Weakly Supervised Source-Specific Sound Level Estimation in Noisy Soundscapes

WASPAA 2021




Aurora Cramer Mark Cartwright Fatemeh Pishdadian Juan Pablo Bello

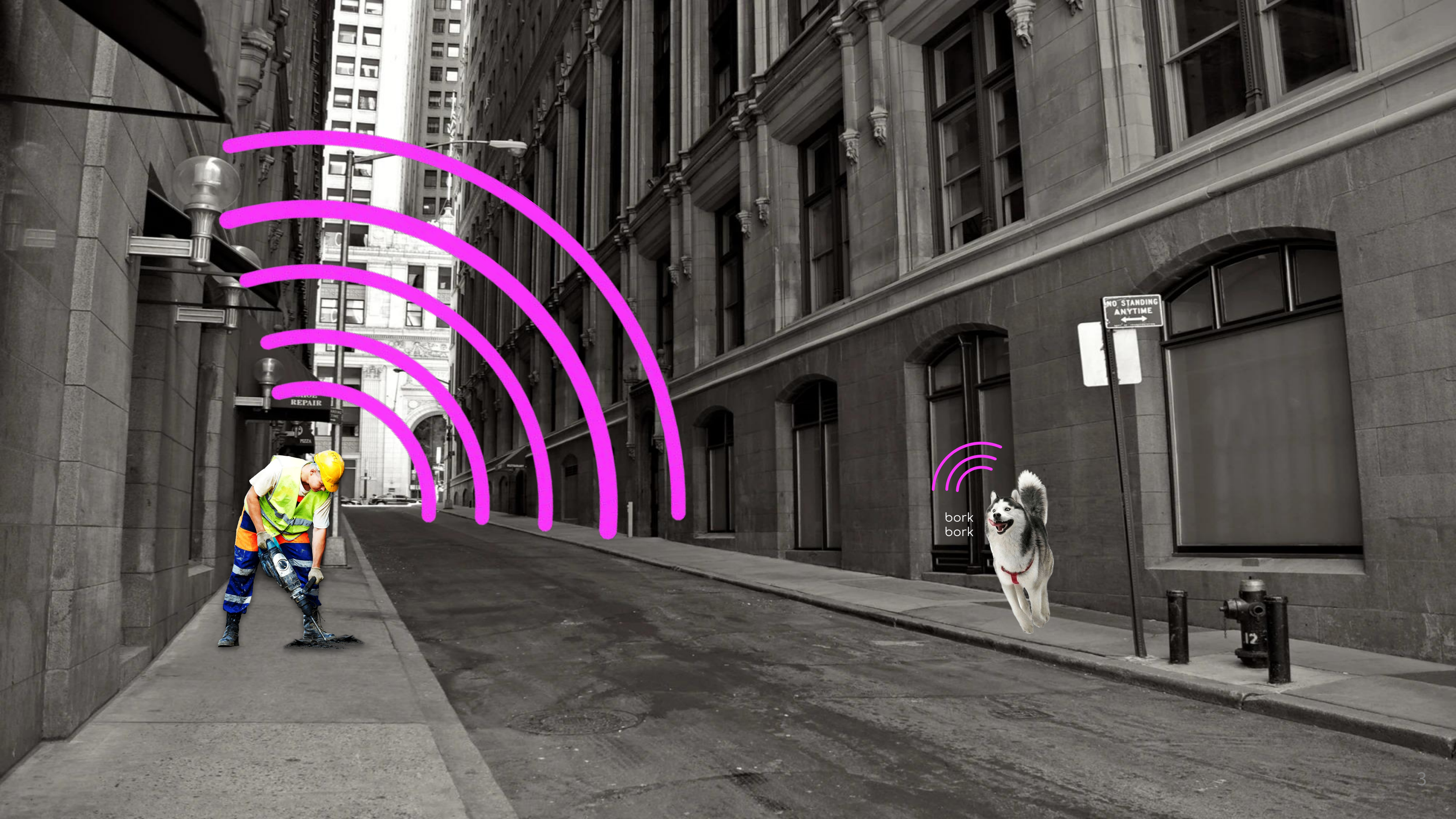
NYU TANDON SCHOOL OF ENGINEERING

MARL

NJIT New Jersey Institute of Technology

Northwestern University

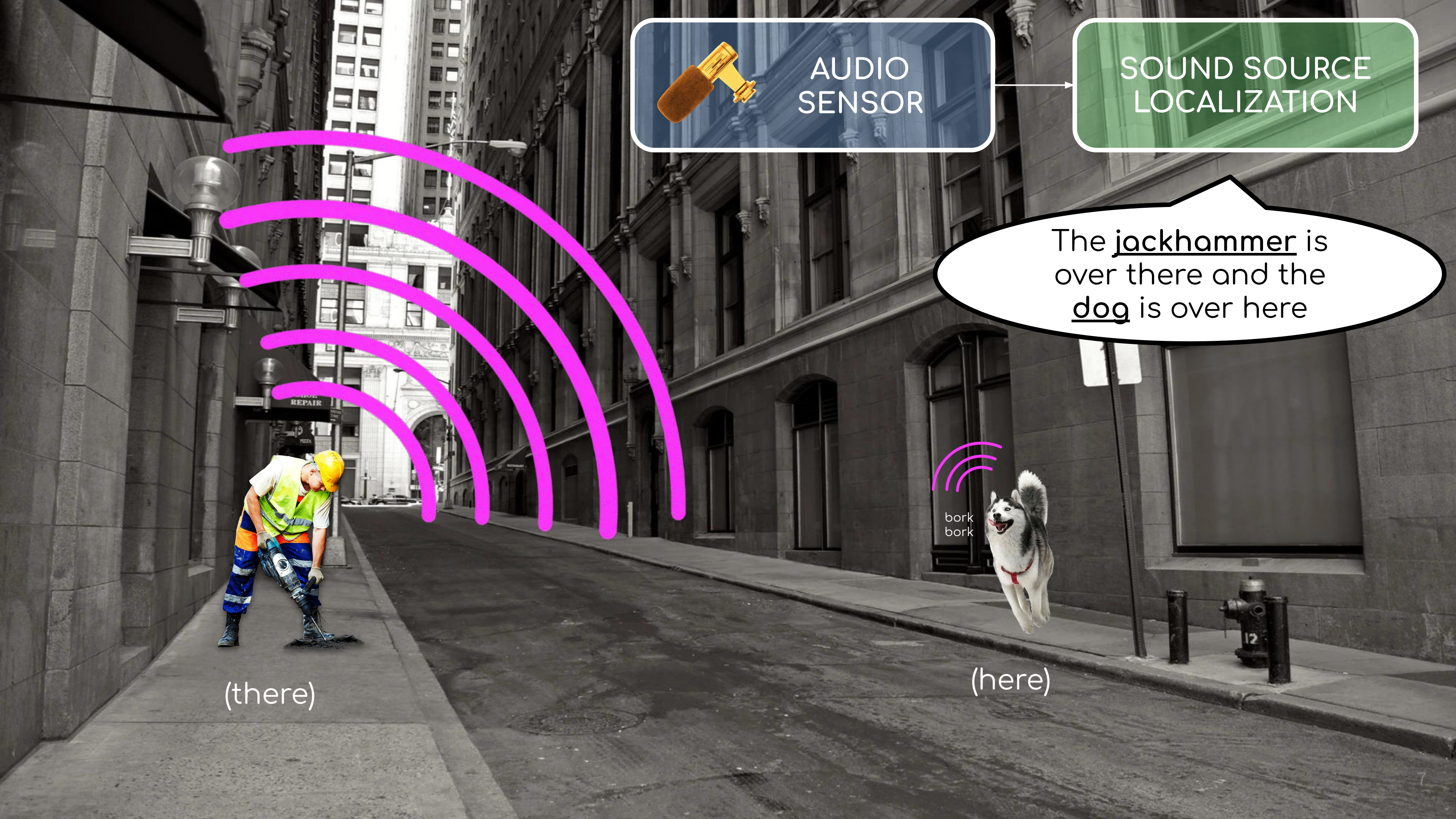
Motivation

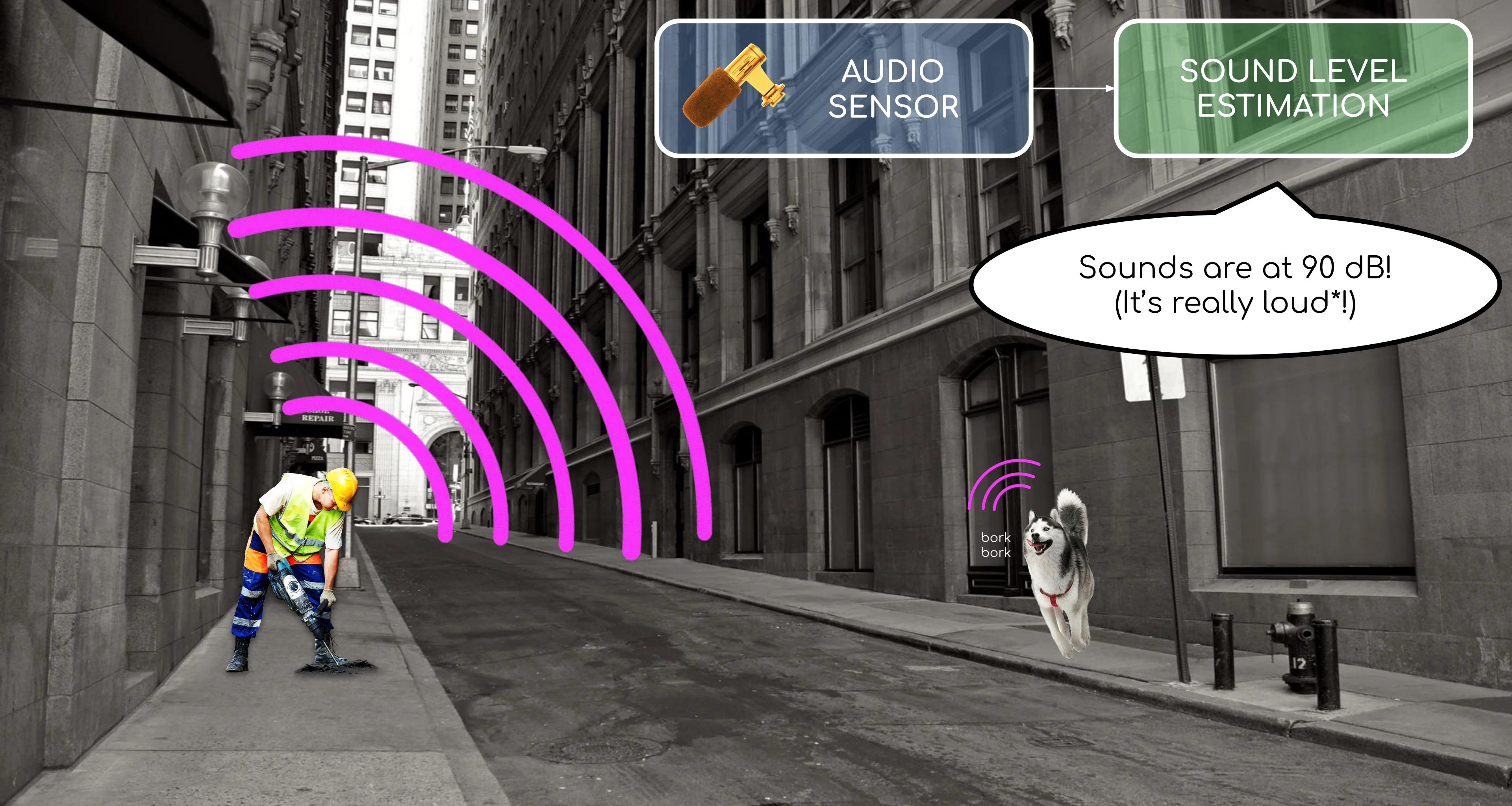

REPAIR
NO STANDING
ANYTIME
bork
bork
12


AUDIO
SENSOR
bork
bork
NO STANDING
ANYTIME
REPAIR


AUDIO SENSOR
MACHINE LISTENING
Q: What's happening?
bork bork
REPAIR
NO STANDING ANYTIME
12


AUDIO SENSOR
SOUND EVENT RECOGNITION
There's a jackhammer and a dog
bork bork


AUDIO SENSOR
SOUND SOURCE LOCALIZATION
