29th EUSIPCO
European Signal Processing Conference
DUBLIN // IRELAND
23–27 AUGUST 2021

eusipco2021.org

[1] National Technical University of Athens, Greece
[2] University of Thessaly, Volos, Greece

Overlapped Sound Event Classification via Multi-Channel Sound Separation Network

EUSIPCO 2021, Dublin, 23-08-2021

P. Giannoulis,[1] G. Potamianos,[2] P. Maragos[1]

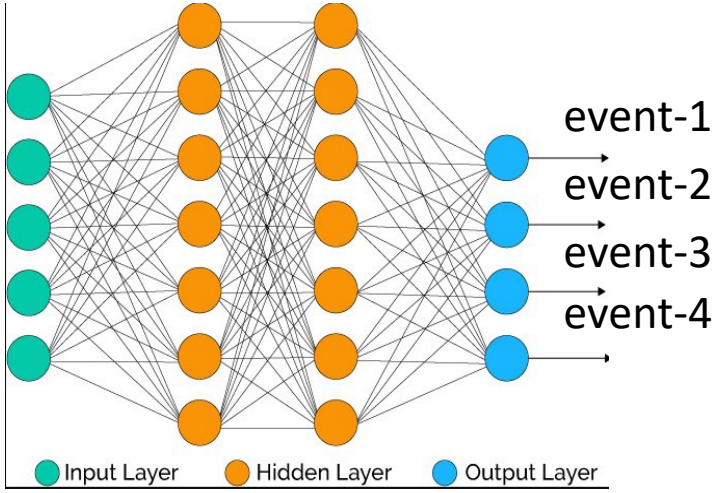
Intro – Deep Learning in Sound Event Classification

- Deep Learning approaches are the **state-of-the-art** in **SEC** since the last ~4 years

  (DCASE'2017–2020 challenges)

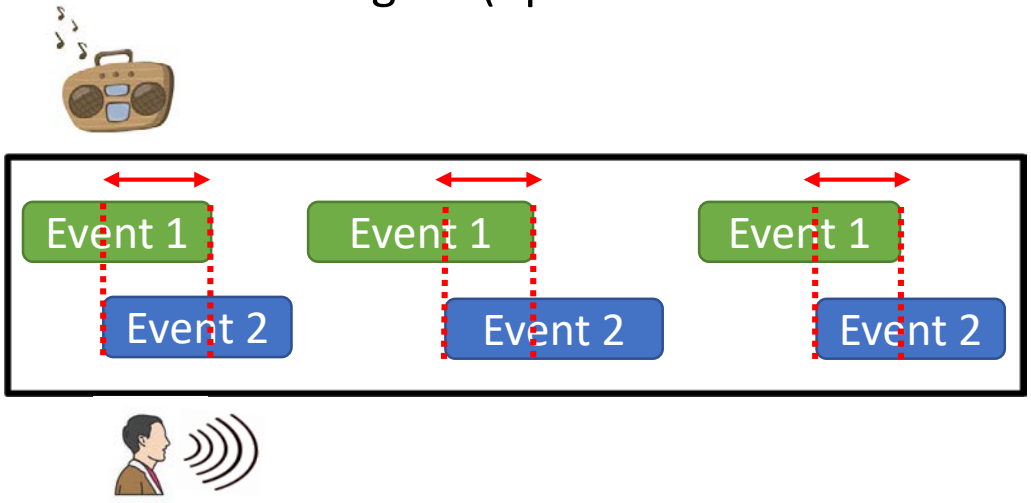
- In **Isolated** case, they have **very good discriminative ability** that outperforms other approaches (HMM, SVM, NMF) when **enough data** are available

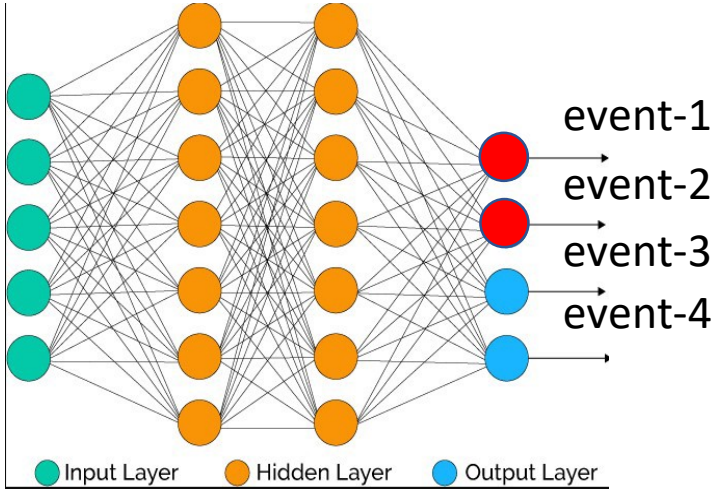
- In **Overlapped** case, they have the natural ability of modeling the presence of multiple events



