Exploiting Attention-based Sequence-to-Sequence Architectures for Sound Event Localization

EUSIPCO 2020

Christopher Schymura, Tsubasa Ochiai, Marc Delcroix, Keisuke Kinoshita, Tomohiro Nakatani, Shoko Araki and Dorothea Kolossa

January 18-22 2021

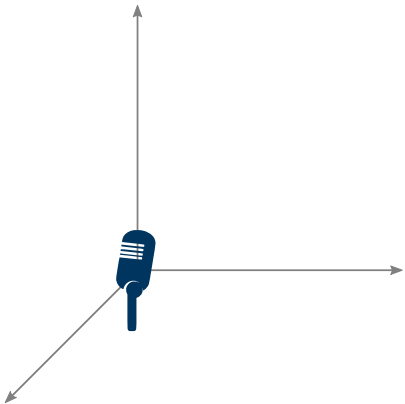
RUHR UNIVERSITÄT BOCHUM

RUB

NTT

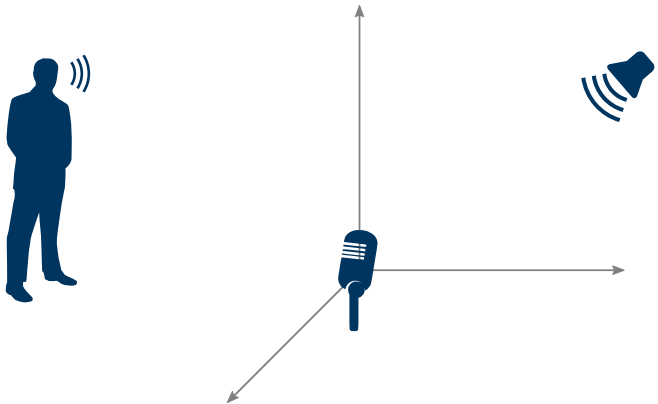
Problem statement

Sound event localization (SEL) aims at finding the positions of *active* sound sources in the environment



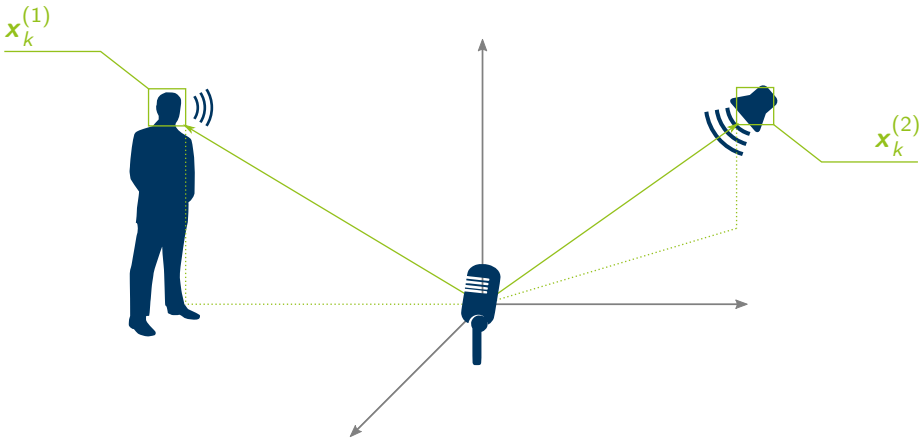
Problem statement

Sound event localization (SEL) aims at finding the positions of *active* sound sources in the environment



