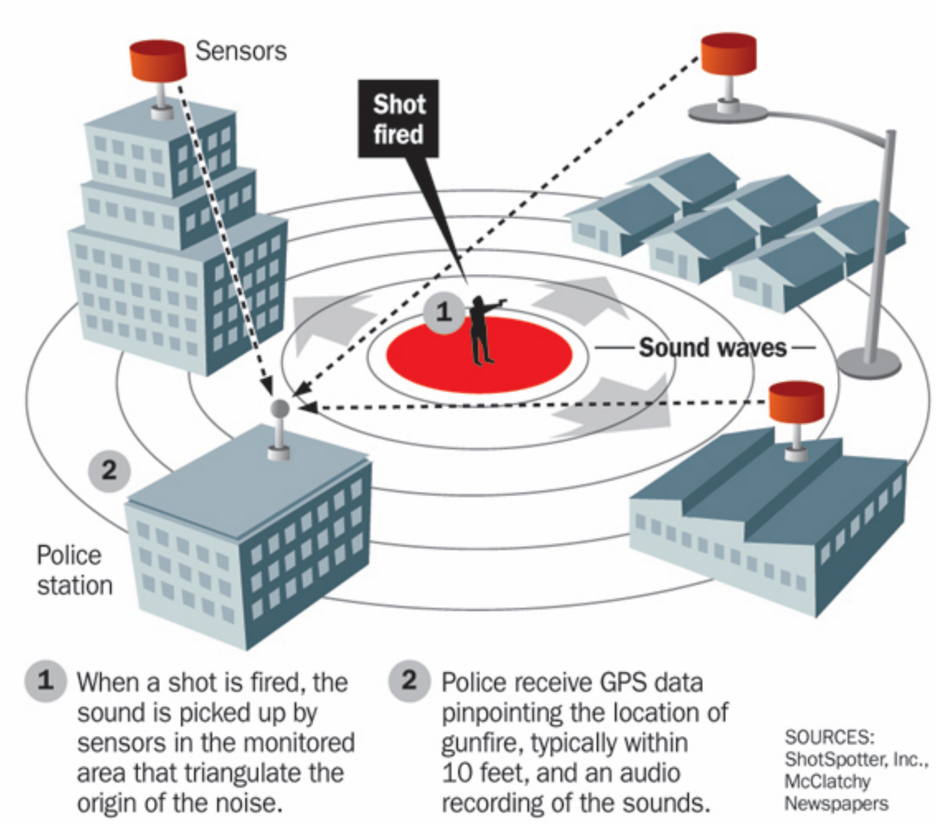
Learning-based Cooperative Sound Event Detection with Edge Computing

Jingrong Wang[†], Kaiyang Liu[†*], George Tzanetakis[†], Jianping Pan[†]

[†]Department of Computer Science, University of Victoria, Victoria, Canada
[*]School of Information Science and Engineering, Central South University, Changsha, China
Email: {jingrongwang, liukaiyang, pan}@uvic.ca, gtzan@cs.uvic.ca