Intro – Deep Learning in Sound Event Classification

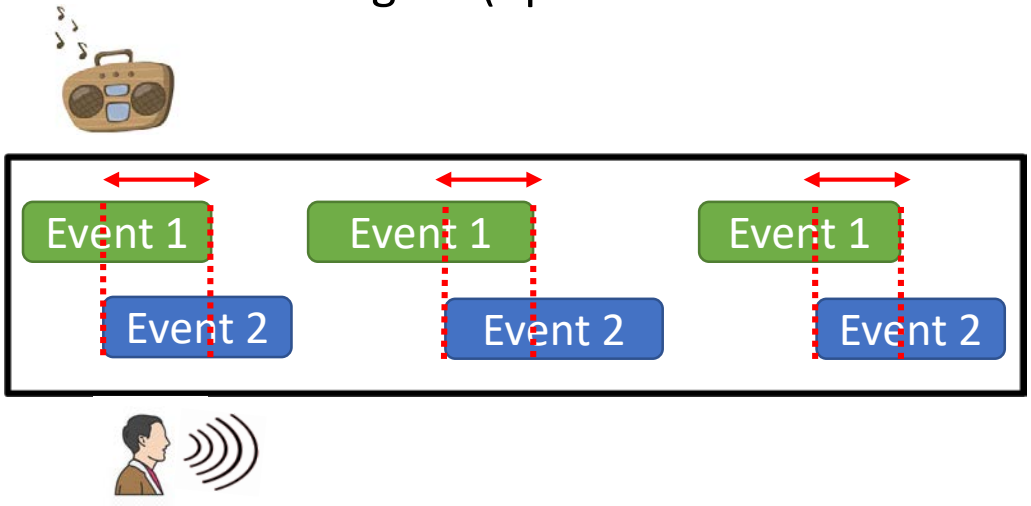
- When dealing with **Overlap**, most deep-learning approaches:
  - rely on **existing overlapped instances** (or data augmentation "mix-up") for their training
  - focus on scenarios with a **few event classes** (~10) and with **polyphony** of 2nd degree (up to 2 simultaneous events)

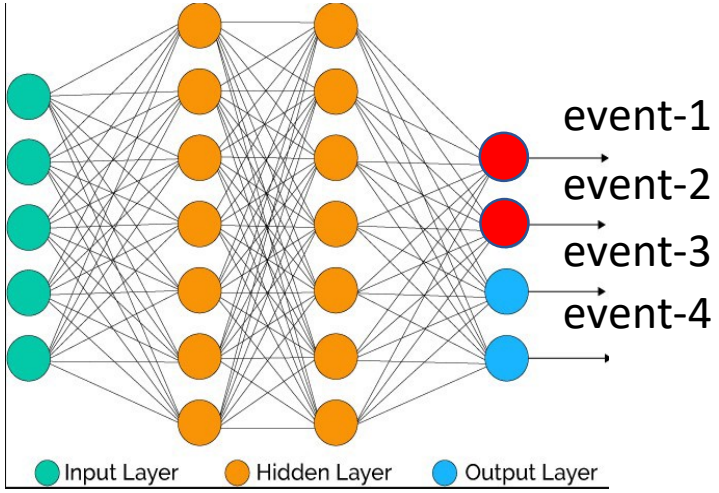




Intro – Deep Learning in Sound Event Classification

- When dealing with **Overlap**, most deep-learning approaches:
  - rely on **existing overlapped instances** (or data augmentation “mix-up”) for their training
  - focus on scenarios with a **few event classes** (~10) and with **polyphony** of 2nd degree (up to 2 simultaneous events)





- Problems:
  - A lot and diverse training data are needed
  - It is hard to obtain enough overlapped data for all combinations when: we have **many event classes** or **polyphony** is high (>2)