The jackhammer is over there and the dog is over here
REPAIR
bork bork
(there)
(here)



* loudness ≠ energy, but they're related

AUDIO SENSOR

SOUND LEVEL ESTIMATION

Sounds are at 90 dB!
(It’s really loud*!)

- **Goal: characterize the energy* of an audio signal**
- **But what if we are interested in the sound level of a specific source?**

* loudness ≠ energy, but they’re related


AUDIO SENSOR
SOURCE-SPECIFIC SOUND LEVEL ESTIMATION
Jackhammer at 100 dB! (It's really loud*!)
Dog at 60dB! (They're a little loud*! (but *very* good))
Siren is -80 dB! (It's ~silent!)
⋮
bork bork
* loudness ≠ energy, but they're related

Why source-specific sound level estimation?

- **Urban noise pollution monitoring:** estimating the loudness of specific sound sources to aid in noise mapping and enforcement [1]
- **Intelligent audio production:** determine (relative) gain of instruments in audio mixes and inform automatic mixing systems that mimic audio engineers [2]
- **Source localization:** could also aid in distance estimation for sources in diverse settings like wildlife monitoring and sound awareness technology

[1] Gloaguen et al., "Road traffic sound level estimation from realistic urban sound mixtures by non-negative matrix factorization," Applied Acoustics, 2019.
[2] Ward et al., "Estimating the loudness balance of musical mixtures using audio source separation," WIMP, 2017.

