PILOT: Introducing Transformers for Probabilistic Sound Event Localization

INTERSPEECH 2021

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

30 August - 3 September 2021

RUHR
UNIVERSITÄT
BOCHUM

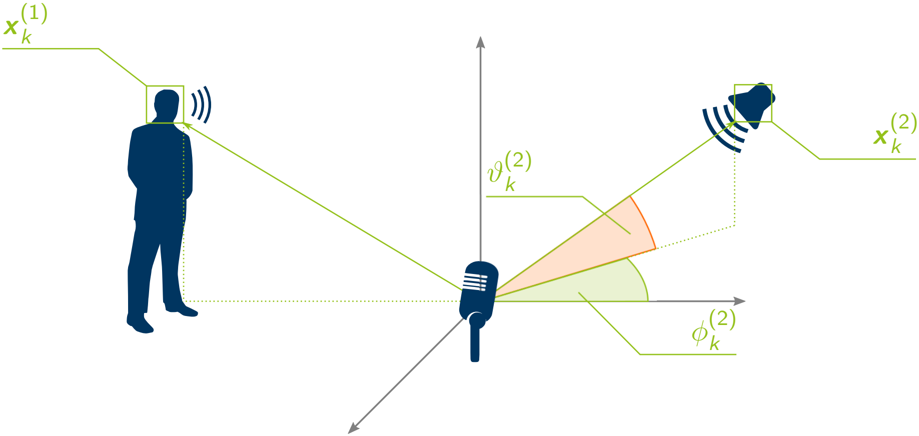
RUB

NTT

30 August – 3 September 2021
INTERSPEECH
2021 BRNO | CZECHIA
Speech everywhere!

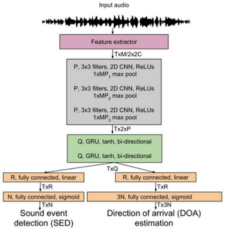
Problem statement

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

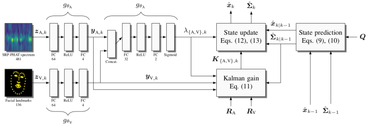


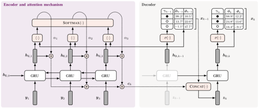
A selection of previous approaches

Different approaches to SEL have been proposed, primarily focusing on recurrent model architectures



Input audio
Feature extractor
TxM/2x2C
P, 3x3 filters, 2D CNN, ReLUs
1xMP$_1$ max pool
P, 3x3 filters, 2D CNN, ReLUs
1xMP$_2$ max pool
P, 3x3 filters, 2D CNN, ReLUs
1xMP$_3$ max pool
Tx2xP
Q, GRU, tanh, bi-directional
Q, GRU, tanh, bi-directional
TxQ
R, fully connected, linear
R, fully connected, linear
TxR
TxR
N, fully connected, sigmoid
3N, fully connected, sigmoid
TxN
Tx3N
Sound event detection (SED)
Direction of arrival (DOA) estimation





SELDNet[1]

Based on RNNs

Not probabilistic

DSW-BPKF[2]

Based on Kalman Filters

Probabilistic

ADRENALINE[3]

Based on RNNs+Attn

Not probabilistic

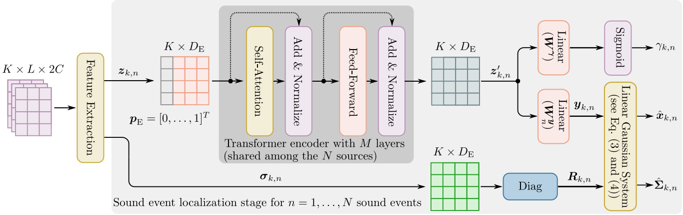
[1] Adavanne et al.: "Sound event localization and detection of overlapping sources using convolutional recurrent neural network" (JSTSP 2018)

[2] Schymura et al.: "A Dynamic Stream Weight Backprop Kalman Filter for Audiovisual Speaker Tracking" (ICASSP 2020)

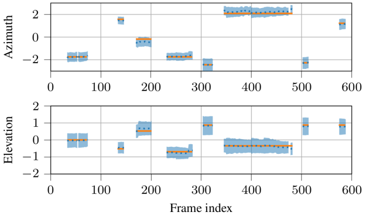
[3] Schymura et al.: "Exploiting Attention-based Sequence-to-Sequence Architectures for Sound Event Localization" (EUSIPCO 2020)

Proposed framework: PILOT[4]

PILOT is a transformer-based SEL system without recurrent structures and probabilistic output stage



$K \times L \times 2C$
Feature Extraction
$\boldsymbol{z}_{k,n}$
$K \times D_{\mathrm{E}}$
$\boldsymbol{p}_{\mathrm{E}} = [0, \dots, 1]^T$
Self-Attention
Add & Normalize
Feed-Forward
Add & Normalize
Transformer encoder with M layers (shared among the N sources)
$K \times D_{\mathrm{E}}$
$\boldsymbol{z}'_{k,n}$
Linear ($\boldsymbol{W}^{\gamma}$)
Sigmoid
$\gamma_{k,n}$
Linear ($\boldsymbol{W}^{\boldsymbol{y}}_n$)
$\boldsymbol{y}_{k,n}$
Linear Gaussian System (see Eq. (3) and (4))
$\hat{\boldsymbol{x}}_{k,n}$
$\boldsymbol{\sigma}_{k,n}$
$K \times D_{\mathrm{E}}$
Diag
$\boldsymbol{R}_{k,n}$
$\hat{\boldsymbol{\Sigma}}_{k,n}$
Sound event localization stage for $n = 1, \dots, N$ sound events



- Multi-head attention instead of recurrent architectures
- Differentiable linear Gaussian system as output stage to represent uncertainty

[4] Probabilistic Localization of Sounds with Transformers

Key results

A transformer-based architecture shows superior SEL performance over recurrent models

- PILOT **consistently outperforms** CNN, modified SELDNet and ADRENALINE baseline systems **with statistically significant differences in DoA error** in both simulated and recorded acoustic conditions.
- PILOT also shows **improved frame recall** (percentage of correctly identified sound sources) compared to SELDNet and ADRENALINE.
- The **probabilistic output stage** allows the model to represent the **estimation reliability** associated with individual sound source DoA estimates.