Focus – Deep Learning in Sound Event Classification

- Focus:
  - Investigation of the performance of DNN methods for **overlapping** sound event classification when the **number of events is large**
  - How and under which **conditions** (i.e. reverberation) the incorporation of a **separation network** can increase the performance in adverse scenarios

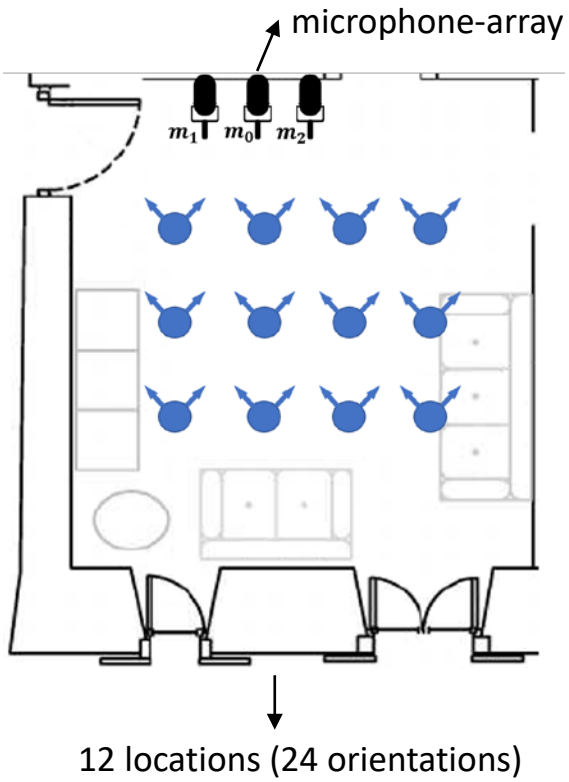
- Contributions:
  - Combination of a **multi-channel source separation** network with a **sound event classification** network shows significant improvements in challenging scenarios

Dataset – Multi-Channel, Overlapping event scenarios

ESC-50 dataset

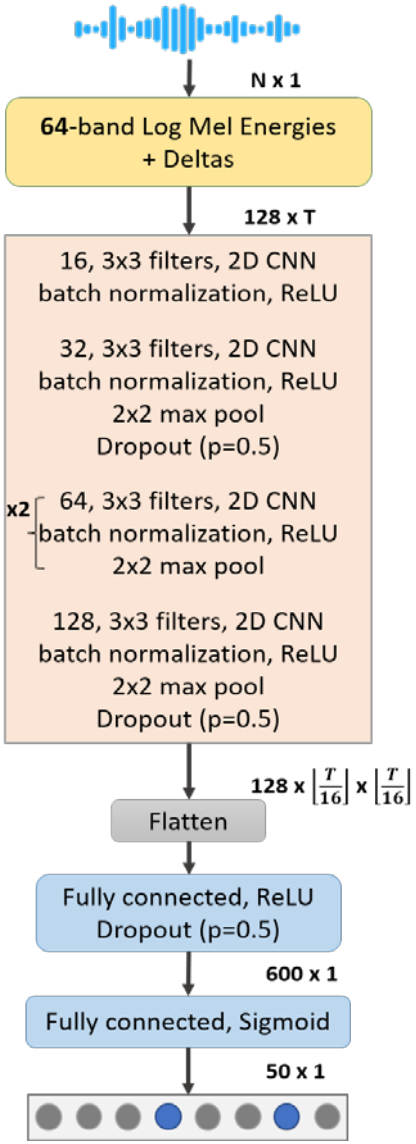
Animals	Natural soundscapes	Human sounds	Domestic sounds	Urban noises
Dog	Rain	Crying baby	Door knock	Helicopter
Rooster	Sea waves	Sneezing	Mouse click	Chainsaw
Pig	Crackling fire	Clapping	Keyboard typing	Siren
Cow	Crickets	Breathing	Door, wood creaks	Car horn
Frog	Chirping birds	Coughing	Can opening	Engine
Cat	Water drops	Footsteps	Washing machine	Train
Hen	Wind	Laughing	Vacuum cleaner	Church bells
Insects (flying)	Pouring water	Brushing teeth	Clock alarm	Airplane
Sheep	Toilet flush	Snoring	Clock tick	Fireworks
Crow	Thunderstorm	Drinking, sipping	Glass breaking	Hand saw

