB $\mathcal{G}$	SLIB $\mathcal{S}$	SVOX $\mathcal{L}$	SVOX $\mathcal{S}$	Test conditions SLIB $\mathcal{G}$	SLIB $\mathcal{S}$	SVOX $\mathcal{L}$	SVOX $\mathcal{S}$
Conditioned*	1	1	100	100	100	100	**11.4**	**11.2**	2.5	**9.1**
PIT (Oracle)*	1	1	100	100	100	100	11.0	10.7	4.6	7.5
In-domain heterogeneous	1	2	50	50			10.9	10.7	−0.5	8.6
					50	50	−0.6	6.2	3.2	6.8
PIT (Oracle)	1	2	50	50			9.5	8.9	**5.6**	6.8
					50	50	5.2	4.5	4.6	5.6
Cross-domain heterogeneous	2	2		50	25	25	−1.4	9.2	4.3	8.2
			25	25		50	9.9	9.9	−0.7	9.0
			50			50	10.1	8.9	−0.9	9.0
				50	50		−0.5	8.4	4.3	6.8
	2	3	25	25	25	25	8.9	8.7	4.4	7.8
PIT (Oracle)	2	3	25	25	25	25	8.0	7.3	5.5	6.5

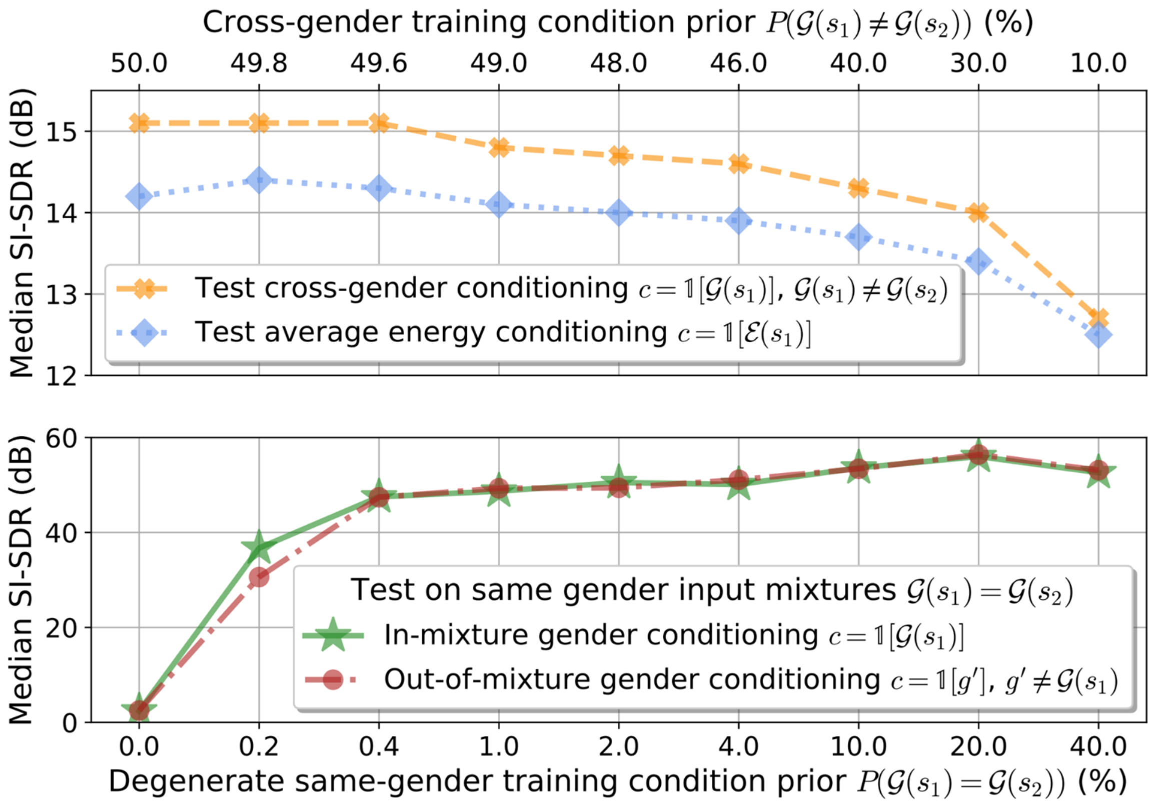
In- and cross-domain results

- Single-conditioned models > PIT
  - Each model trained and evaluated on the corresponding condition
- Heterogeneous training > PIT
  - For all conditions except language
  - For **in-domain data**
  - For **cross-domain** evaluation

