Learning To Separate Sounds From Weakly Labeled Scenes

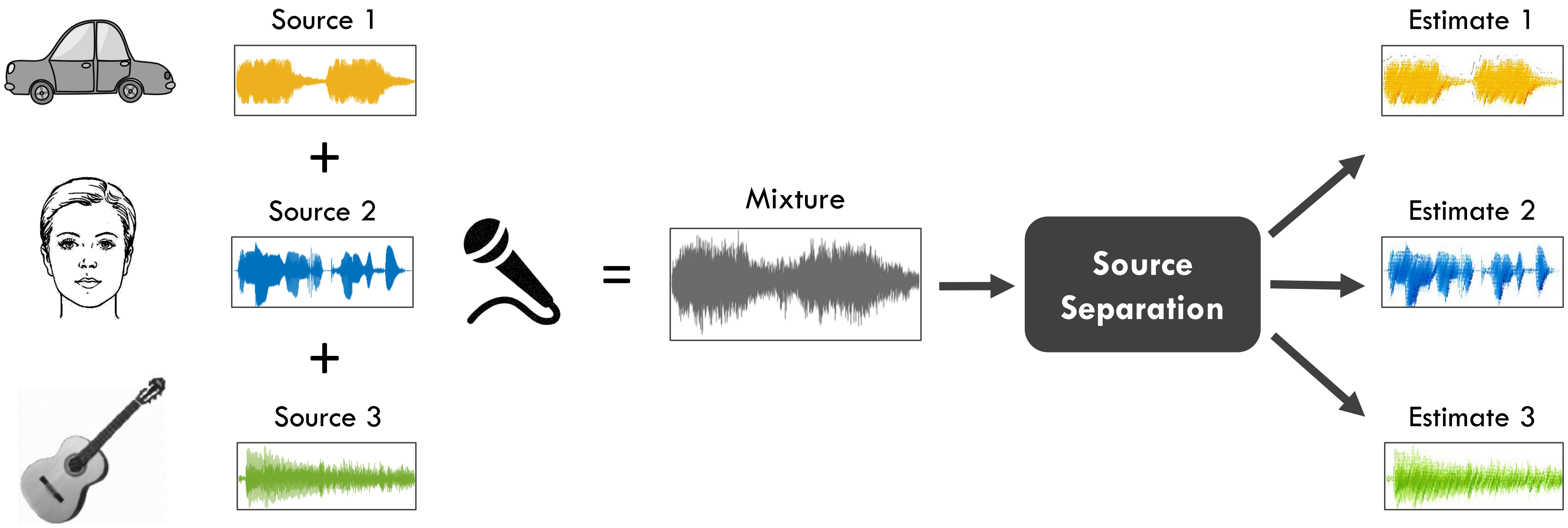
Fatemeh Pishdadian, Gordon Wichern, Jonathan Le Roux

ICASSP 2020

MITSUBISHI ELECTRIC RESEARCH LABORATORIES (MERL)
Cambridge, Massachusetts, USA
http://www.merl.com