DIRHA dataset



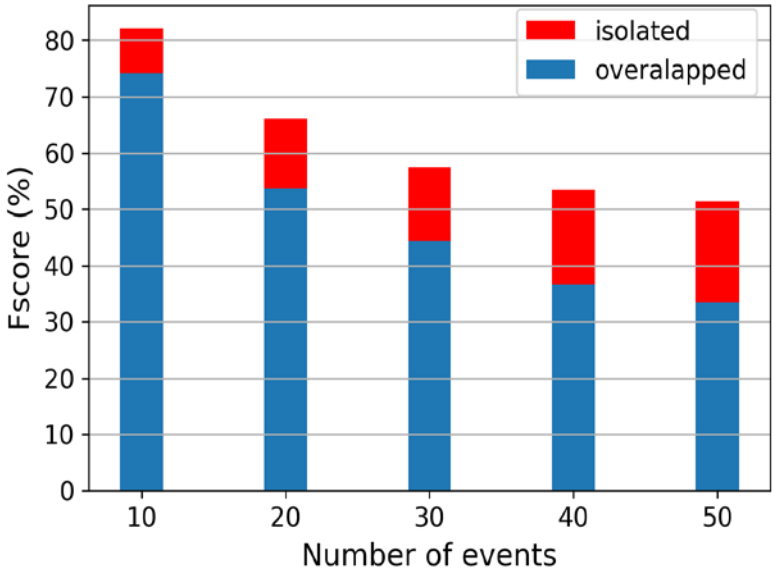
- **ESC-50** dataset: 2000 clips of 5sec (40 clips per event)
- **DIRHA** Livingroom **RIRs** ($\boldsymbol{T_{60}} \in [0.58s, 0.81s]$)
- Create **Isolated** & **Overlapped** instances **1.5sec**-long from different positions

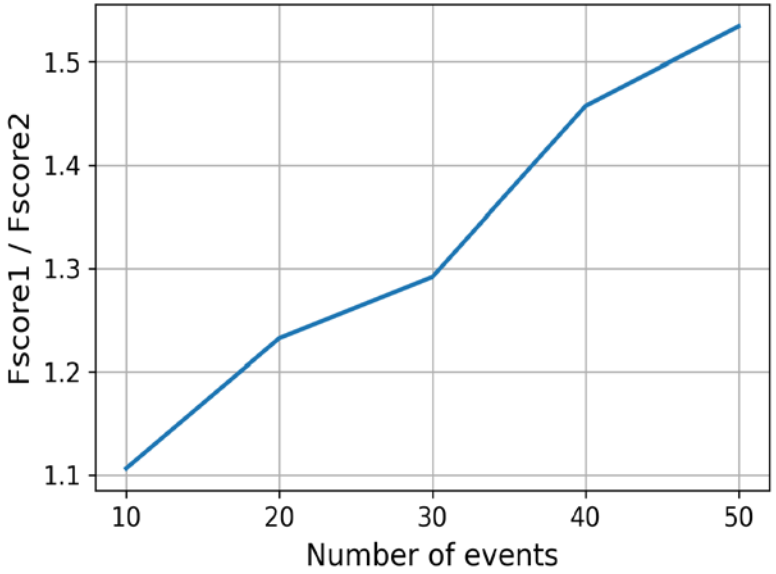
DNN SEC network – Single-channel Baseline Network



Motivation – Increasing number of event classes

- How the number of possible event classes affects the performance of multi-label DNNs in Overlapped SEC?