Training method	$\lvert\mathcal{D}\rvert$	$\lvert\mathcal{C}\rvert$	Train condition priors (%) SLIB $\mathcal{G}$	SLIB $\mathcal{S}$	SVOX $\mathcal{L}$	SVOX $\mathcal{S}$	Test conditions SLIB $\mathcal{G}$	SLIB $\mathcal{S}$	SVOX $\mathcal{L}$	SVOX $\mathcal{S}$
Conditioned*	1	1	100	100	100	100	**11.4**	**11.2**	2.5	**9.1**
PIT (Oracle)*	1	1	100	100	100	100	11.0	10.7	4.6	7.5
In-domain heterogeneous	1	2	50	50			10.9	10.7	−0.5	8.6
					50	50	−0.6	6.2	3.2	6.8
PIT (Oracle)	1	2	50	50			9.5	8.9	**5.6**	6.8
					50	50	5.2	4.5	4.6	5.6
Cross-domain heterogeneous	2	2		50	25	25	−1.4	9.2	4.3	8.2
			25	25		50	9.9	9.9	−0.7	9.0
			50	✖		50	10.1	8.9	−0.9	9.0
				50	50	✖	−0.5	8.4	4.3	6.8
	2	3	25	25	25	25	8.9	8.7	4.4	7.8
PIT (Oracle)	2	3	25	25	25	25	8.0	7.3	5.5	6.5

Robustness under degenerate conditions

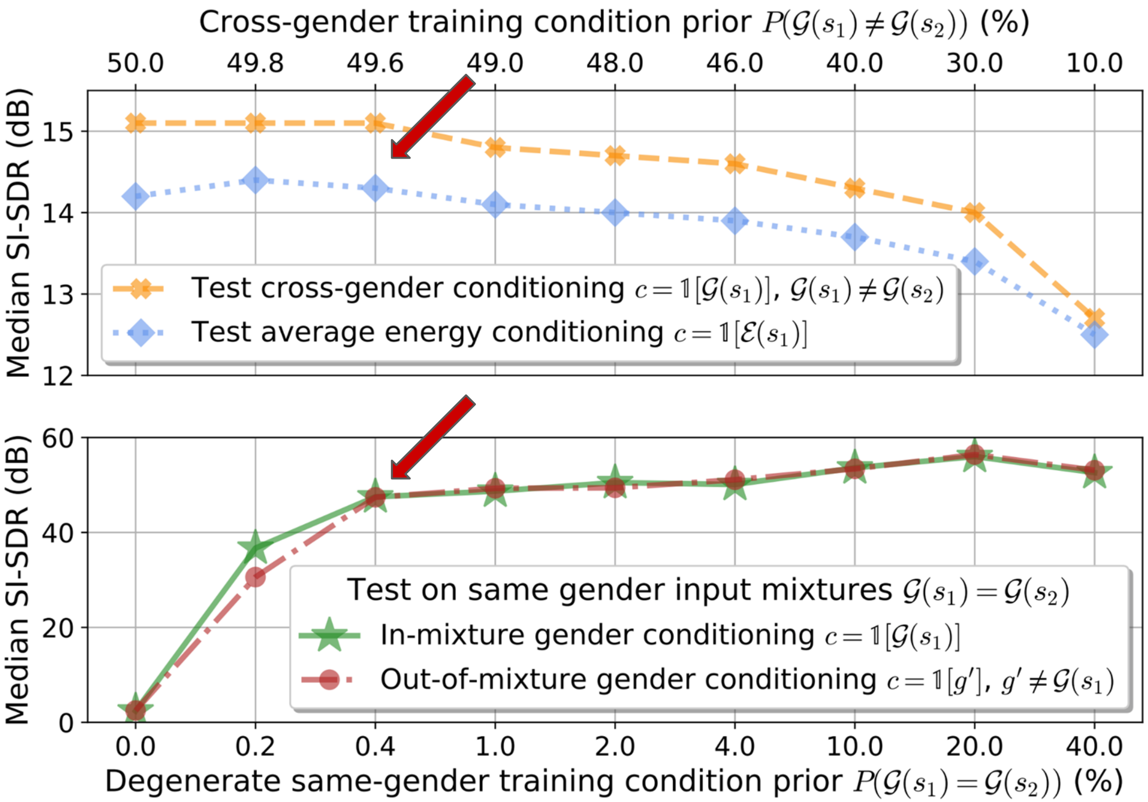
- Trade-off between the percentage of:
  - Same gender conditioning
  - Cross-gender



UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Robustness under degenerate conditions

- Trade-off between the percentage of:
  - Same gender conditioning
  - Cross-gender
- Optimal point for both gender and energy conditions
  - Using only 0.2-0.4% of same-gender mixtures
  - Also learns the degenerate case



UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Bridge conditioning ablation

Training method	Train condition priors (%)				Test conditions			
	WSJ		SLIB		WSJ		SLIB	
	$\mathcal{G}$	$\mathcal{E}$	$\mathcal{G}$	$\mathcal{E}$	$\mathcal{G}$	$\mathcal{E}$	$\mathcal{G}$	$\mathcal{E}$
Proposed	25	25		50	13.3	12.4	7.1	8.8

- Learn a harder discriminative concept (e.g. gender on SLIB)
  - No access to SLIB gender metadata about the speakers
  - Learn using the energy concept as a "bridge" condition
    - Possible available metadata for the WSJ anechoic dataset

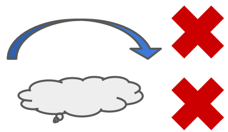
Bridge conditioning ablation

Training method	Train condition priors (%)				Test conditions			
	WSJ		SLIB		WSJ		SLIB	
	$\mathcal{G}$	$\mathcal{E}$	$\mathcal{G}$	$\mathcal{E}$	$\mathcal{G}$	$\mathcal{E}$	$\mathcal{G}$	$\mathcal{E}$
Proposed	25	25	✗	50	13.3	12.4	7.1	8.8
(-) Bridge condition	50		✗	50	14.5	7.4	5.5	9.2

- Learn a harder discriminative concept (e.g. gender on SLIB)
  - No access to SLIB gender metadata about the speakers
  - Learn using the energy concept as a "bridge" condition
    - Possible available metadata for the WSJ anechoic dataset

Bridge conditioning ablation

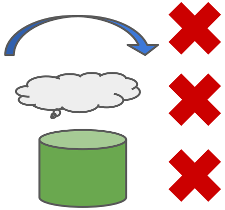
Training method	Train condition priors (%)				Test conditions			
	WSJ		SLIB		WSJ		SLIB	
	$\mathcal{G}$	$\mathcal{E}$	$\mathcal{G}$	$\mathcal{E}$	$\mathcal{G}$	$\mathcal{E}$	$\mathcal{G}$	$\mathcal{E}$
Proposed	25	25	✖	50	13.3	12.4	$\boxed{7.1}$	8.8
(-) Bridge condition	50		✖	50	14.5	7.4	$\boxed{5.5}$	9.2
(-) Exclude amb. $\mathcal{E}$ cases	25	25	✖	50	13.0	11.8	$\boxed{6.2}$	8.4



- Learn a harder discriminative concept (e.g. gender on SLIB)
  - No access to SLIB gender metadata about the speakers
  - Learn using the energy concept as a “bridge” condition
    - Possible available metadata for the WSJ anechoic dataset

Bridge conditioning ablation

Training method	Train condition priors (%) WSJ $\mathcal{G}$	WSJ $\mathcal{E}$	SLIB $\mathcal{G}$	SLIB $\mathcal{E}$	Test conditions WSJ $\mathcal{G}$	WSJ $\mathcal{E}$	SLIB $\mathcal{G}$	SLIB $\mathcal{E}$
Proposed	25	25	✖	50	13.3	12.4	7.1	8.8
(-) Bridge condition	50		✖	50	14.5	7.4	5.5	9.2
(-) Exclude amb. $\mathcal{E}$ cases	25	25	✖	50	13.0	11.8	6.2	8.4
(-) In-domain data	100		✖		**17.3**	−2.4	5.8	−2.3
	50	50	✖		15.2	**14.3**	4.2	3.0