Single-channel Audio source separation

- Isolating individual sounds in a complex auditory scene

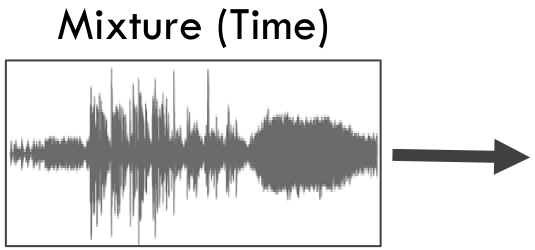


Masking-based audio source separation

- A common approach: time-frequency mask inference

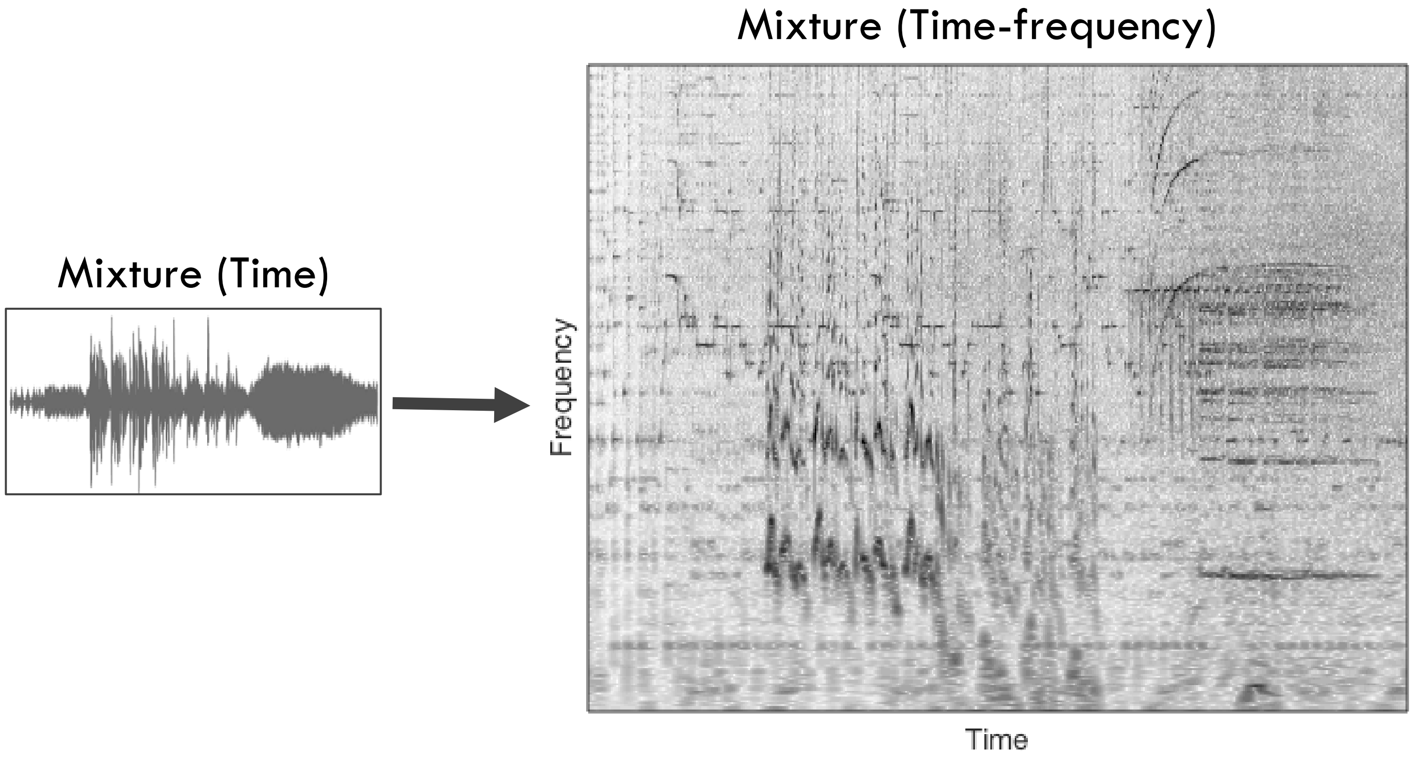
Masking-based audio source separation

- A common approach: time-frequency mask inference



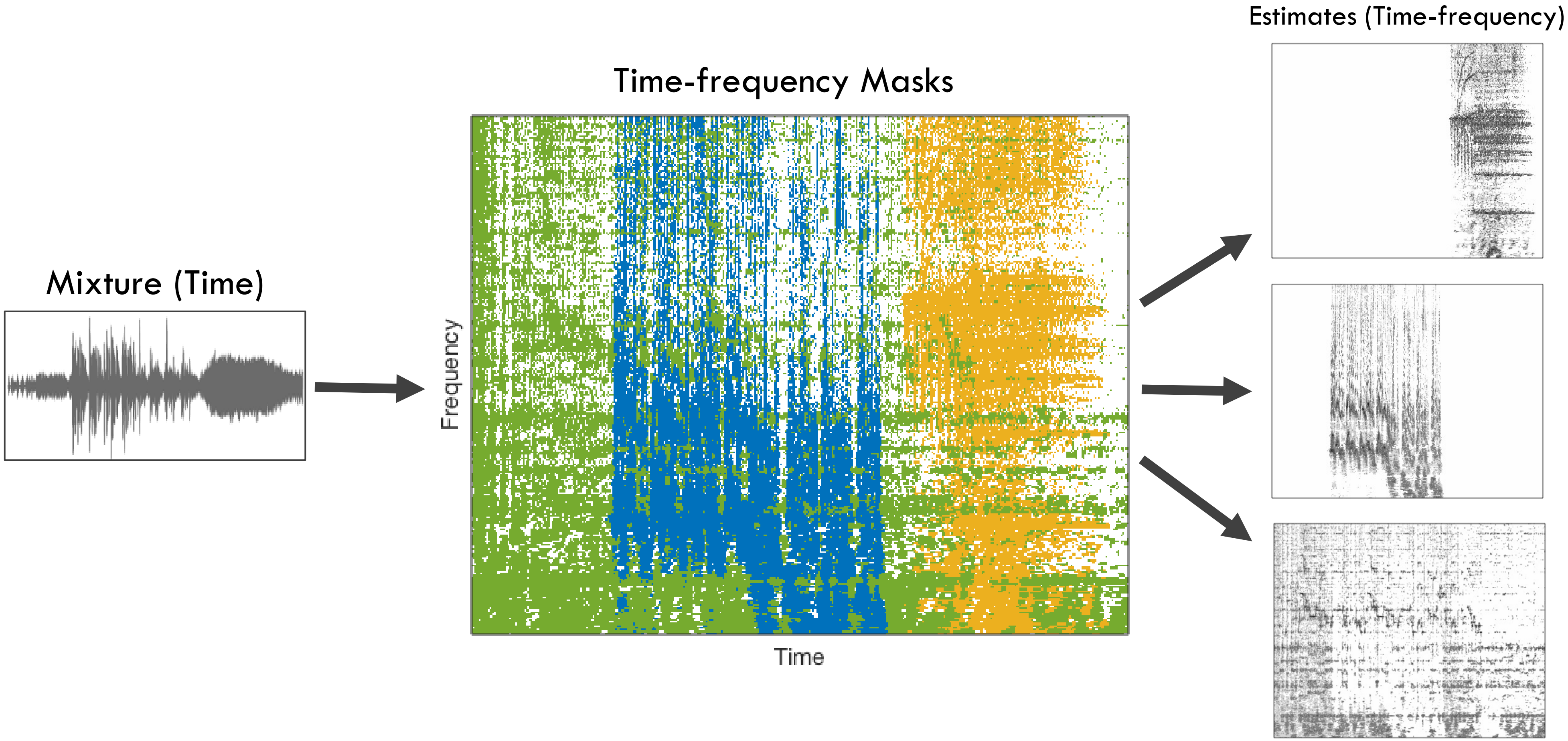
Masking-based audio source separation

- A common approach: time-frequency mask inference



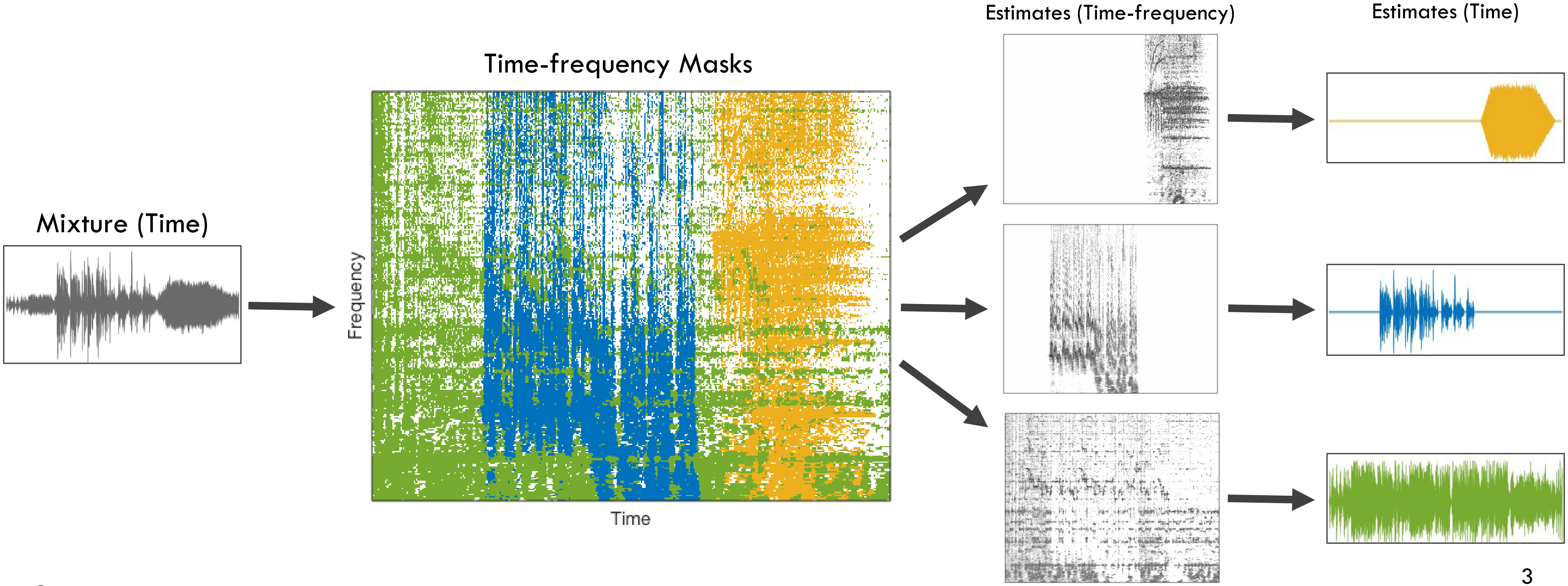
Masking-based audio source separation

- A common approach: time-frequency mask inference

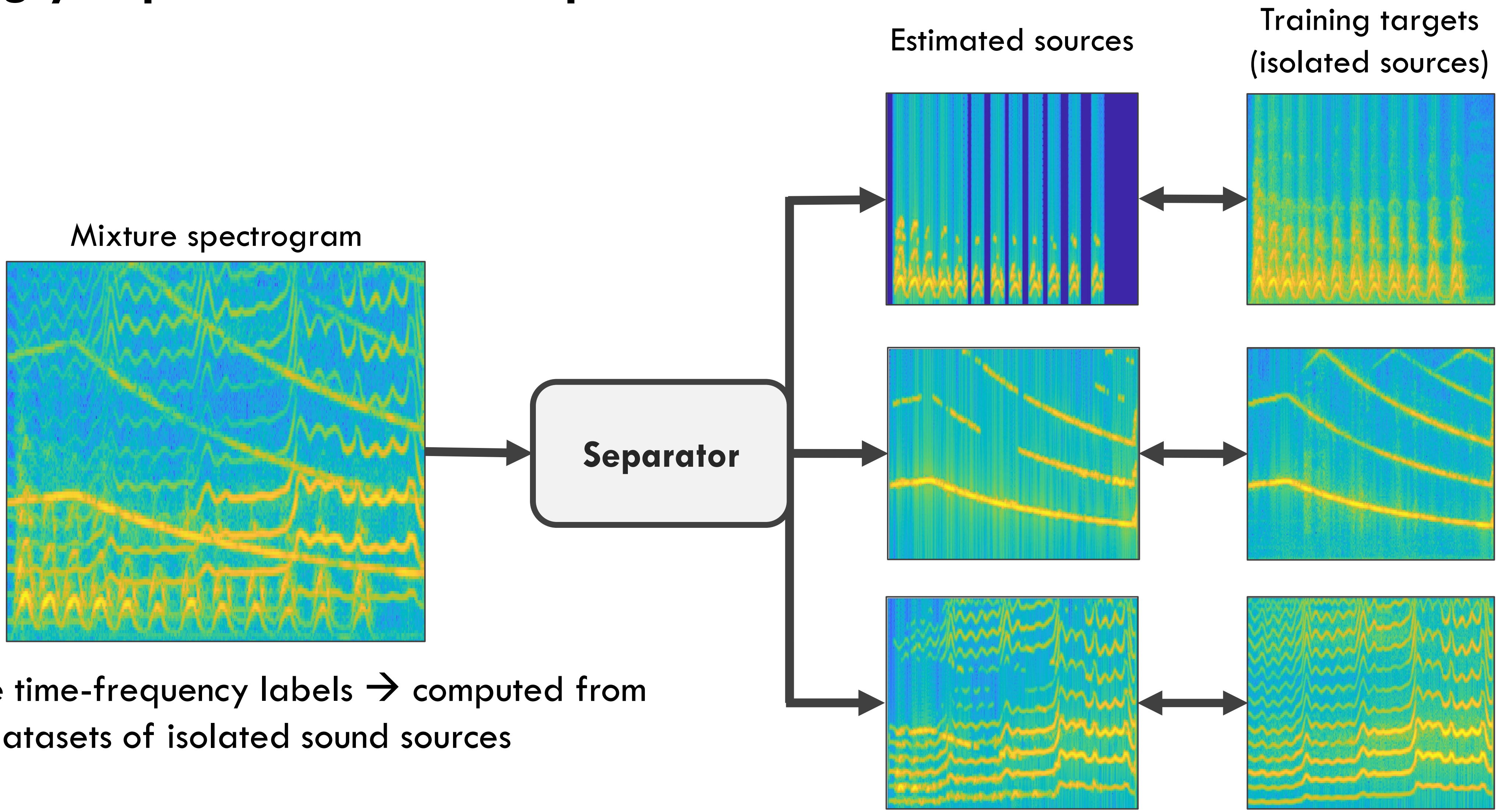


Masking-based audio source separation

- A common approach: time-frequency mask inference



Strongly supervised source separation



Require time-frequency labels → computed from large datasets of isolated sound sources

Strongly supervised source separation

- Deep learning methods
  - Good performance in speech/music source separation
  - Require time-frequency labels → computed from large datasets of isolated sound sources