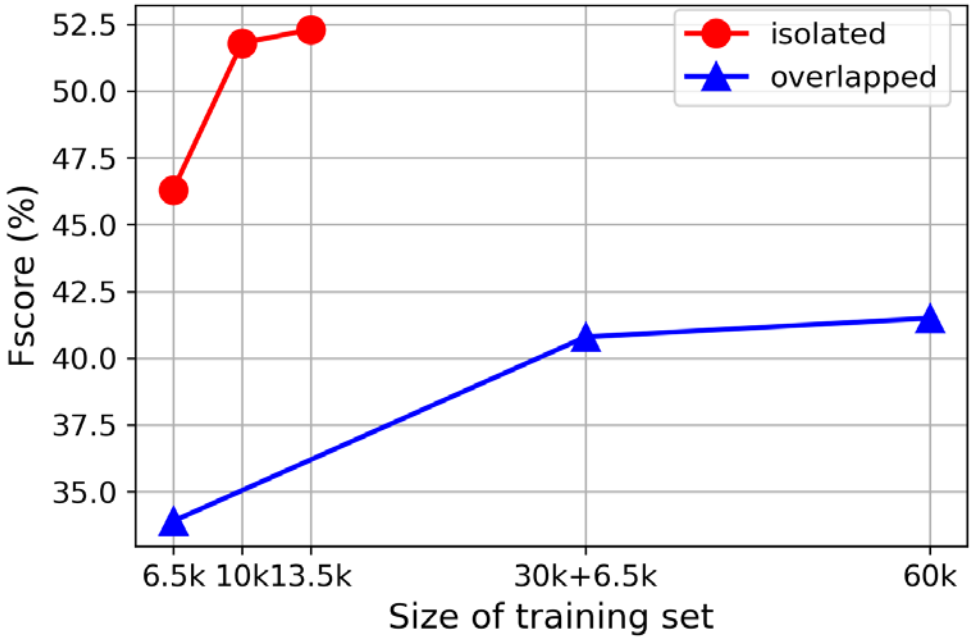




- As # of event classes increases → performance drops both in isolated and overlapped scenarios (left fig.)
- In **Overlapped case,** the performance deteriorates more, and the gap increases. (i.e. when N=10 events, we observe similar performances)

Motivation – Training data size

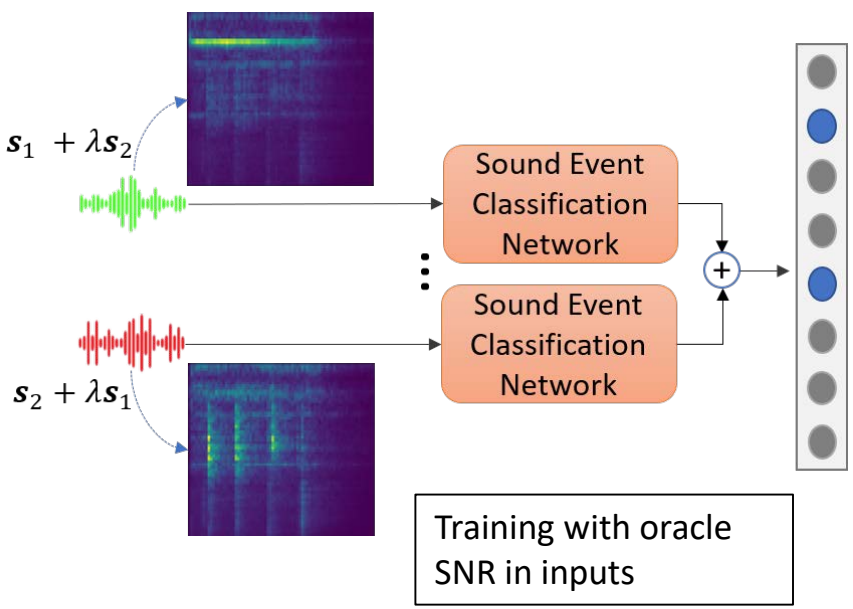
- Can we increase the performance of **overlapped scenario** by **creating overlapped instances** (augmented)?