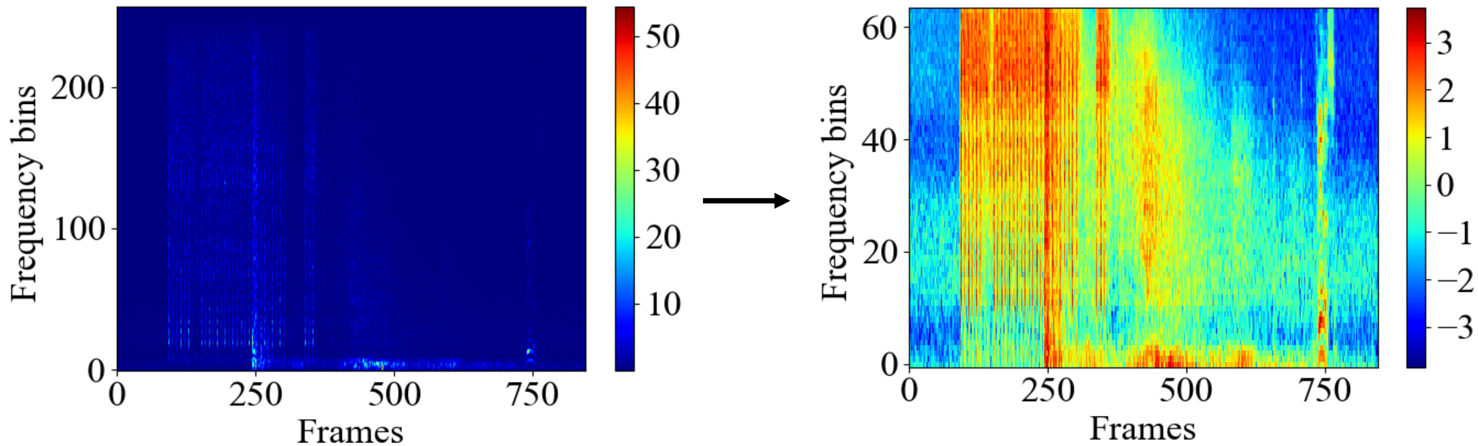
Problem and Motivation

- Gunshot violence increasing…
  - 6,000+ reported last year in US but 80% more unreported
  - Slow response time: about 10 minutes since incoming 911 calls
  - Lives and evidence lost
- New services, e.g., ShotSpotter
  - Sensors installed in certain places
  - Audio clips sent to cloud for ID
  - 90% identified in about 1 minute
  - Cost and scalability problem



How to identify a sound event?

- First, extract the audio features



(1) Spectrogram

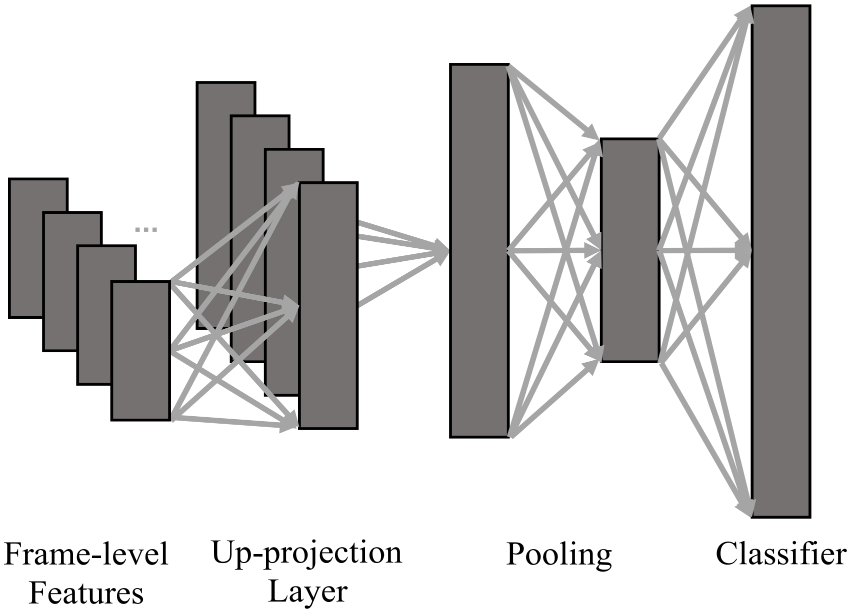
(2) Log-scale mel spectrogram

"Short-Time Fourier Transform" + "Log-Mel Spectrogram"

Then classification based on extracted features

E.g., Deep Bag-of-Frames learning-based approach [1]

Randomly choose 128 batches as audio features →



NVIDIA GTX 970 4GB: "4h+" + "300MB+"