- Obtaining isolated sound sources
  - Expensive
  - Require complicated recording setups
  - Not practical in some situations → difficult to record sounds in isolation e.g., isolating natural sounds or the sound of a machine part when the machine is running

Strongly supervised source separation

- Deep learning methods
  - Good performance in speech/music source separation
  - Require time-frequency labels → computed from large datasets of isolated sound sources

- Obtaining isolated sound sources
  - Expensive
  - Require complicated recording setups
  - Not practical in some situations → difficult to record sounds in isolation e.g., isolating natural sounds or the sound of a machine part when the machine is running

Our approach

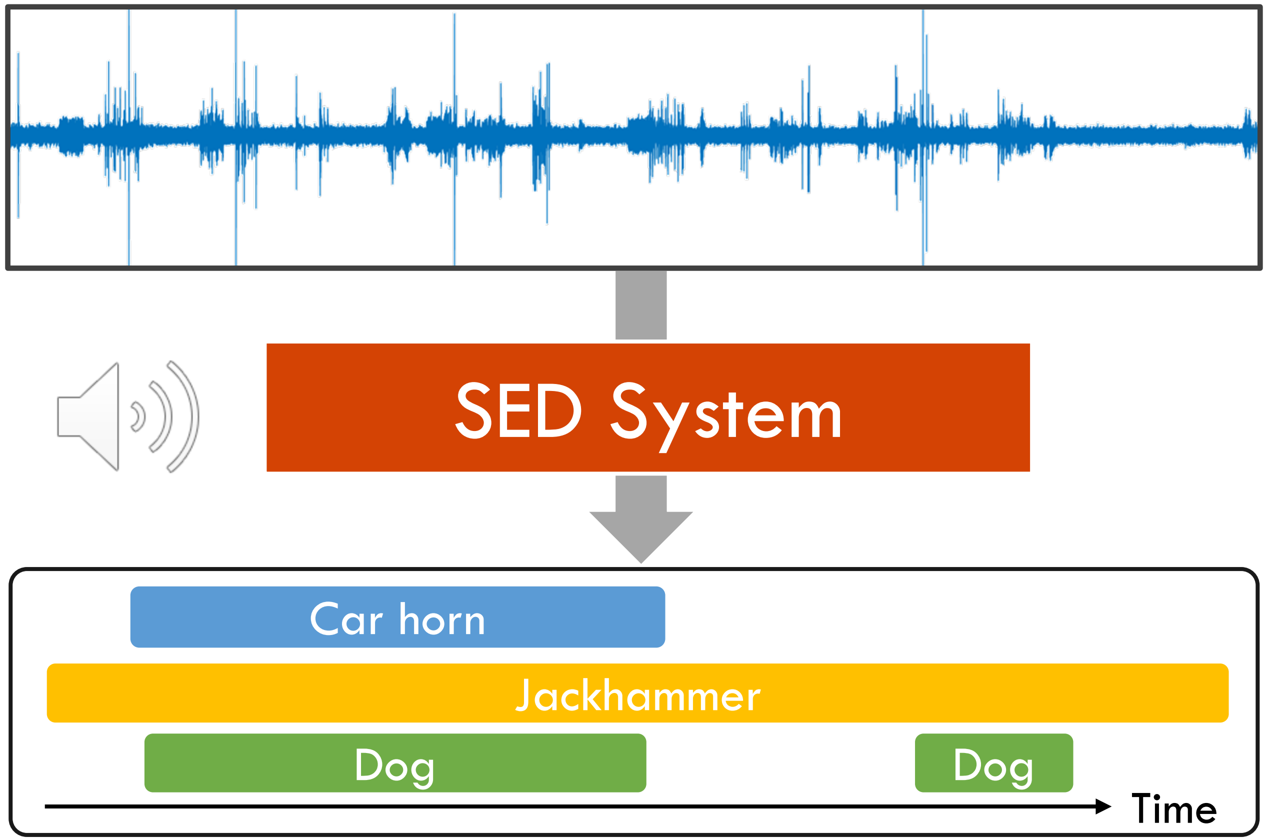
- Train a source separation system with labels that are easier to collect in realistic conditions, e.g., information on each source's activity over time

- Predicting such information is typically the goal of a **Sound Event Detection (SED)** system → we hope to use such a system as a bridge