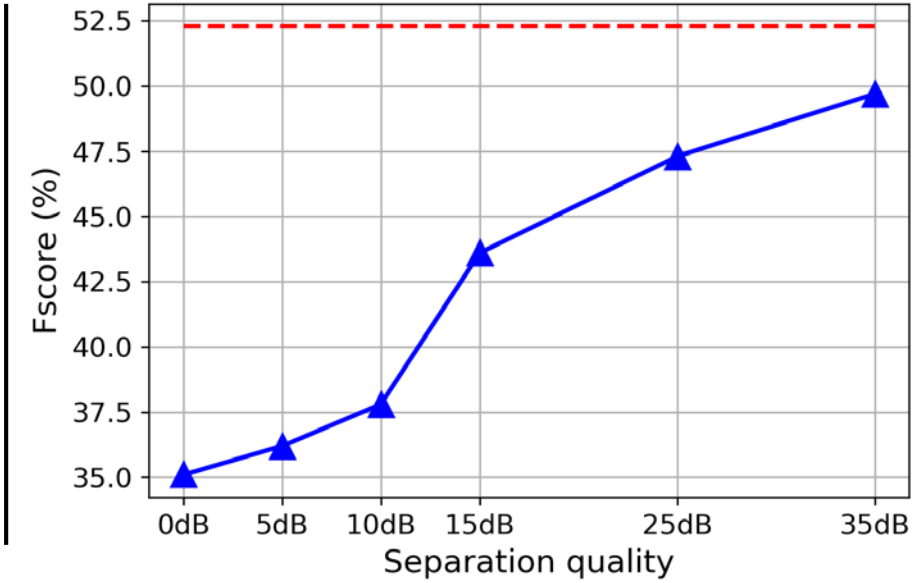


- Yes, we can improve performance (33.5% → 41%) but **up to a point**.
- Still the **gap** between overlapped and isolated scenarios **can be large**.

Motivation – Separation & Classification

- Can we approximate the isolated performance in overlapped scenarios?
- **Idea:** Combine **Sound Separation** & **Event Classification**
  - Separate acoustic events (pre-processing step: overlapped → set of isolated inputs)
  - Exploit good performance of DNNs for isolated AED





- Given **<u>adequate</u> quality of separation**, performance gap can be reduced **significantly**!

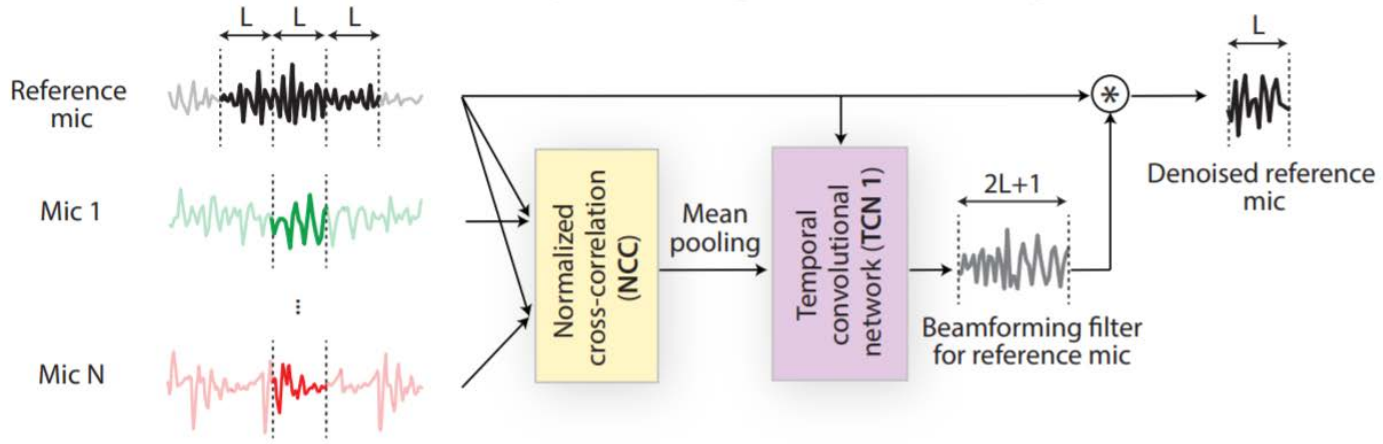
Method – Sound Source Separation (I)

- State-of-the-art **Speech Separation** methods based on Single-Channel (!) Deep learning, achieve impressive performance for 2&3 simultaneous speakers.
- **Challenges:**
  - Transition Speech Separation → Arbitrary Sound Event Separation
  - Usually trained & tested in **>4sec** audio segments
  - **Reverberation** time affects the performance (tested usually in up to T_{60}=0.5sec)
  - Achieve **sufficiently good** performance (>10dB?)

