Hierarchical Musical Instrument Separation

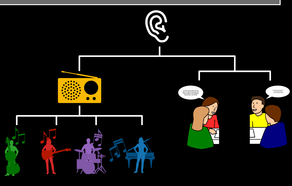
Ethan Manilow, Gordon Wichern, and Jonathan Le Roux

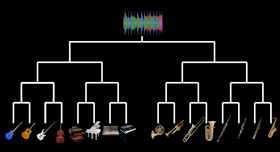
Northwestern ENGINEERING

MITSUBISHI ELECTRIC

Problem Statement

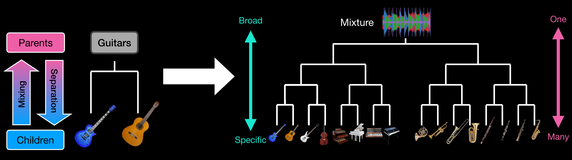
- Many auditory scenes are hierarchical & humans perceive them as hierarchical
- Source Separation usually assumes the scene is **flat**





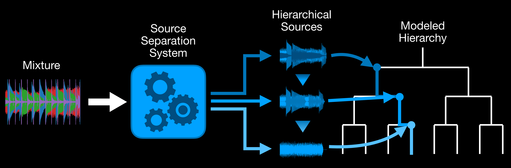
- Musical instruments have long been classified hierarchically
- ***Can we separate musical mixes at multiple hierarchical levels?***

Auditory Hierarchies



- We assume sources can be combined and separated hierarchically
- A **path** is a sequence of sources between nodes at different levels

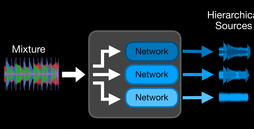
Hierarchical Source Separation



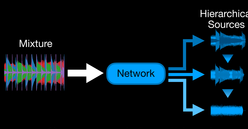
Source separation is **hierarchical** if it separates along a hierarchical path

Here, we outline paradigms for **hierarchical source separation**:

Single Level vs Multi Level



One network **per** level
Trained **separately**

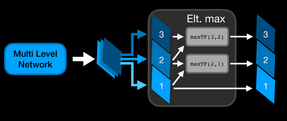


One network **for all** levels
Trained **jointly**

Multi Level Hierarchical Constraints

Want net to learn to put more energy in parent sources

Element-wise max across TF bins at adjacent hierarchy levels



Single Instrument vs Multi Instrument

Source Specific Separation (SSS)
Only separates one path

Query-by-Example (QBE)
Separate any path based on query

Experimental Design

Dataset & Evaluation

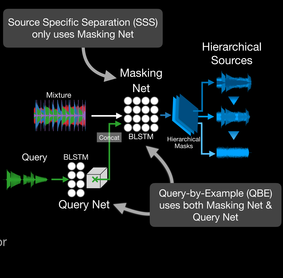
- Use *Slakh2100-split2* for train/val/test
- 3 hierarchy levels (+ mix)
- 10 sec clips. 10k train, 5k val, 3k test.
- Report SI-SDR Improvement (in dB)

Source Specific Separation (SSS) Network Hierarchical Levels	
Level	Submixes to be separated
3	Keyboards, guitars, and orchestral strings
2	All guitars (both clean and effected)
1	Only clean guitars (both electric and acoustic)

Query-by-Example (QBE) Network Hierarchical Levels
34 instrument categories
Full hierarchy description: https://git.io/J4gx

Network Details

- Masking Network (SSS & QBE)
  - 4 Layer BLSTM network
    - 600 units, 0.3 dropout
  - FC layer w/ sigmoid
    - Makes Masks
- Query Network (Just QBE)
  - 2 Layer BLSTM network
    - 600 units, 0.3 dropout
  - FC layer
    - Makes Query Embedding Anchor

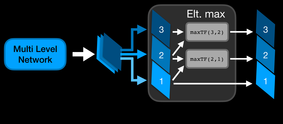


Experiment 1:

Does the Hierarchical Constraint help?

Main Idea

Test Multi Level Networks
with & *without*
the Hierarchical Constraint (HC)



Results

Single Instrument (SSS) or Multi Instrument (QBE)

Use Hierarchical Constraint?

Grandparent Source
Parent Source
Child Source

Model Type	HC	Level 3	Level 2	Level 1
SSS (Guitar)		3.5	4.0	4.0
SSS (Guitar)	✓	3.2	3.6	3.8
QBE		3.2	2.4	0.2
QBE	✓	3.3	2.1	1.6

SI-SDR Improvement (dB)
↑ is better

HC does not help as much with SSS. We think the net learns the "contour" of the instrument's mask better

HC helps much more in Multi Instrument case esp. at leaf level. It gives the QBE net a boost.