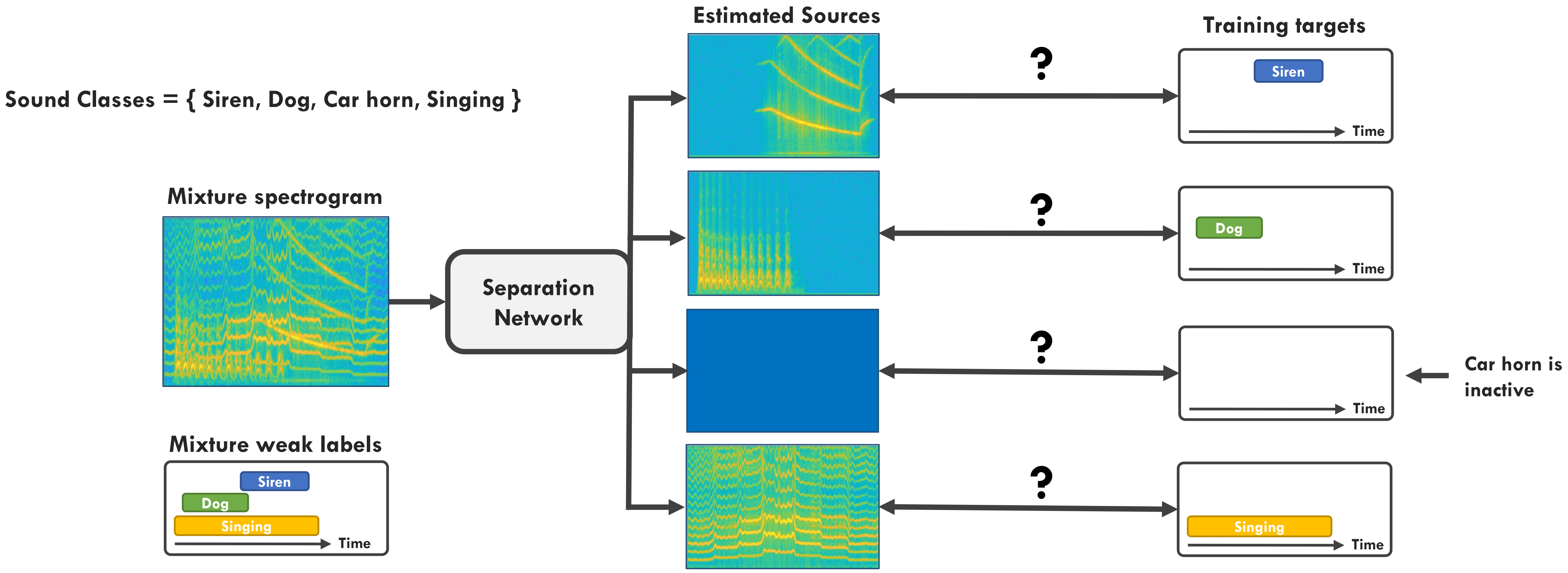
Sound event detection

- Sound Event Detection (SED) system
  - Predicts start and end time of each event
  - Classifies event into predefined categories

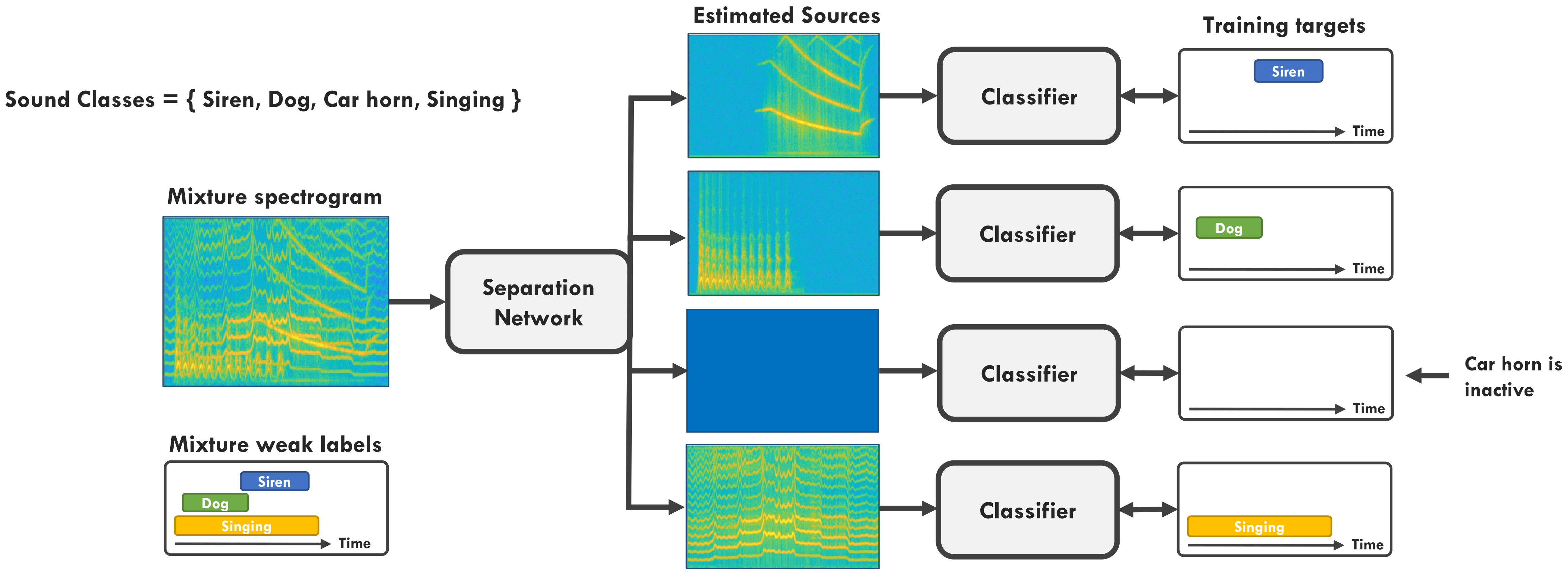
- Typical SED system
  1. Feature extraction
  2. Classification



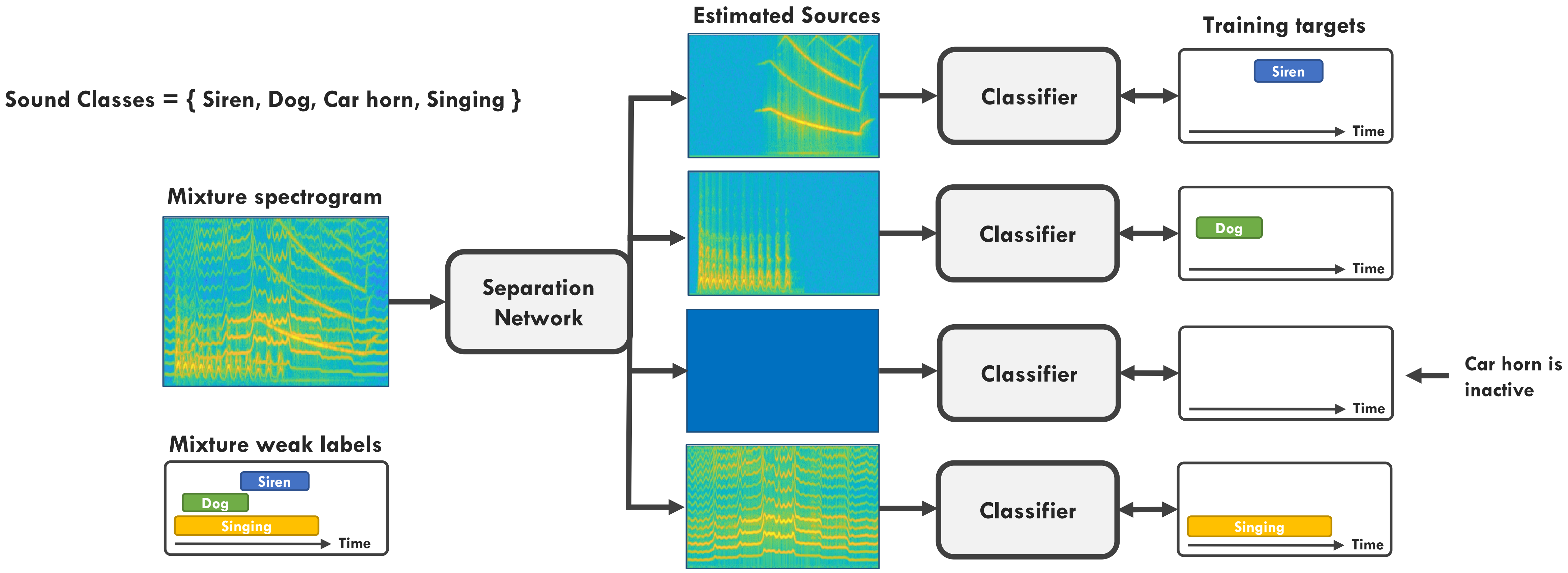
Frame-level weakly supervised source separation