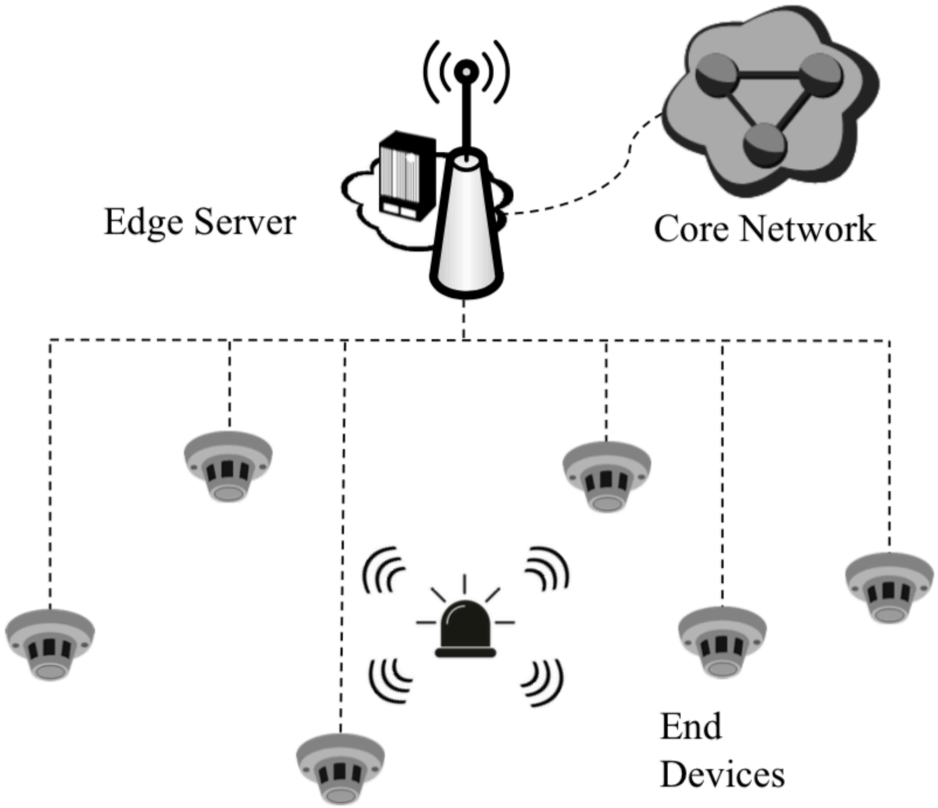
[1] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8M: A large- scale video classification benchmark," *arXiv preprint arXiv:1609.08675*, 2016.

Challenges

- Delay-sensitive + computation-intensive
  - Front-end devices → limited computation capabilities [2]
  - Cloud → high communication latencies [3]
  - Communication among devices, or through an access point
- Edge computing
  - Enhances and extends the cloud services at the edge of the network
  - Deploys computation capacity closer to where the data is captured
  - Breakdown between devices, edge and cloud?

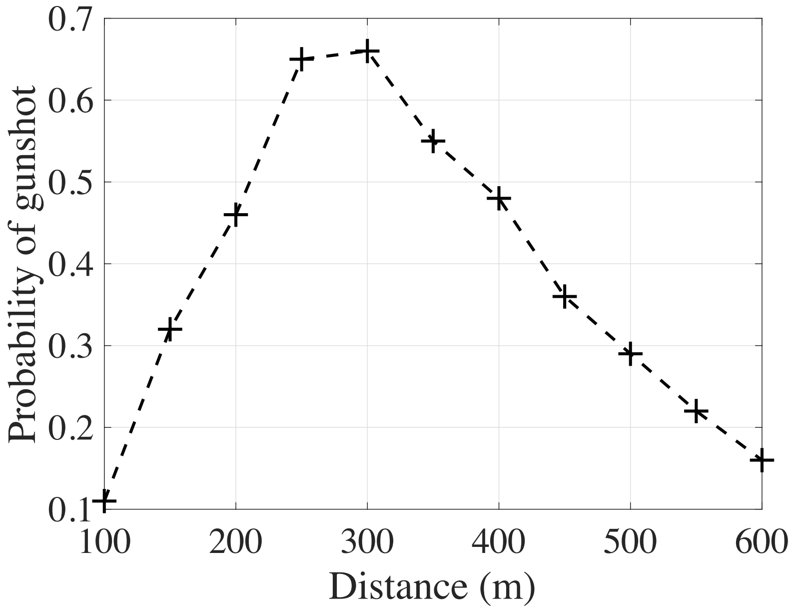
[2] X. Ran, H. Chen, X. Zhu, Z. Liu, and J. Chen, "DeepDecision: A mobile deep learning framework for edge video analytics," in *Proc. of IEEE INFOCOM*, 2018.
[3] K. Hong, D. Lillethun, U. Ramachandran, B. Ottenwälder, and B. Koldehofe, "Mobile fog: A programming model for large-scale applications on the internet of things," in *Proc. of ACM SIGCOMM workshop on Mobile cloud computing*, 2013, pp. 15–20.

Edge computing system setup



- Front-end acoustic devices
  - Slow local execution
- Edge server
  - Wireless comm. overhead
- Cloud server
  - Backbone congestion

Why multiple acoustic sensors?