- In our work, we propose for the **1st time** the combination of **Multi-channel** Sound Source Separation Network with Event Classification network.
  - Takes advantage of **Spectral** content like Single-channel methods
  - Takes advantage of **Spatial** separation of sound sources

Method – Sound Source Separation (II)

- **TAC-FaSNet**: Multi-channel **Neural Beamforming** (diagram of original FaSNet core idea)



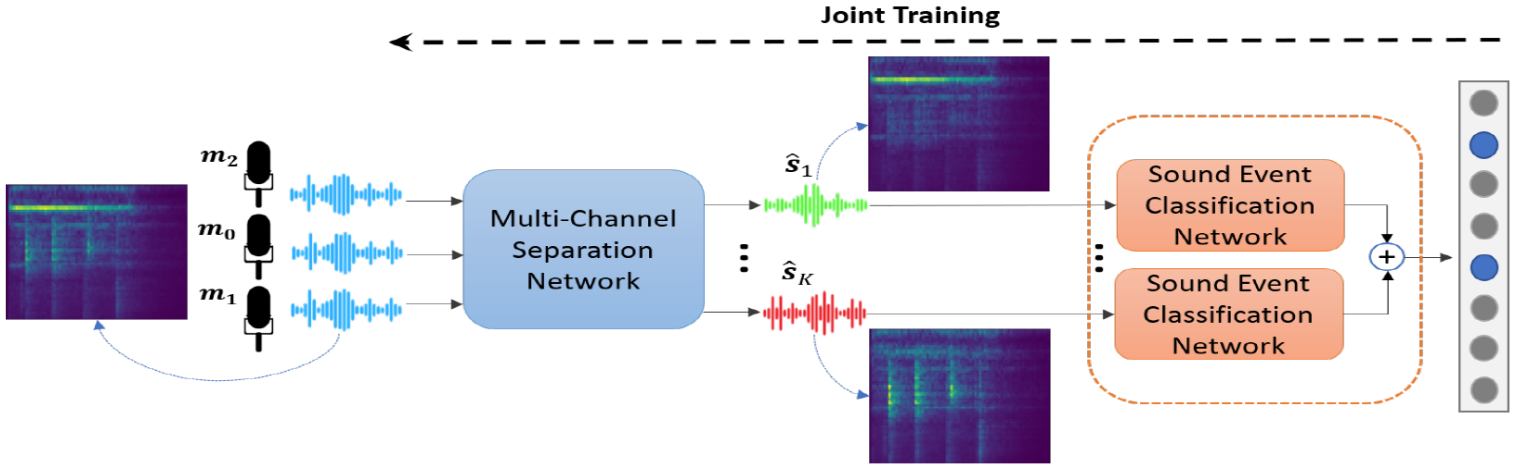
- Loss Function: **SI-MSE** (scale invariant MSE) for Log-Mel Spectrograms

$$\begin{cases} \mathbf{Y}_c = \left| \mathrm{STFT}\left(\frac{\mathbf{y}_c}{\|\mathbf{y}_c\|_2} \right) \right| \\ \mathbf{Y}_c^* = \left| \mathrm{STFT}\left(\frac{\mathbf{y}_c^*}{\|\mathbf{y}_c^*\|_2} \right) \right| \end{cases}$$

$$\mathcal{L}_{obj} = \frac{1}{C} \sum_{c=1}^{C} \mathrm{MSE}(\mathbf{Y}_c\mathbf{M}, \mathbf{Y}_c^*\mathbf{M})$$

- *Y. Luo, C. Han, N. Mesgarani, E. Ceolini, and S. C. Liu, "**FaSNet: Low latency adaptive beamforming for multi-microphone audio processing**," in ASRU, 2019, pp. 260–267.*
- *Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "**End-to-end microphone permutation and number invariant multi-channel speech separation**," arXiv preprint arXiv:1910.14104, 2020.*

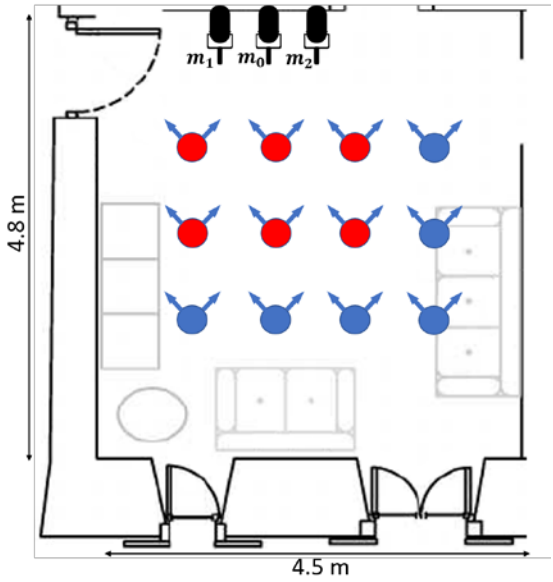
Method – Proposed Pipeline



Pipeline of the proposed system for overlapped sound event classification.