Frame-level weakly supervised source separation

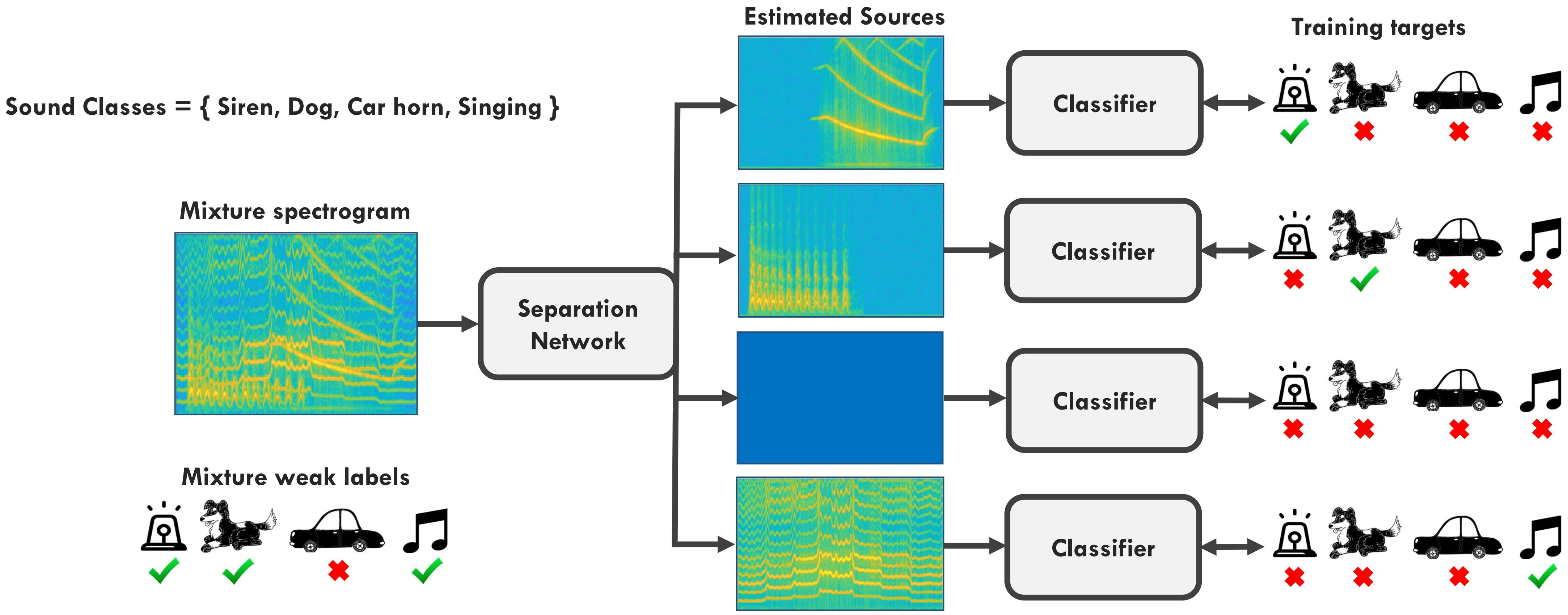


Frame-level weakly supervised source separation



- Training Objective: A pre-trained SED classifier should find only a single source at correct times in the estimated source spectrogram
- Only **time periods when sources are active** required for training, **not** isolated sources

Clip-level weakly supervised source separation



- Training Objective: A pre-trained sound event detection classifier should find only a single source in the estimated source spectrogram
- Only information on **presence or absence of sources within a clip** is required for training, **not** isolated sources

Classification objective

- Classification loss for mixture **frame** : τ

$$\mathcal{L}_{\text{f-class}}(\boldsymbol{X}, \tau) = \sum_{i=1}^{n} W_{i,\tau} H(l_{i,\tau}, p_{i,\tau}(\boldsymbol{X}))$$

cross-entropy loss function

$$H(l,p) = -l\log(p) - (1-l)\log(1-p)$$

Classification objective

- Classification loss for mixture **frame** : $\mathcal{T}$

$$\mathcal{L}_{\text{f-class}}(\boldsymbol{X}, \tau) = \sum_{i=1}^{n} W_{i,\tau} H(l_{i,\tau}, p_{i,\tau}(\boldsymbol{X}))$$

cross-entropy loss function

$$H(l,p) = -l\log(p) - (1-l)\log(1-p)$$

- Class activity priors:

$$W_{i,\tau} = \begin{cases} \gamma_i^{-1} & i \in \mathcal{A}_\tau, \\ (1-\gamma_i)^{-1} & i \notin \mathcal{A}_\tau, \end{cases}$$

Classification objective

- Classification loss for mixture **frame** : τ

$$\mathcal{L}_{\text{f-class}}(\boldsymbol{X},\tau)=\sum_{i=1}^{n} W_{i,\tau} H(l_{i,\tau},p_{i,\tau}(\boldsymbol{X}))$$

cross-entropy loss function

$$H(l,p)=-l\log(p)-(1-l)\log(1-p)$$

- Class activity priors:

$$W_{i,\tau}=\begin{cases}\gamma_i^{-1} & i\in\mathcal{A}_\tau,\\ (1-\gamma_i)^{-1} & i\notin\mathcal{A}_\tau,\end{cases}$$

prior probability for the activation of the i-th source

Classification objective

- Classification loss for mixture **frame** : τ

$$\mathcal{L}_{\text{f-class}}(\boldsymbol{X}, \tau) = \sum_{i=1}^{n} W_{i,\tau} H(l_{i,\tau}, p_{i,\tau}(\boldsymbol{X}))$$

cross-entropy loss function

$$H(l, p) = -l \log(p) - (1 - l) \log(1 - p)$$

- Class activity priors:

$$W_{i,\tau} = \begin{cases} \gamma_i^{-1} & i \in \mathcal{A}_\tau, \\ (1 - \gamma_i)^{-1} & i \notin \mathcal{A}_\tau, \end{cases}$$

prior probability for the activation of the i-th source

set of active source indices at frame τ

Using the classification loss to train the separator

- Classification loss for the i-th estimated source at **frame** : τ

$$\mathcal{L}_{\text{f-class}}(\hat{\boldsymbol{S}}_i, \tau) = W_{i,\tau} H(l_{i,\tau}, p_{i,\tau}(\hat{\boldsymbol{S}}_i)) + \sum_{j \neq i} W_{j,\tau} H(0, p_{j,\tau}(\hat{\boldsymbol{S}}_i))$$