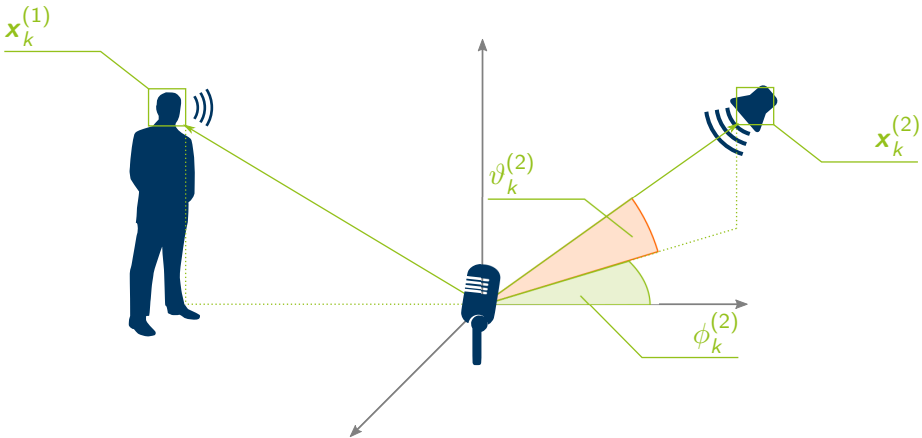
Problem statement

Sound event localization (SEL) aims at finding the positions of *active* sound sources in the environment



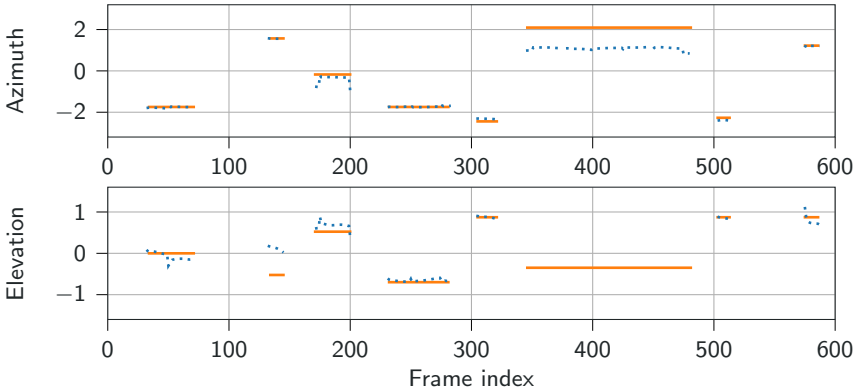
Problem statement

Sound event localization (SEL) aims at finding the positions of *active* sound sources in the environment



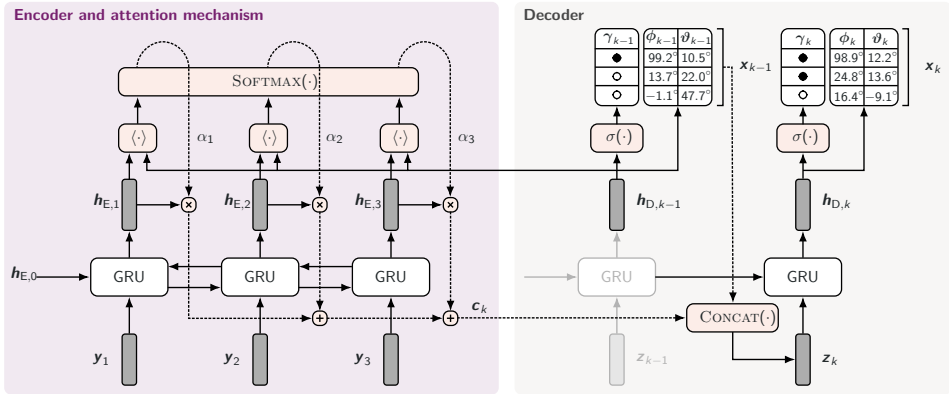
A deeper look into sound event localization

Source activity and position can change over time and multiple sources can be active simultaneously



Proposed framework: ADRENALINE[1]

A sequence-to-sequence model with attentions is a suitable framework to handle temporal context



[1] Attention-based Deep Recurrent Neural-Network for Localizing Sound Events

Datasets and baseline models for evaluation

Experimental evaluation covers multi-source SEL datasets with various acoustic conditions

- A subset from the **TUT Sound Events 2018**[2] corpus was used for evaluation:
  - ANSYN: Anechoic and synthetic impulse responses
  - RESYN: Reverberant and synthetic impulse responses
  - REAL: Reverberant and real-life impulse responses