The state of SSSLE

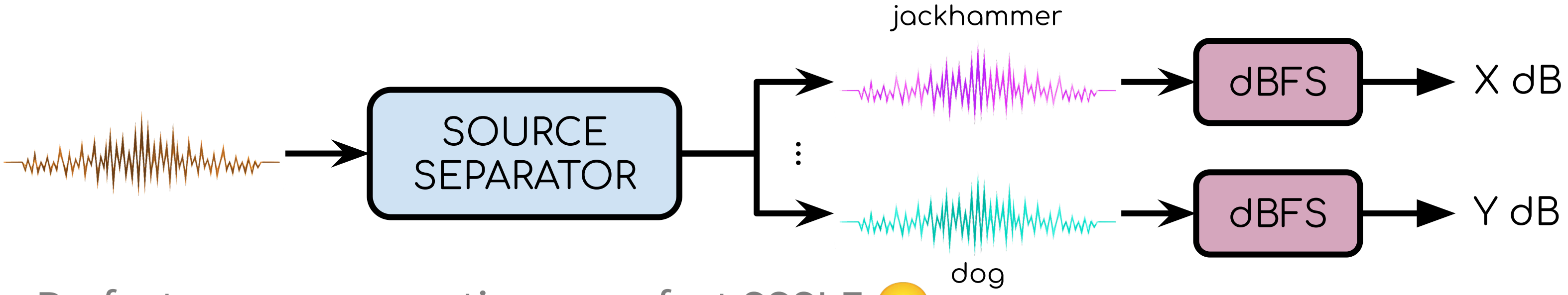
- **SSSLE has been understudied** compared to other machine listening tasks

- Most existing approaches **require access to isolated sources** which are hard to reliably acquire in realistic recording scenarios

- **Obtaining ground truth sound levels for sources is generally impractical or infeasible** in realistic settings

- No accounting for **background noise and out-of-vocabulary sources** that are generally present in recordings

What if we just use source separation?

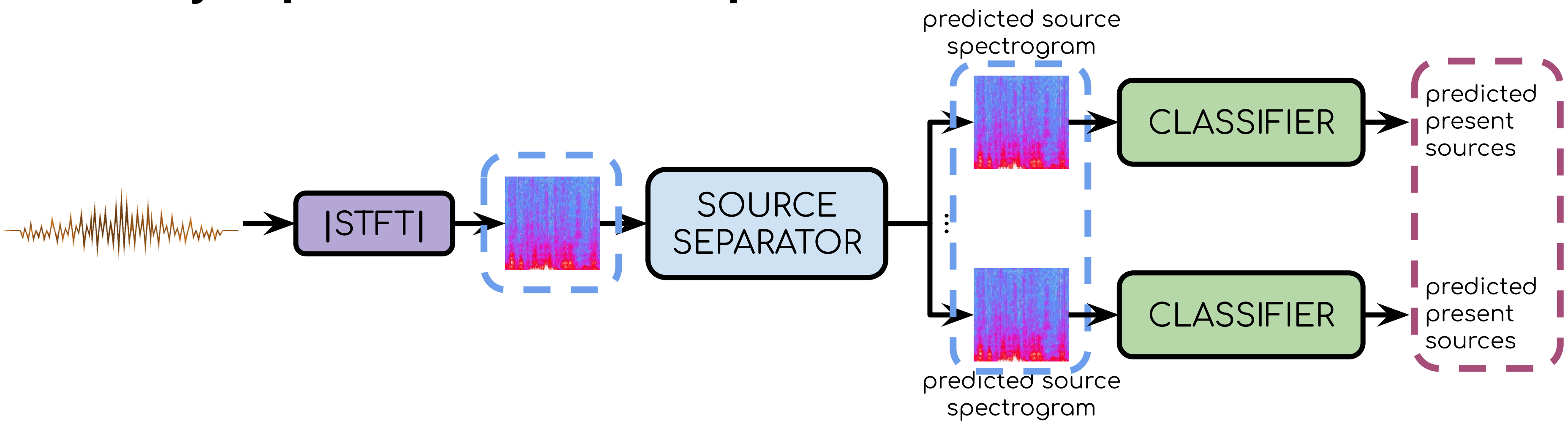


- Perfect source separation → perfect SSSLE 🙂
- Often impractical or infeasible to effectively train a fully-supervised deep source separation model for the target application 😦
- Recent methods have been developed to require less supervision for deep source separation 😊
    - Weakly supervised: joint separation and classification **(Pishdadian et al. '20)**[3], (Kong et al. '19, '20)[4, 5]
    - Unsupervised: MixIT [6]

[3] Pishdadian, G.Wichern, and J. Le Roux, "Finding strength in weakness: Learning to separate sounds with weak supervision," TASLP, 2020.
[4] Kong et al "Sound event detection and time–frequency segmentation from weakly labelled data," TASLP, 2019.
[5] Kong et al., "Source separation with weakly labelled data: An approach to computational auditory scene analysis," ICASSP, 2020
[6] Wisdom et al., "Unsupervised speech separation using mixtures of mixtures," ICML 2020 Workshop on Self-supervision in Audio and Speech, 2020.