Using the classification loss to train the separator

- Classification loss for the i-th estimated source at **frame** : τ

$$\mathcal{L}_{\text{f-class}}(\hat{\boldsymbol{S}}_i, \tau) = W_{i,\tau} H(l_{i,\tau}, p_{i,\tau}(\hat{\boldsymbol{S}}_i)) + \sum_{j \neq i} W_{j,\tau} H(0, p_{j,\tau}(\hat{\boldsymbol{S}}_i))$$

activity of the i-th source should match the frame labels

Using the classification loss to train the separator

- Classification loss for the i-th estimated source at **frame** : τ

$$\mathcal{L}_{\text{f-class}}(\hat{\boldsymbol{S}}_i,\tau) = W_{i,\tau} H(l_{i,\tau}, p_{i,\tau}(\hat{\boldsymbol{S}}_i)) + \sum_{j\neq i} W_{j,\tau} H(0, p_{j,\tau}(\hat{\boldsymbol{S}}_i))$$

activity of the i-th source should match the frame labels

all sources except the i-th source should be inactive

Joint separation-classification objective

- Training with only the classification loss → the separator network only needs to isolate the TF features necessary for classification, not signal reconstruction

Joint separation-classification objective

- Training with only the classification loss → the separator network only needs to isolate the TF features necessary for classification, not signal reconstruction

- Adding a mixture loss forces the separator to produce masks that reconstruct sources.

$$\mathcal{L}_{\mathrm{mix}}(\tau) = \sum_{\omega} \left| X_{\omega,\tau} - \sum_{i \in \mathcal{A}_\tau} \hat{S}_{i,\omega,\tau} \right| + \sum_{\omega} \sum_{i \notin \mathcal{A}_\tau} \left| \hat{S}_{i,\omega,\tau} \right|$$

Joint separation-classification objective

- Training with only the classification loss → the separator network only needs to isolate the TF features necessary for classification, not signal reconstruction

- Adding a mixture loss forces the separator to produce masks that reconstruct sources.

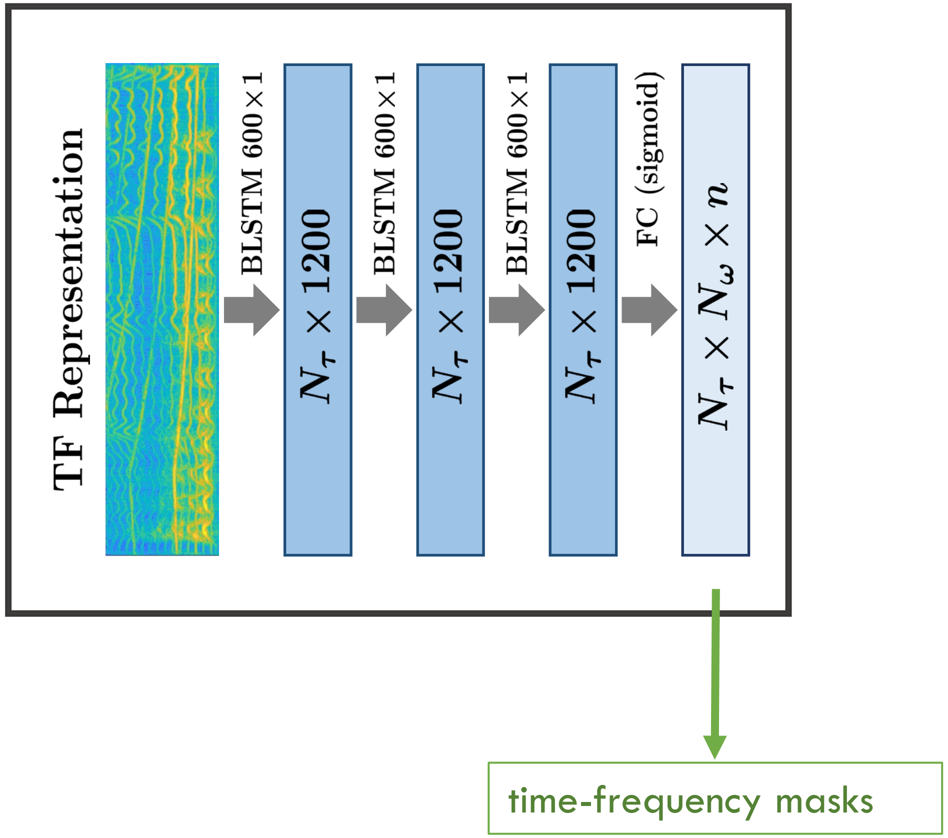
$$\mathcal{L}_{\text{mix}}(\tau) = \sum_{\omega} \left| X_{\omega,\tau} - \sum_{i \in \mathcal{A}_\tau} \hat{S}_{i,\omega,\tau} \right| + \sum_{\omega} \sum_{i \notin \mathcal{A}_\tau} \left| \hat{S}_{i,\omega,\tau} \right|$$

- Total loss for separation training: weighted sum of classification and mixture loss

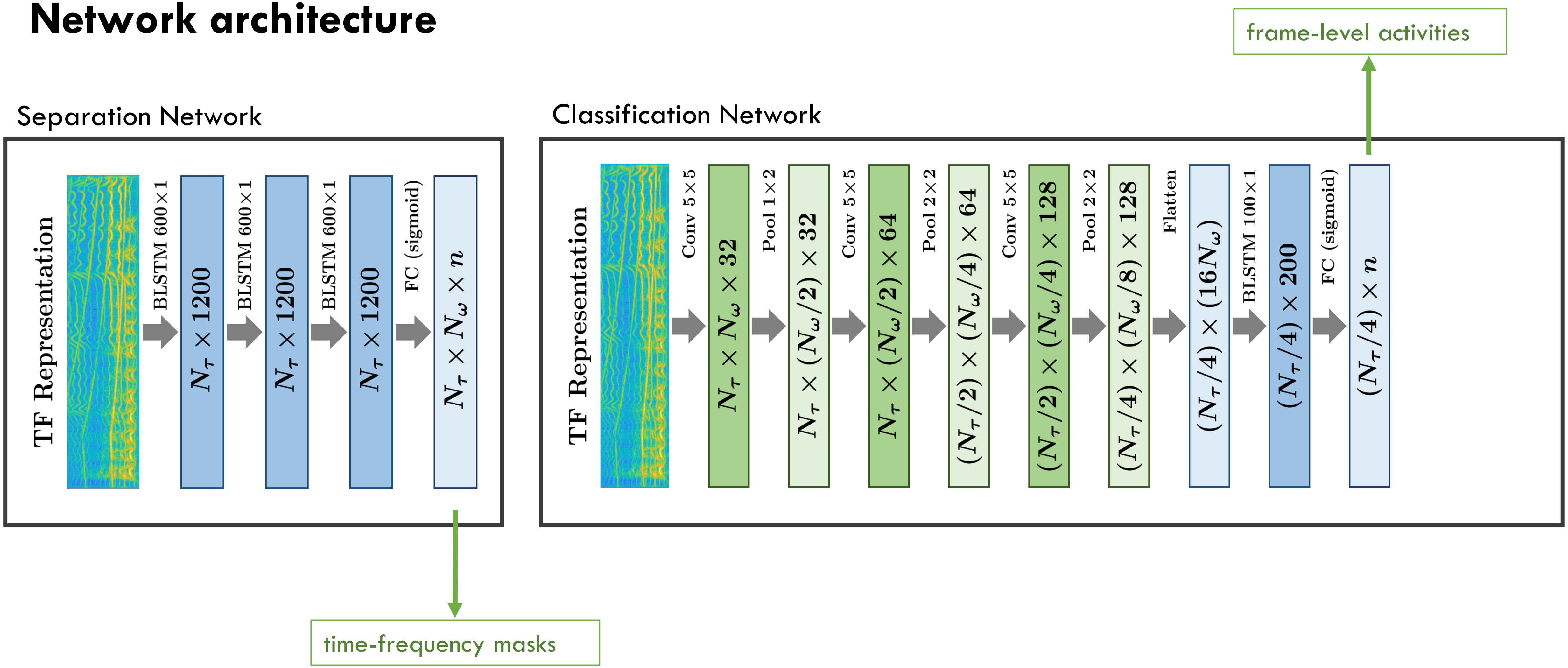
$$\mathcal{L}_{\text{f}} = \sum_{\tau,i} \mathcal{L}_{\text{f-class}}(\hat{\boldsymbol{S}}_i, \tau) + \alpha \sum_{\tau} \mathcal{L}_{\text{mix}}(\tau)$$