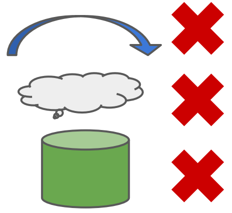


- Learn a harder discriminative concept (e.g. gender on SLIB)
  - No access to SLIB gender metadata about the speakers
  - Learn using the energy concept as a “bridge” condition
    - Possible available metadata for the WSJ anechoic dataset

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Bridge conditioning ablation

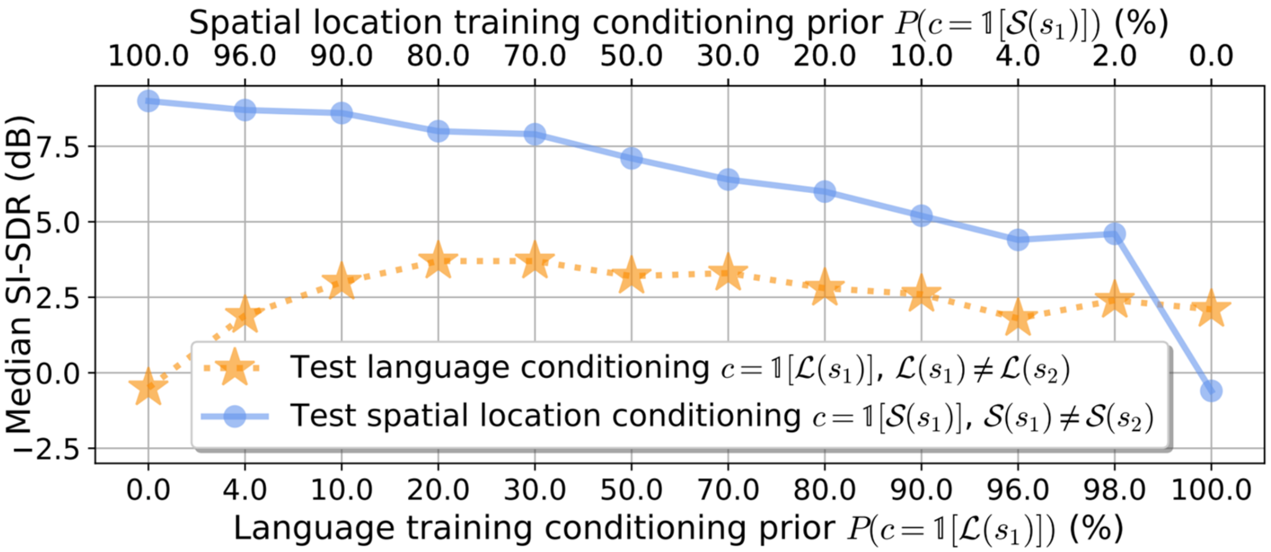
Training method	Train condition priors (%)				Test conditions			
	WSJ		SLIB		WSJ		SLIB	
	$\mathcal{G}$	$\mathcal{E}$	$\mathcal{G}$	$\mathcal{E}$	$\mathcal{G}$	$\mathcal{E}$	$\mathcal{G}$	$\mathcal{E}$
Proposed	25	25	✖	50	13.3	12.4	7.1	8.8
(-) Bridge condition	50		✖	50	14.5	7.4	5.5	9.2
(-) Exclude amb. $\mathcal{E}$ cases	25	25	✖	50	13.0	11.8	6.2	8.4
(-) In-domain data	100		✖		**17.3**	−2.4	5.8	−2.3
	50	50	✖		15.2	**14.3**	4.2	3.0
PIT (Oracle)*	100	100	100	100	**17.3**	13.6	**10.9**	**10.2**
PIT (Oracle)	25	25	25	25	12.9	11.9	9.3	8.5



- Learn a harder discriminative concept (e.g. gender on SLIB)
  - No access to SLIB gender metadata about the speakers
  - Learn using the energy concept as a “bridge” condition
    - Possible available metadata for the WSJ anechoic dataset