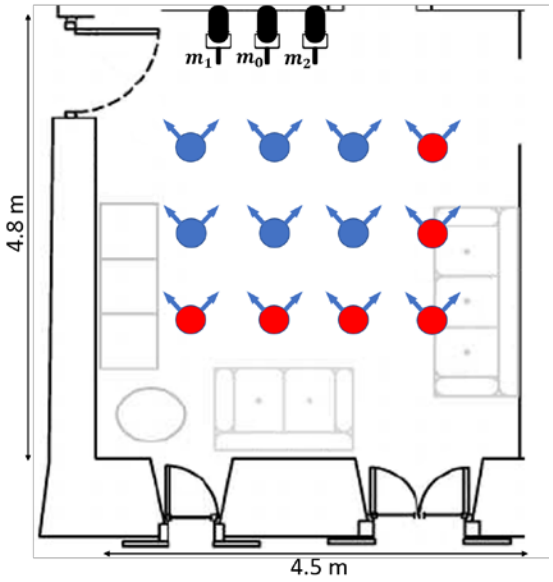
- 2 ways of **Training:**
  - **Sequential Training:**
    - Train Separation network on artificial mixtures of isolated instances
    - Train SEC network on the separated outputs of Separation network
  - **Joint Training:** Joint Re-Training of the 2 networks
    - Parameters of both networks are fine-tuned towards the final objective of SEC
- **Multi-channel Polyphony Network:** to apply the proposed system only in overlapped segments.
  - Spatial content (GCC features)
  - Spectral content (Log-Mel features)

Experiments – 2 Reverberation scenarios



Mid Reverberation

$\bar{T}_{60}$= 0.61s



High Reverberation

$\bar{T}_{60}$= 0.80s

- It is well-known that reverberation affects the quality of the separation systems

Results – Sound event classification

TABLE I
PERFORMANCE OF THE VARIOUS SYSTEMS FOR THE OVERLAPPED-EVENT SCENARIO, IN TERMS OF FSCORE.

System	Fscore (%)	
	$\overline{T}_{60}$=0.61s	$\overline{T}_{60}$=0.80s
(A) Baseline (1 channel)	41.26	39.05
(B) Baseline (3 channels)	41.45	**39.33**
(C) Proposed - Sequential	44.72	38.41
(D) Proposed - Joint	**47.46**	38.75
Late Fusion (B+C)	46.20	41.52
Late Fusion (B+D)	**48.95**	**41.95**

- **Table I**

Mid Reverberation:

- Proposed pipeline (C,D) **improves** the baseline performance
- Joint Training (D) better than Sequential (C)
- Fusion of proposed (D) with baseline (B) achieves **7.7%** absolute improvement compared to the baseline (A)

High Reverberation:

- Proposed pipeline (C,D) **fails** to improve the baseline performance – due to inadequate performance of the separation network
- Still, the **fusion schemes** achieve improvements over the baseline

Results – Polyphony classification

TABLE II
PERFORMANCE OF THE POLYPHONY CLASSIFICATION NETWORK FOR DIFFERENT FEATURE SETS, IN TERMS OF ACCURACY.

Features	Notation	Accuracy (%)
logFBE	$[\mathbf{X}^{(s_0)}]$	95.59
GCC	$[\mathbf{GCC}^{(s_0,s_1)}; \mathbf{GCC}^{(s_0,s_2)}]$	98.68
logFBE + GCC	$[\mathbf{X}^{(s_0)}; \mathbf{GCC}^{(s_0,s_1)}; \mathbf{GCC}^{(s_0,s_2)}]$	**99.27**

- **Table II**
  - Almost perfect performance (99.27%) when we combine GCC with Log-Mel
  - Estimated close to Oracle (ensures applicability of our approach)

Conclusions

- We examined the combination of sound source separation with overlapped sound event classification in a multi-channel setup with a large variety of event classes

- Our results showcase the potential of incorporating separation methods to SEC systems, for overlapping scenarios

- High reverberation can be a limiting factor for the efficiency of the proposed pipeline

- With the continuous improvement of separation networks in the recent years, systems similar to the proposed pipeline can soon play a major role in the overlapping sound event detection task

Thank You!