Network architecture

Separation Network



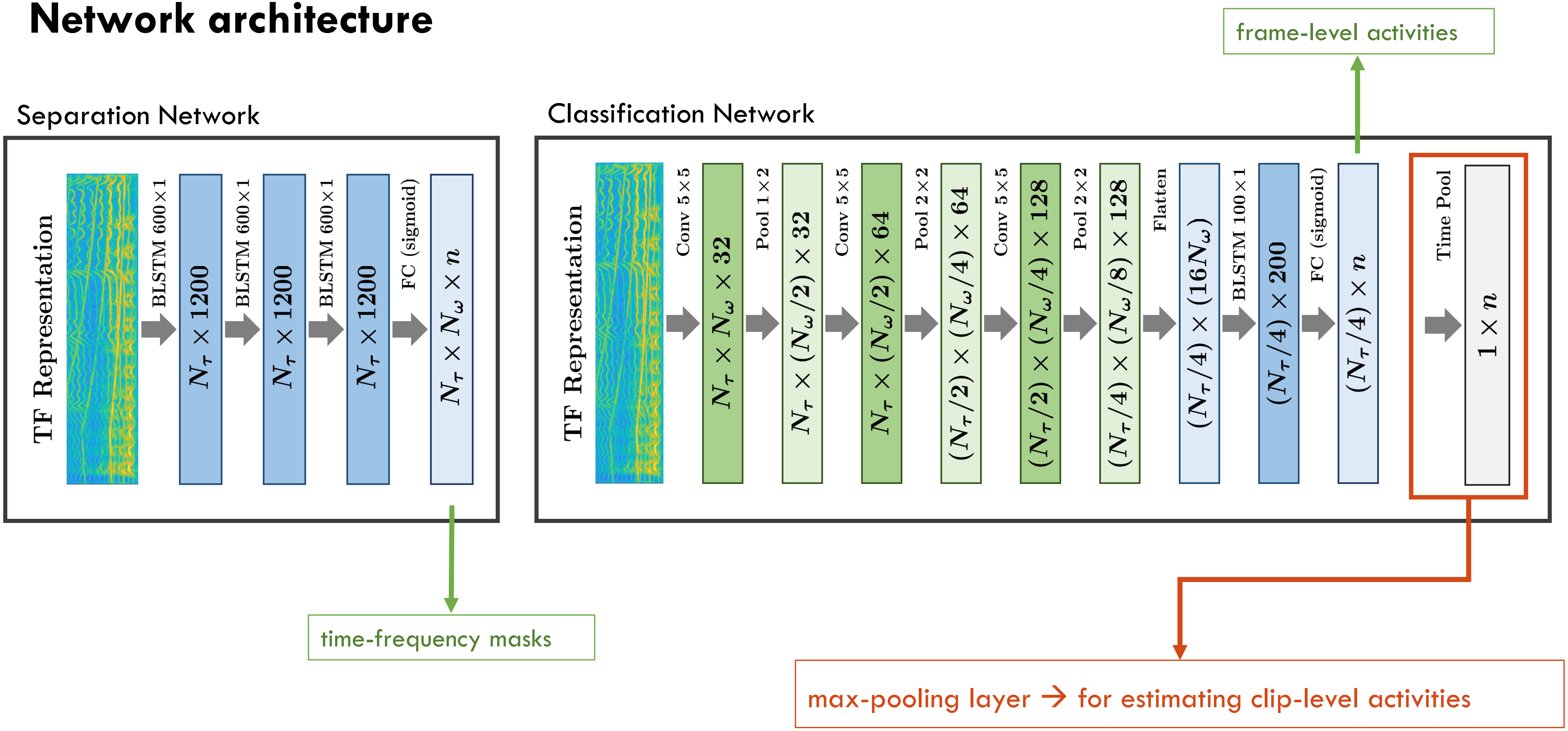
Network architecture



MITSUBISHI ELECTRIC
Changes for the Better

for a greener tomorrow

Network architecture



Experiments

- Dataset
  - Urbansound8K: short excerpts of field recordings
  - Selected classes: car horn, dog bark, gun shot, jackhammer, siren
  - Audio mixtures:
    - Length: 4-sec
    - Sampling rate: 16 kHz
    - Each mixture includes at least 1 sound event
  - Training/validation/test: 20K, 5K, 5K samples

Number of sources	Per-frame distribution	Per-clip distribution
0	0.17	0.00
1	0.28	0.06
2	0.30	0.20
3	0.18	0.34
4	0.06	0.30
5	0.01	0.10

Experiments

- Dataset
  - Urbansound8K: short excerpts of field recordings
  - Selected classes: car horn, dog bark, gun shot, jackhammer, siren
  - Audio mixtures:
    - Length: 4-sec
    - Sampling rate: 16 kHz
    - Each mixture includes at least 1 sound event
  - Training/validation/test: 20K, 5K, 5K samples

Number of sources	Per-frame distribution	Per-clip distribution
0	0.17	0.00
1	0.28	0.06
2	0.30	0.20
3	0.18	0.34
4	0.06	0.30
5	0.01	0.10