Methods

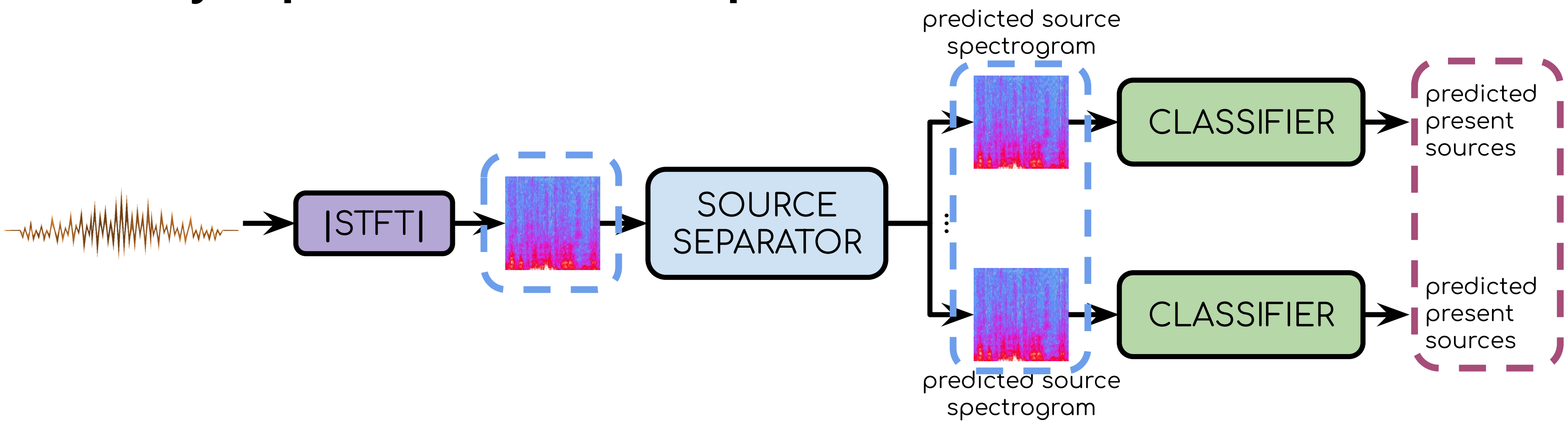
Weakly supervised source separation (Pishdadian et al. 2020)



- **Energy consistency:** energy (in each TF-bin) from active sources should sum to mixture
- **Classifier critic:** separated (true) active sources should contain only that source type, separated (true) inactive sources should not contain any relevant sources

- Training a reasonable source-separation is possible with only clip-level labels!

Weakly supervised source separation (Pishdadian et al. 2020)



- **Energy consistency:** $\frac{1}{TF}\|\mathbf{M}_E \odot \mathbf{R}_{\text{active}}\|_1 + \frac{1}{TF}\|\mathbf{M}_E \odot \mathbf{R}_{\text{inactive}}\|_1$

  non-silence mask

  residual between mixture and sum of active sources

  residual between silence and sum of inactive sources

- **Classifier critic:** $\mathcal{L}_{\text{cls-mix}} + \sum_i \mathcal{L}_{\text{cls-mix},i}$

  BCE: mixture contains all true active sources

  BCE: separated active source contain only that source; separated inactive sources contain no (known) sources

Remaining concerns:

1. We are training the model for source separation, but we really care about SSSLE!

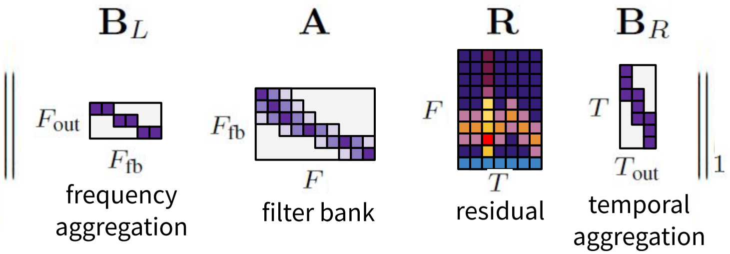
2. We still need to account for background noise and out-of-vocabulary sources!