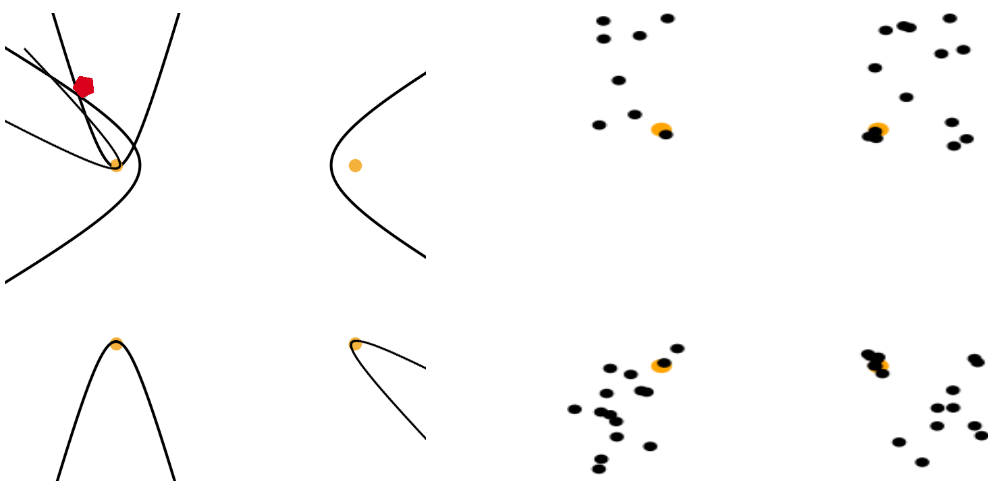
- Localization by triangulation
- Classification accuracy is affected by:
  - Training data (Google Audioset)
  - Learning algorithm (DBof)
  - Distance
    - Near field
    - Reverberant field
- Joint localization and classification needed

Localization

- Least-squares formulation
  - Time difference of arrival (TDOA)
  - Minimize the quadratic difference between the predicted and the actual value

$$A^* = \arg\min_{A} \sum_{i=1}^{N} \sum_{j=1}^{M} \left\{ \left\| \begin{pmatrix} x_i^* \\ y_i^* \end{pmatrix} - \begin{pmatrix} a_j^* \\ b_j^* \end{pmatrix} \right\|_2 - \left\| \begin{pmatrix} a_j^* \\ b_j^* \end{pmatrix} \right\|_2 - D_{i,j}^* \right\}^2$$

- Deadzone
  - Hyperbolas + measurement noise



- End devices
- Deadzone

Aggregated classifier

- Merge multiple learners to obtain a more accurate prediction than any individual learner alone
  - Ensemble learning → Majority vote

Algorithm 1 EC algorithm

Predict the labels of a sound event instance m aggregated from each end device and record the confidence of the predicted class p, that is, (21) and $v_{n,p}$.
Calculate the total vote for each predicted class $V(p) = \sum_{n=1}^{N} v_{n,p}$.
if $\boxed{\max C'(m,p) > \epsilon}$ OR $\boxed{V(p) >= N/2}$ **then**
  Class p is added to the final decision.
else
  Class p is not considered in the final decision.
end if