[2] A. Adavanne, J. Politis, J. Nikunen and T. Virtanen, "*Sound event localization and detection of overlapping sources using convolutional recurrent neural networks*", IEEE JSTSP, 2019

Datasets and baseline models for evaluation

Experimental evaluation covers multi-source SEL datasets with various acoustic conditions

- A subset from the **TUT Sound Events 2018**[2] corpus was used for evaluation:
  - ANSYN: Anechoic and synthetic impulse responses
  - RESYN: Reverberant and synthetic impulse responses
  - REAL: Reverberant and real-life impulse responses
- Each subset contains **3 cross-validation splits** with 240 files for training and 60 files for validation. Each file represents Ambisonic audio signals of 30 s duration.

[2] A. Adavanne, J. Politis, J. Nikunen and T. Virtanen, "*Sound event localization and detection of overlapping sources using convolutional recurrent neural networks*", IEEE JSTSP, 2019

Datasets and baseline models for evaluation

Experimental evaluation covers multi-source SEL datasets with various acoustic conditions

- A subset from the **TUT Sound Events 2018**[2] corpus was used for evaluation:
  - ANSYN: Anechoic and synthetic impulse responses
  - RESYN: Reverberant and synthetic impulse responses
  - REAL: Reverberant and real-life impulse responses
- Each subset contains **3 cross-validation splits** with 240 files for training and 60 files for validation. Each file represents Ambisonic audio signals of 30 s duration.
- The signals are composed of **up to 3 simultaneously active sound sources** at different locations, taken from **11 different sound event classes**.

[2] A. Adavanne, J. Politis, J. Nikunen and T. Virtanen, "*Sound event localization and detection of overlapping sources using convolutional recurrent neural networks*", IEEE JSTSP, 2019

Datasets and baseline models for evaluation

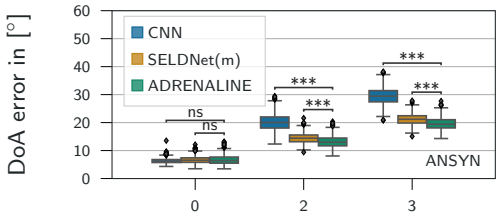
Experimental evaluation covers multi-source SEL datasets with various acoustic conditions

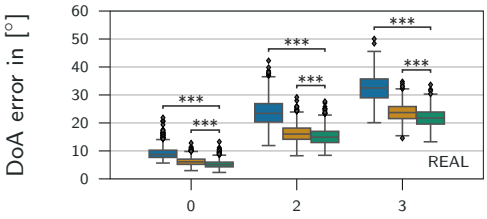
- A subset from the **TUT Sound Events 2018**[2] corpus was used for evaluation:
  - ANSYN: Anechoic and synthetic impulse responses
  - RESYN: Reverberant and synthetic impulse responses
  - REAL: Reverberant and real-life impulse responses
- Each subset contains **3 cross-validation splits** with 240 files for training and 60 files for validation. Each file represents Ambisonic audio signals of 30 s duration.
- The signals are composed of **up to 3 simultaneously active sound sources** at different locations, taken from **11 different sound event classes**.
- ADRENALINE uses the same **CNN-based feature extraction** initially proposed for **SELDNet**[2], which are also used as baseline methods here.

[2] A. Adavanne, J. Politis, J. Nikunen and T. Virtanen, "*Sound event localization and detection of overlapping sources using convolutional recurrent neural networks*", IEEE JSTSP, 2019

Results

ADRENALINE outperforms baseline methods in terms of DoA error and yields comparable frame recall