Our work attempts to address these two concerns

Connecting source separation to SSSLE

- Use the relationship between source separation and SSSLE to bridge the gap
- Observations:
  - Sound level estimation can be formulated as enforcing ***global*** energy consistency
  - Energy consistency terms are of the form: $\frac{1}{TF}\|\mathbf{R}\|_1$
- Idea: generalize these expressions

$$\frac{1}{TF_{\text{fb}}}\|h_\Phi(\mathbf{R})\|_1 = \frac{1}{TF_{\text{fb}}}\left\| \mathbf{B}_L \, \mathbf{A} \, \mathbf{R} \, \mathbf{B}_R \right\|_1$$

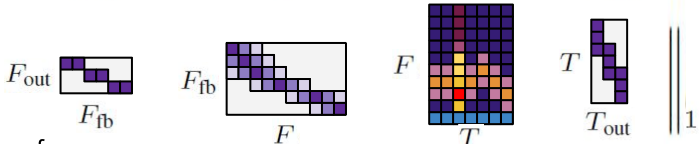


- Different choices of $\Phi = (\mathbf{A}, \mathbf{B}_L, \mathbf{B}_R)$ apply energy consistency at different time-frequency resolutions

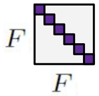
Parameterizing energy consistency

$$\frac{1}{TF_{\text{fb}}}\|h_\Phi(\mathbf{R})\|_1 = \frac{1}{TF_{\text{fb}}}\left\| \mathbf{B}_L \, \mathbf{A} \, \mathbf{R} \, \mathbf{B}_R \right\|_1$$

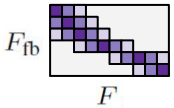
$(\mathbf{A}, \mathbf{B}_L, \mathbf{B}_R)$



$\mathbf{A} = \mathbf{A}_{\text{linear}} =$



$\mathbf{A} = \mathbf{A}_{\text{mel}} =$



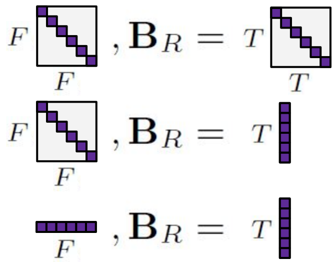
triangular mel frequency filter bank, 40 bands

✖

time-frequency energy consistency: $\mathbf{B}_L =$, $\mathbf{B}_R =$

spectrum energy consistency: $\mathbf{B}_L =$, $\mathbf{B}_R =$

global energy consistency: $\mathbf{B}_L =$, $\mathbf{B}_R =$



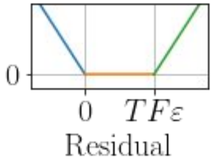
$\mathcal{P} \longrightarrow$ We apply energy consistencies at multiple time-frequency resolutions!

$$\frac{1}{|\mathcal{P}|}\sum_{\Phi\in\mathcal{P}} \frac{1}{TF_{\text{fb}}}\|h_\Phi(\mathbf{R})\|_1$$

Accounting for background

- Sum of sources no longer adds up to the mixture, but what if it almost adds up to the mixture?
- Idea: Introduce an asymmetric margin to the **active** energy consistency loss to allow for background and out-of-vocabulary sources

$$\|\mathbf{R}\|_1^{(\text{asym},T,F,\varepsilon)} = \left[\left\|[\mathbf{R}]_+\right\|_1 - TF\varepsilon\right]_+ + \left\|[-\mathbf{R}]_+\right\|_1$$



asymmetry allows for underestimating mixture energy
while penalizing overestimating mixture energy

- Ensure residual "background" signal does not contain any in-vocabulary sources

$$\hat{\mathbf{M}}_{\text{bkgr}} = \left[1 - \textstyle\sum_i \hat{\mathbf{M}}_i\right]_+$$

background mask = complement of estimated source masks

$$\mathcal{L}_{\text{cls-bkgr}} = \sum_i H\left(0, \hat{y}_i^{(\text{bkgr})}\right)$$

classifier should predict all zeros for background

Putting it all together!

set of energy consistency configurations

relative importance hyperparameter

$$\mathcal{L}^{\boxed{\mathcal{P}}}_{\text{weak, sssle}} = \frac{\boxed{\alpha}}{|\mathcal{P}|} \sum_{\Phi \in \mathcal{P}} \underbrace{\boxed{\mathcal{L}^{\Phi}_{\text{mix, sssle}}}}_{\text{energy consistency loss}} + \underbrace{\boxed{\mathcal{L}_{\text{cls, sssle}}}}_{\text{classification loss}}$$

$$\mathcal{L}^{\Phi}_{\text{mix, sssle}} = \frac{1}{TF_{\text{fb}}} \| \boxed{h_\Phi}(\mathbf{M}_E \odot \mathbf{R}_{\text{active}}) \|_1^{\boxed{(\text{asym}, T, F_{\text{fb}}, \varepsilon)}} + \frac{1}{TF_{\text{fb}}} \| \boxed{h_\Phi}(\mathbf{M}_E \odot \mathbf{R}_{\text{inactive}}) \|_1$$

asymmetric margin

generalized energy consistency

$$\mathcal{L}_{\text{cls, sssle}} = \mathcal{L}_{\text{cls-mix}} + \sum_i \mathcal{L}_{\text{cls-mix}, i} + \boxed{\beta} \boxed{\mathcal{L}_{\text{cls-bkgr}}}$$