High confidence

Majority vote

Performance evaluation: Scenario and metrics

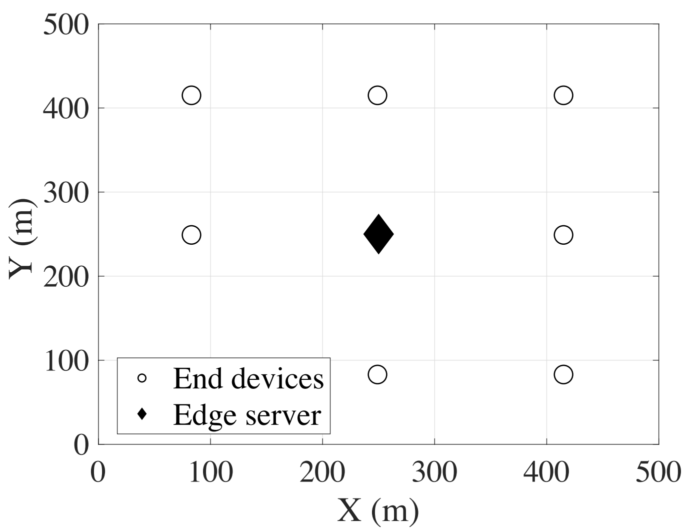


Fig. 1 Grid deployment

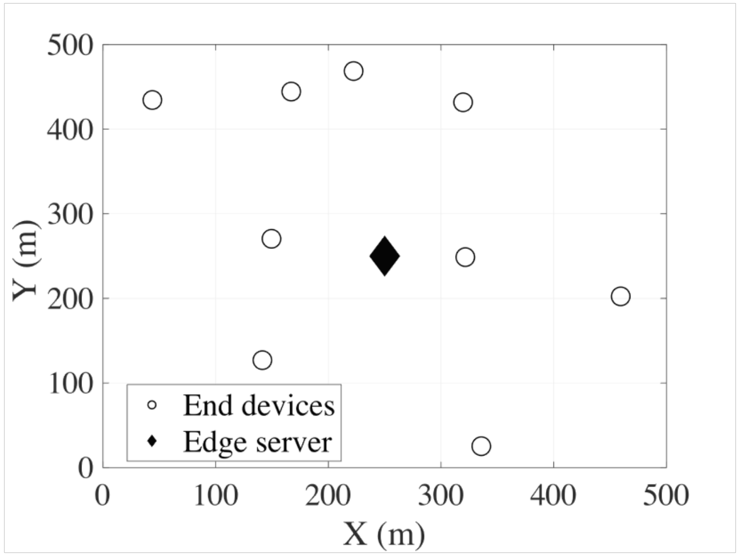


Fig. 2 Random deployment

Parameter	Value	Parameter	Value
Area	500 m×500 m	r	100 m
W	20 MHz	D	3840 kbit
P^{TX}	23 dBm	N_0	-174 dBm
$1/\eta$	4.28 [24]	σ_1	3.6
σ_2	1	γ^{I}	[2, 10] Mbps

Tab. 1 System parameters

- Metrics
  - –Response time (RT)
  - –Classification accuracy (CA)
  - –Localization error (LE)
  - –Dead zone ratio (DZ)