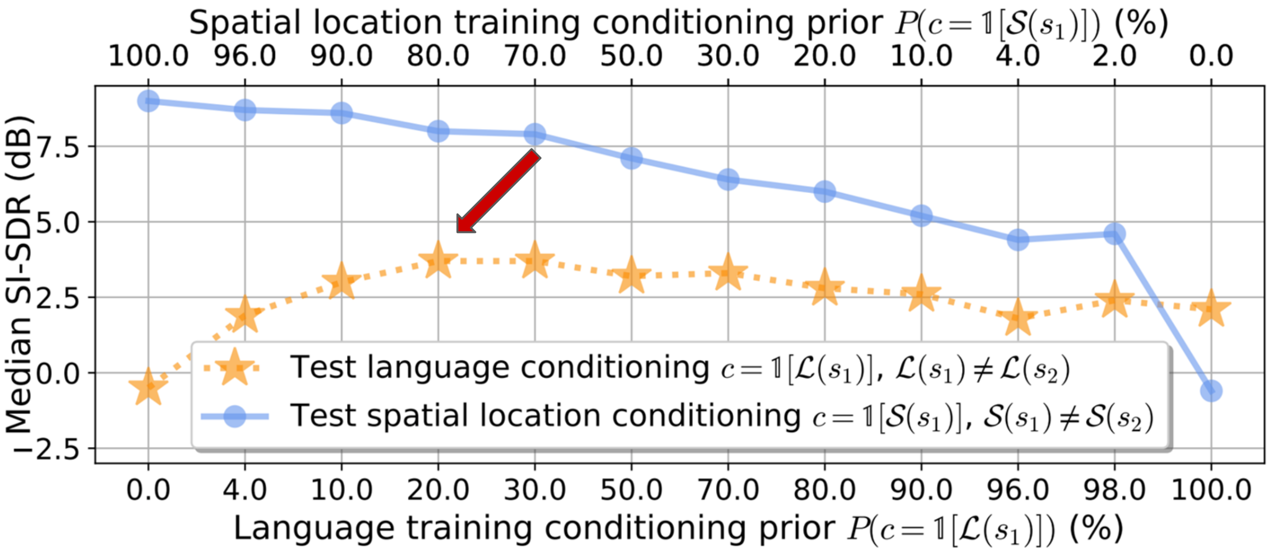
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Using a bridge semantic condition



- Learn a hard condition using an easier one
  - Learn how to condition on a specific **language** using the **spatial location**

Using a bridge semantic condition



- Learn a hard condition using an easier one
  - Learn how to condition on a specific **language** using the **spatial location**
  - Best model for both conditions appears to be in between the two extremes
    - The training conditioning prior is key

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Conclusions & Highlights

- Heterogeneous target source separation
  - A new paradigm in source separation
  - Slicing acoustic scenes based on deviant:
    - **Non-mutually exclusive** signal characteristic conditions
      - One can also consider using **AND** and **OR** conditions

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Conclusions & Highlights

- Heterogeneous target source separation
  - A new paradigm in source separation
  - Slicing acoustic scenes based on deviant:
    - **Non-mutually exclusive** signal characteristic conditions
      - One can also consider using **AND** and **OR** conditions
- Heterogeneous condition training
  - **Improves upon oracle permutation invariant training**
  - Improves cross-domain **generalization**
  - **Robust** under degenerate cases

Conclusions & Highlights

- Heterogeneous target source separation
  - A new paradigm in source separation
  - Slicing acoustic scenes based on deviant:
    - **Non-mutually exclusive** signal characteristic conditions
      - One can also consider using **AND** and **OR** conditions
- Heterogeneous condition training
  - **Improves upon oracle permutation invariant training**
  - Improves cross-domain **generalization**
  - **Robust** under degenerate cases
- In the future
  - We want to apply our method towards a **variable number of sources**
  - Make our method require **less supervision**
  - Extend out method to work with natural language queries

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Thank you!

Any questions?

https://github.com/etzinis/heterogeneous_separation

Efthymios Tzinis
etzinis2@illinois.edu
etzinis.com