background classification loss

hyperparameter enabling/disabling background classification

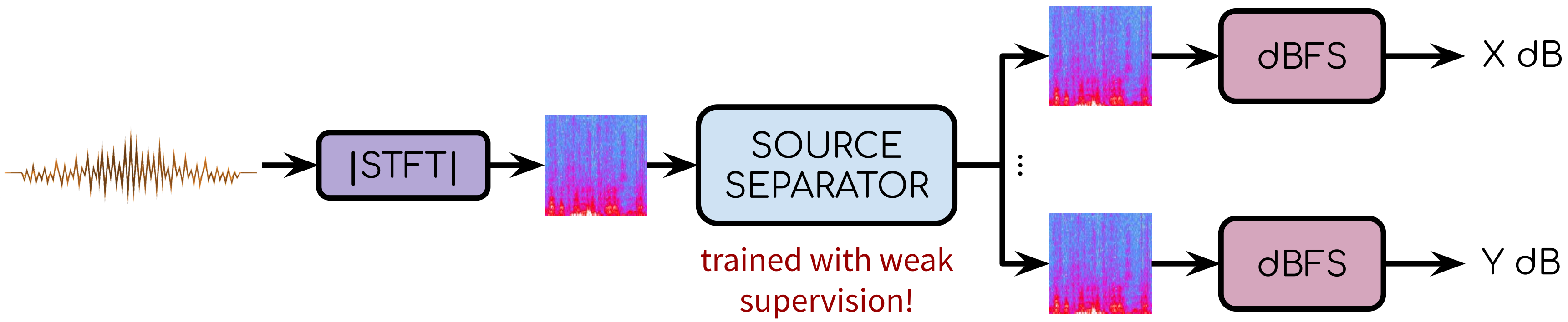
<u>hyperparameter choices</u>

α = average margin of training set

$\mathcal{P}$ = {mel filterbank} x {TF, energy, global consistency}

β = 1 (enabled)

Estimating sound levels



Experiments

Data

- Start with synthetic dataset used by Pishdadian et al.
    - 4 second mixtures (@ 16kHz) w/ sources sampled from subset of UrbanSound8K [7]
    - train/valid/test: 50k/10k/10k mixtures

- Add backgrounds noise from city soundscapes recordings obtained from an urban noise monitoring sensor network (SONYC)
    - SONYC-Backgrounds: https://doi.org/10.5281/zenodo.5129078

- Create datasets from mixtures and backgrounds at **-50**/**-20**/**0 dB LUFS** (**weak**/**moderate**/**strong** background)**,** as well as and **no background**

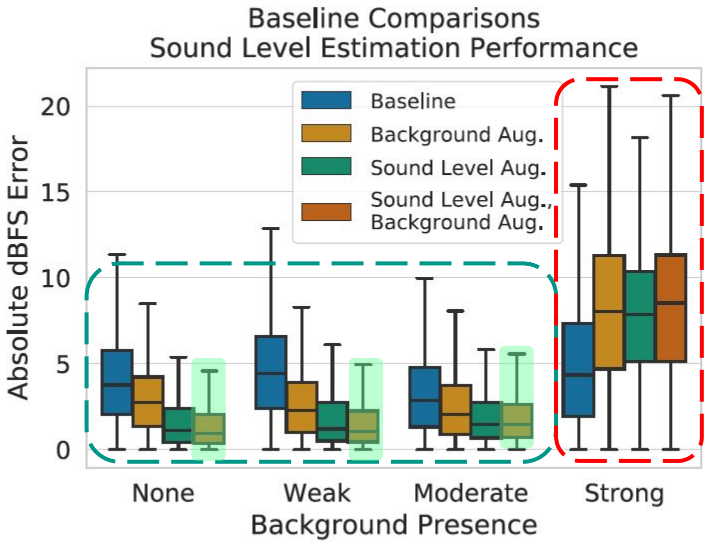
[6] Salamon et al., "A dataset and taxonomy for urban sound research," ACM Multimedia, 2014.