Performance evaluation – Response time

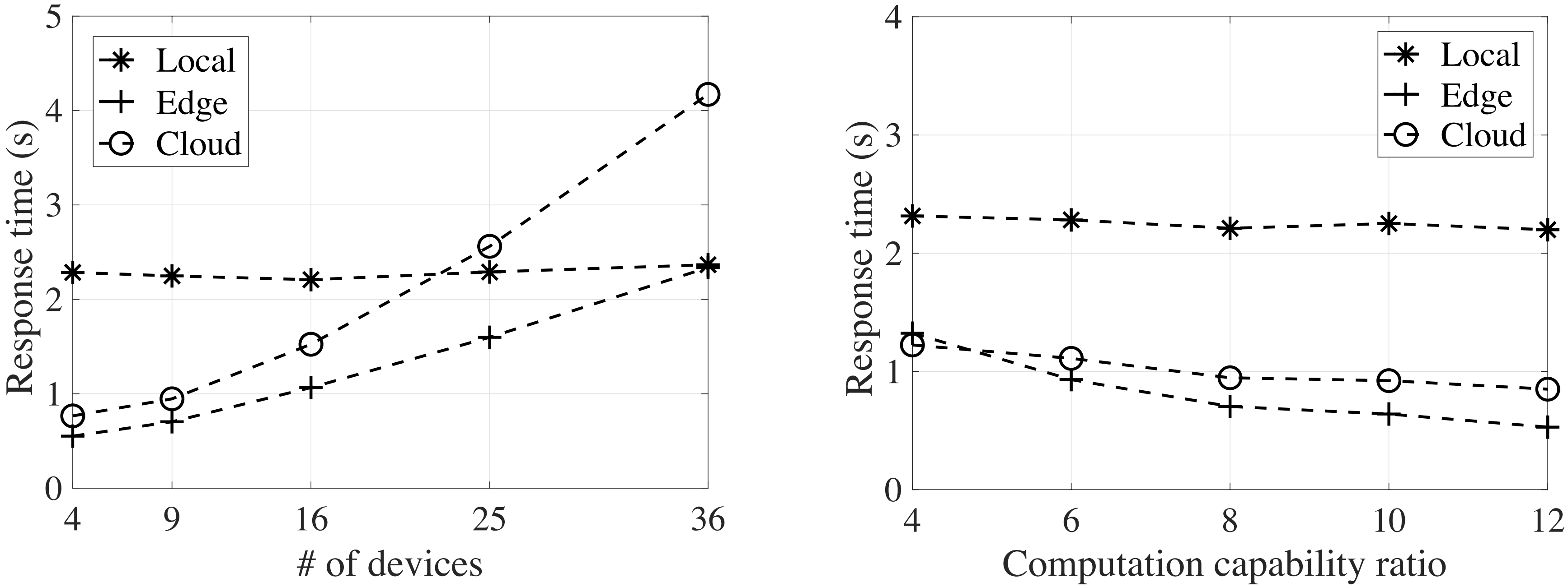


Fig. 3 Response time

Performance evaluation – Classification accuracy

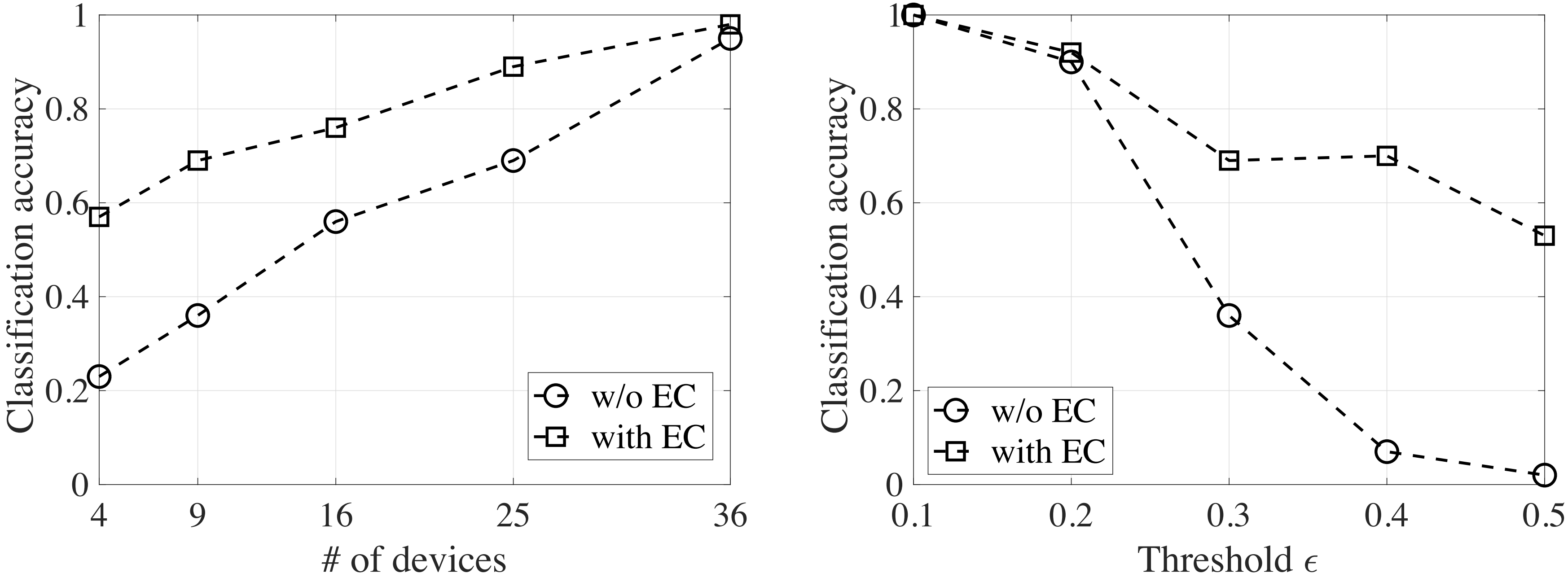


Fig. 4 Classification accuracy

Performance evaluation – Localization & random deployment

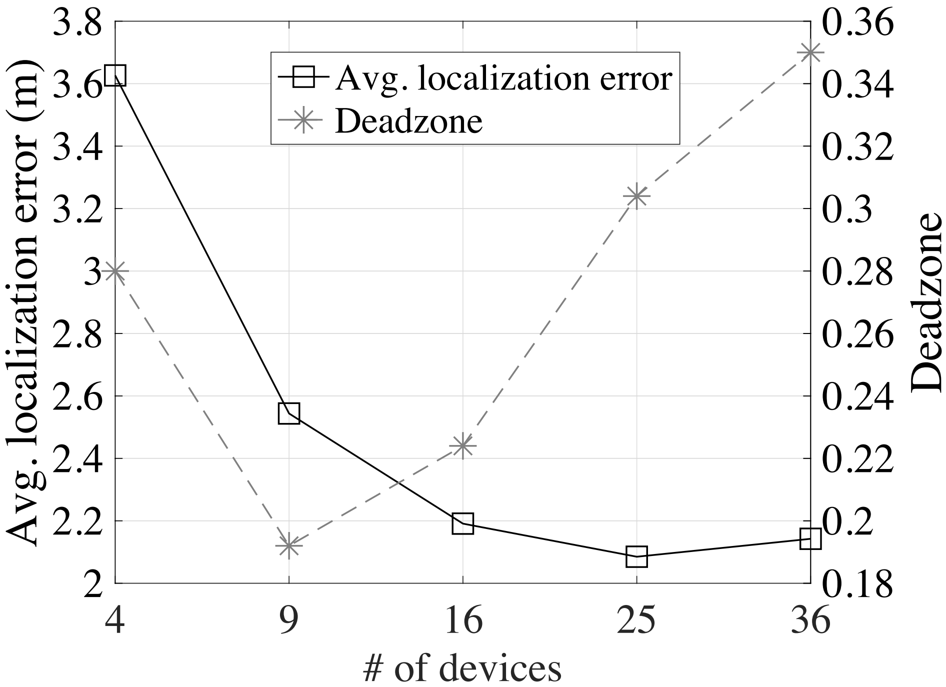
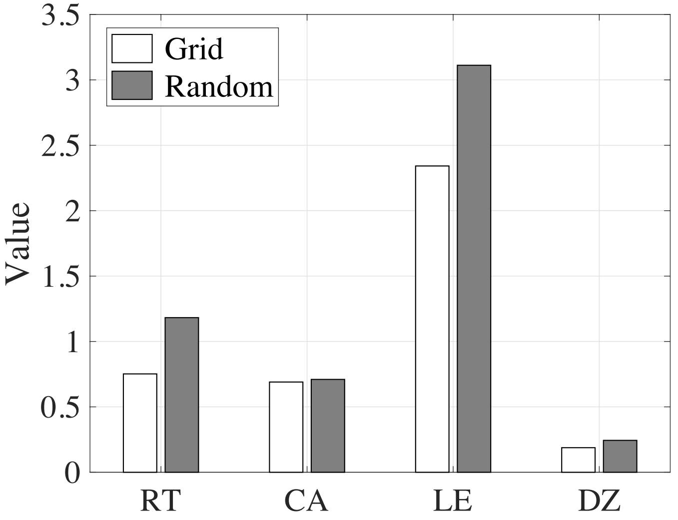


Fig. 5 Localization performance



Values	RT	CA	LE	DZ
Grid	0.752 s	69 %	2.34 m	18 %
Random	1.182 s	71 %	3.12 m	24 %

Tab. 2 Impact of deployment

Conclusion and future work

- Edge-assisted sound event detection framework
  – Computation capacity at the edge of the network
- Ensemble-based cooperative processing
  – Aggregates information for a more accurate result
- Future work
  – Realistic sound propagation model + complex acoustic scenario
  – Distance-weighted differentiation

Q&A

Thanks!

Wireless communication model

- Path loss model

$$PL_n = PL(d_0) + 10\theta\log\left(\frac{d_n}{d_0}\right)$$

  - d_n (in m) > d_0 is the distance between the base station and device n
  - θ is the path loss exponent
  - d_0 is the reference distance for the antenna far-field propagation effect

- Received signal strength

$$P_n = P^{\mathrm{TX}}\text{-}PL_n\text{-}X_{\sigma_1}$$

  - P^{TX} (in dBm) is the transmitted power of device n
  - X_{σ_1} denotes the shadowing fading (in dB) subject to the Gaussian distribution with zero mean and standard deviation σ_1

- Maximum uplink transmission rate

$$r_n^{\mathrm{TX}} = W\log_2(1 + \frac{10^{P_n/10}}{I_n + N_0})$$

  - W is the channel bandwidth, N_0 (in mW) is the noise power
  - I_n (in mW) is the interference signal from other devices