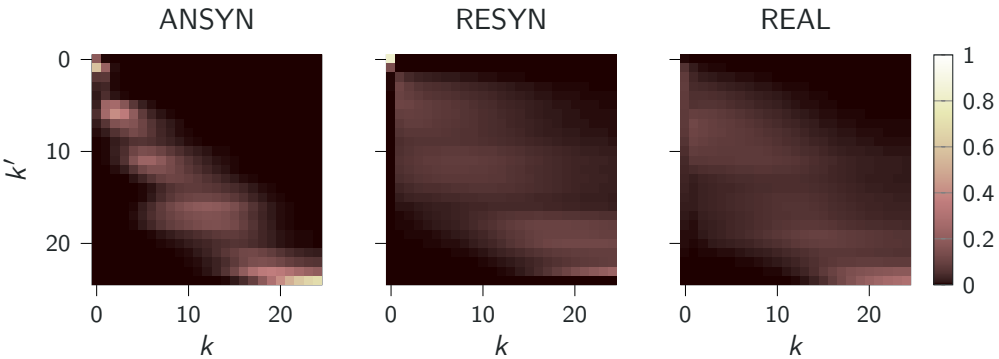




Frame recall	ANSYN	RESYN	REAL
CNN	**87.48**	71.91	72.07
SELDNet(m)	85.78	**72.46**	69.63
ADRENALINE	84.83	71.18	**72.08**

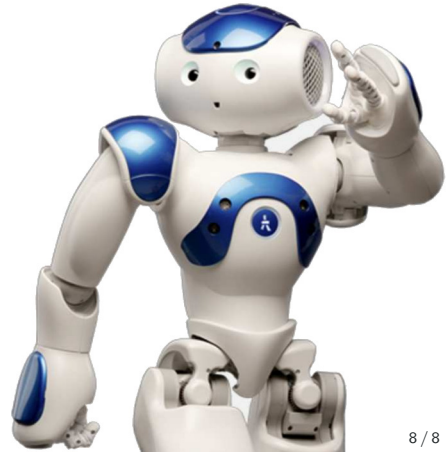
What can the attentions tell us?

Reverberant environments force the model to utilize larger temporal context for SEL



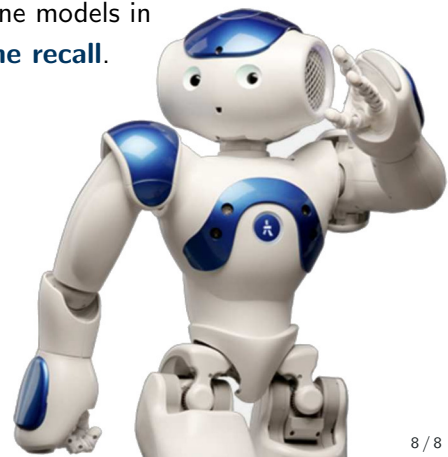
Conclusions

- An **attention-based sequence-to-sequence model** (ADRENALINE) was proposed for multi-source sound event localization tasks.



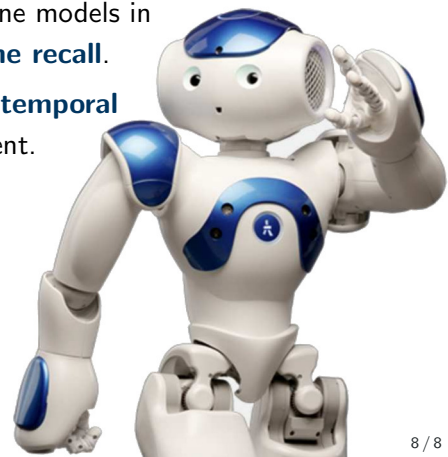
Conclusions

- An **attention-based sequence-to-sequence model** (ADRENALINE) was proposed for multi-source sound event localization tasks.
- The proposed framework **outperformed** the baseline models in terms of **DoA error** and yielded **comparable frame recall**.



Conclusions

- An **attention-based sequence-to-sequence model** (ADRENALINE) was proposed for multi-source sound event localization tasks.
- The proposed framework **outperformed** the baseline models in terms of **DoA error** and yielded **comparable frame recall**.
- Attentions enable the model to **adapt** the utilized **temporal context size** depending on the acoustic environment.



Conclusions

- An **attention-based sequence-to-sequence model** (ADRENALINE) was proposed for multi-source sound event localization tasks.
- The proposed framework **outperformed** the baseline models in terms of **DoA error** and yielded **comparable frame recall**.
- Attentions enable the model to **adapt** the utilized **temporal context size** depending on the acoustic environment.

Thank you for your attention!