Evaluation

- Metric: **absolute dBFS error**: characterizes the sound level estimation error
- Compare with:
  - Weakly supervised source separation (no augmentations)
  - Only energy consistency augmentations
  - Only background augmentations

Baseline Comparison

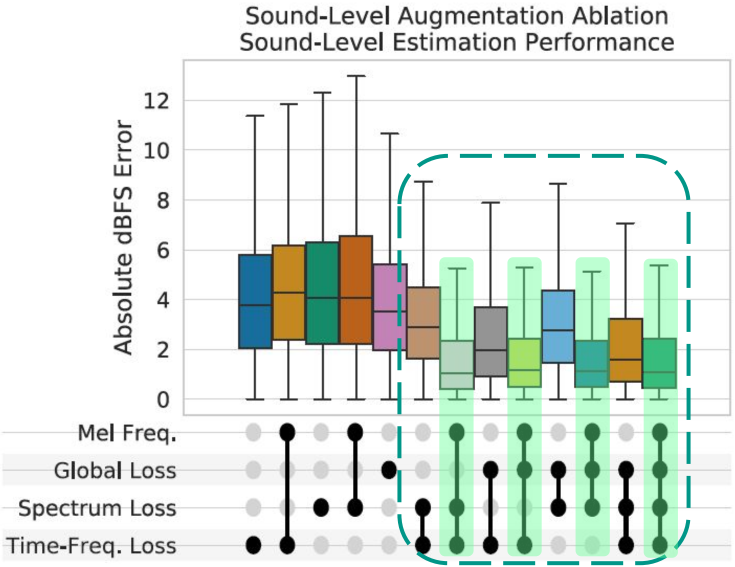
- Both augmentations yield best improvements in up to moderate background
- However, strong background breaks energy margin assumptions



Ablation studies

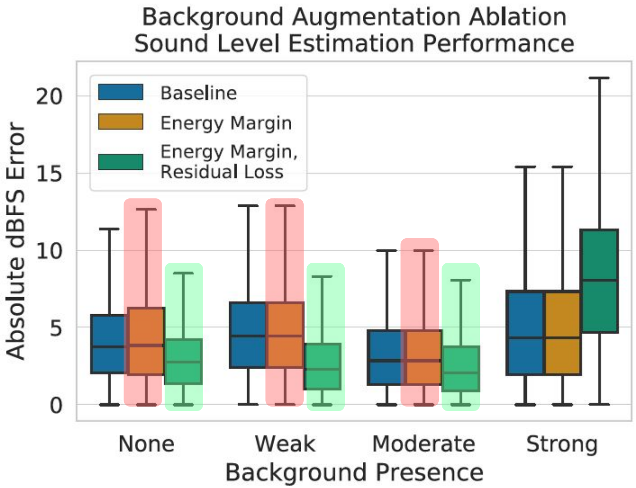
Ablation study: sound-level augmentations

- Multiple time frequency resolutions improve sound level estimation
- Best performance with at least 2 time-frequency resolutions and mel scale

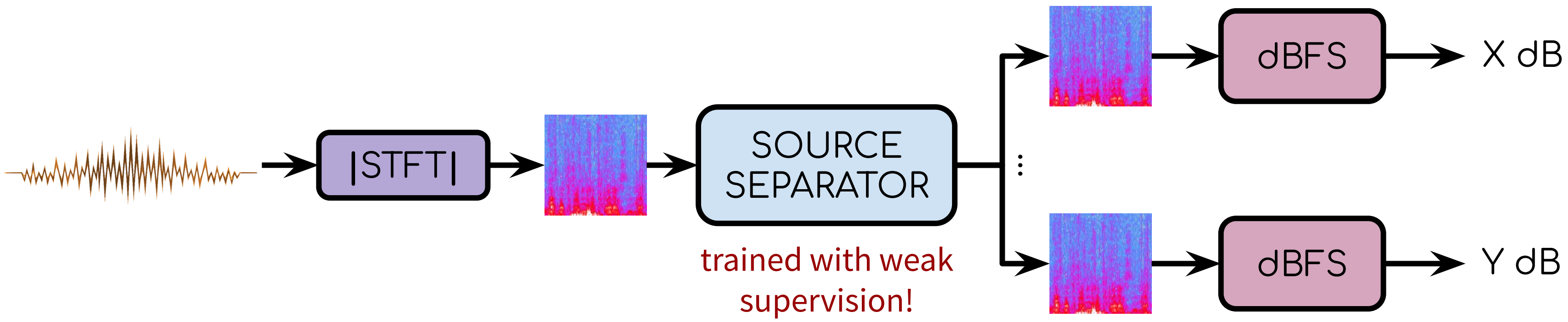


Ablation study: background augmentations

- Both the energy margin and residual background classification loss improve performance in up to moderate background
- Background classification is important for the margin to be effective



Estimating sound levels



Future work

- Addressing fixed margin
- Better background modeling
- Open question: how to evaluate SSSLE for real recordings?

In summary:

- We extended weakly supervised source separation to **more directly address sound level estimation** and to **account for background, improving SSSLE performance in up to moderate background conditions**
- **New dataset**: SONYC-Backgrounds (https://doi.org/10.5281/zenodo.5129078)
- SSSLE models can be trained from **only clip-level class presence annotations**
- **SSSLE is possible in practical scenarios!**

https://github.com/sonyc-project/weakly-supervised-sssle

Thank you!