Experiment 2:
Are Single Level or Multi Level Networks better?

Main Idea

Test **Single Level** vs. **Multi Level** Networks in two cases:

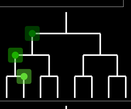
1) **Single Instrument**
Source Specific Separation (SSS) Net

2) **Multi Instrument**
Query-by-Example (QBE) Net

Four Cases in Total:

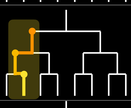
Single Instrument
Single Level

3 SSS models
- 1 per level
- 1 path



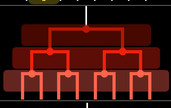
Single Instrument
Multi Level

1 SSS models
- All levels
- 1 path



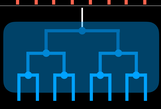
Multi Instrument
Single Level

3 QBE models
- 1 per level
- Any path



Multi Instrument
Multi Level

1 QBE model
- All levels
- Any path



Results

Unprocessed SI-SDR
Net's SI-SDR
SI-SDR Change

Grandparent ← → Child

x3 Single Level Models

Single Instrument (SSS) or Multi Instrument (QBE)

Model Type	lvls #	All Levels Mix	All Levels SI-SDR	All Levels Δ	Level 3 Mix	Level 3 SI-SDR	Level 3 Δ	Level 2 Mix	Level 2 SI-SDR	Level 2 Δ	Level 1 Mix	Level 1 SI-SDR	Level 1 Δ
SSS (Guitar)	1	−3.9	−2.1	1.8	0.9	4.1	3.2	−5.9	−3.2	2.7	−6.6	−7.3	−0.7
SSS (Guitar)	3	−3.9	0.0	3.9	0.9	4.3	3.4	−5.9	−1.9	4.0	−6.6	−2.6	4.0
QBE	1	−4.9	−3.9	1.0	−1.3	2.0	3.3	−5.3	−3.9	1.4	−8.0	−9.8	−1.9
QBE	3	−4.9	−2.5	2.3	−1.3	2.0	3.3	−5.3	−3.2	2.1	−8.0	−6.4	1.6

SI-SDR (dB) — ↑ is better

x1 Multi Level Model

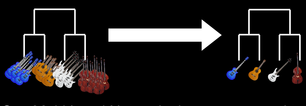
We observe that Multi Level Networks always match or beat Single Level Networks, despite Single Level Networks having **3x** as many learnable parameters! This implies that the Multi Level Nets are able to leverage hierarchical knowledge about the mixture.

Experiment 3:
Do we need labeled data for all leaf nodes?

Main Idea

Training hierarchical networks requires *lots* of data, especially for leaf nodes.

Can we train models with diminished access to ground truth leaf data?



Can Multi Level Networks leverage ground truth data at coarser hierarchy levels for finer grained separation?

Results

% → Percent of Data Removed
"type" → Whether **all** data was removed or just **leaf** data

	Data Reduction %	Data Reduction type	Levels All	Levels Level 3	Levels Level 2	Levels Level 1
SSS (Guitar)	0	-	3.8	3.5	4.0	4.0
	50	all	3.3	3.1	3.4	3.4
	50	leaf	3.5	3.3	3.6	3.6
	90	all	0.1	1.5	−0.7	−0.5
	90	leaf	3.6	3.4	3.7	3.7
Mix			−3.9	0.9	−5.9	−6.6

SI-SDR (dB) Improvement over Mix
↑ is better

Multi Level
Source Specific Separation

SSS Nets are able to retain almost all of their performance when 90% of leaf data is missing!

Multi Level
Query-by-Example

QBE Nets retain half of their performance with 90% leaf data missing. Amazingly, they still separate under harsh requirements!

	Data Reduction %	Data Reduction type	Levels All	Levels Level 3	Levels Level 2	Levels Level 1
QBE	0	-	2.3	3.3	2.1	1.6
	50	all	−1.5	−2.1	−1.4	−1.1
	50	leaf	2.2	3.4	2.1	1.1
	90	all	−1.8	−2.1	−1.8	−1.5
	90	leaf	1.9	3.1	1.7	0.8
Mix			−4.9	−1.3	−5.3	−8.0

SI-SDR (dB) Improvement over Mix
↑ is better

Thanks for Scrolling!

We are happy to take questions in the chat 😄

Audio samples are available at this link.