Experiments

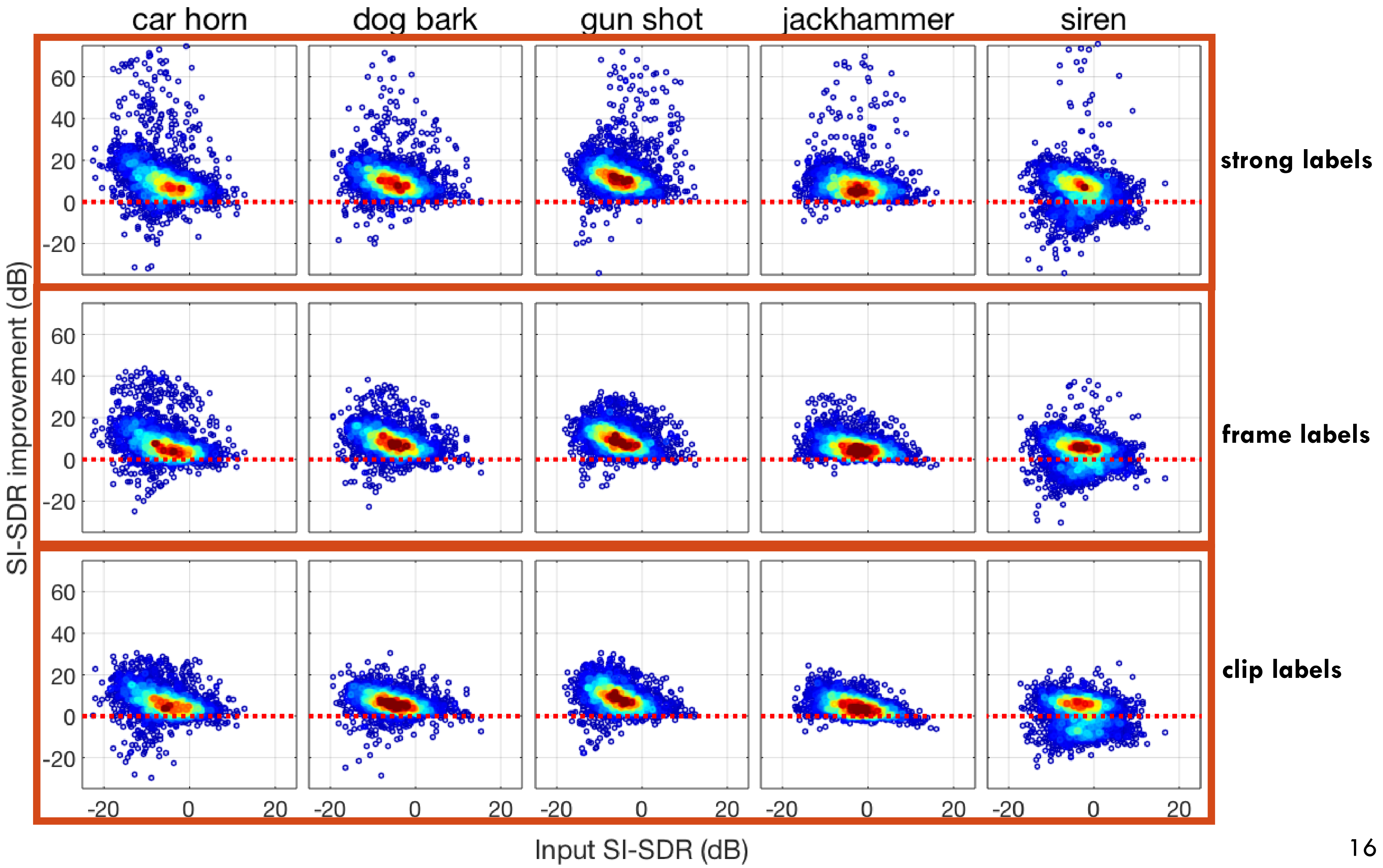
- Dataset
  - Urbansound8K: short excerpts of field recordings
  - Selected classes: car horn, dog bark, gun shot, jackhammer, siren
  - Audio mixtures:
    - Length: 4-sec
    - Sampling rate: 16 kHz
    - Each mixture includes at least 1 sound event
  - Training/validation/test: 20K, 5K, 5K samples

Number of sources	Per-frame distribution	Per-clip distribution
0	0.17	0.00
1	0.28	0.06
2	0.30	0.20
3	0.18	0.34
4	0.06	0.30
5	0.01	0.10

Experiments

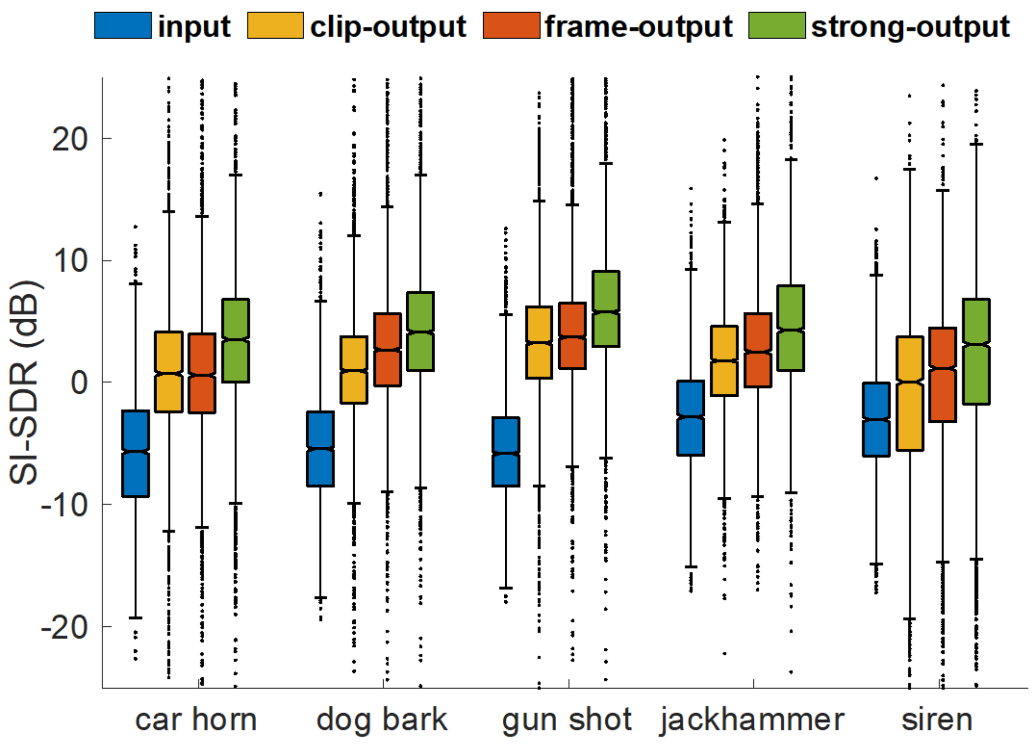
- Training
  - Classifier trained **only** on mixtures (may include isolated cases)
  - Classifier weights **fixed** when training the separator
  - If trained jointly from scratch, the two networks co-adapt, resulting in degradation of separation performance.
- Evaluation measures
  - Separation: scale-invariant source to distortion ratio (SI-SDR)

Results


car horn
dog bark
gun shot
jackhammer
siren
strong labels
frame labels
clip labels
SI-SDR improvement (dB)
60
40
20
0
-20
-20
0
20
Input SI-SDR (dB)

Results

- Siren is the most difficult class in our dataset → contains a more diverse set of sounds (e.g., police siren vs. ambulance siren)
- Distributions of weakly supervised results are very close to strongly supervised results except at the very high SI-SDR range



	Car horn	Dog bark	Gun shot	Jackhammer	Siren	Overall
Input SI-SDR	-5.8 ± 5.1	-5.4 ± 4.8	-5.5 ± 4.4	-2.9 ± 4.8	-3.0 ± 4.6	-4.5 ± 4.9
ΔSI-SDR-clip	6.5 ± 6.1	6.4 ± 4.4	8.8 ± 5.5	4.6 ± 3.8	1.8 ± 6.7	5.6 ± 5.9
ΔSI-SDR-frame	7.0 ± 7.4	8.3 ± 5.6	9.7 ± 5.4	5.7 ± 4.2	3.1 ± 6.4	6.8 ± 6.3
ΔSI-SDR-strong	9.9 ± 10.1	10.0 ± 7.1	12.5 ± 8.0	7.8 ± 6.6	4.9 ± 8.9	9.0 ± 8.6

Audio examples

Mixture

Separated Car Horn

Separated Dog Bark

Separated Jackhammer

Future directions

- Extension to other types of masking, e.g., phase sensitive masking
- Considering unlabeled sounds from other classes in addition to labeled sounds
- Training on datasets with fine-grained labels, e.g., bird songs of different species
- Exploring application of this technique to speech and/or music

MITSUBISHI ELECTRIC
Changes for the Better

for a greener tomorrow

MITSUBISHI ELECTRIC
Changes for the Better