# MULTI-SCALE NETWORK FOR SOUND EVENT LOCALIZATION AND DETECTION

Technical Report

*Patrick Emmanuel, Nathan Parrish, Mark Horton*

Johns Hopkins University Applied Physics Laboratory
11100 Johns Hopkins Road
Laurel, Maryland 20723-6099, USA
{patrick.emmanuel, nathan.parrish, mark.horton}@jhuapl.edu

## ABSTRACT

This report describes a multi-scale approach to the DCASE 2021 Sound Event Localization and Detection with Directional Interference task. The goal of this task is to detect, classify, and localize in time and space events from twelve sound event classes in varying reverberant acoustic environments in the presence of interfering sources. We train a network that jointly performs detection, localization, and classification using multi-channel magnitude spectral data and intensity vectors derived from first order ambisonics time-series. We implement a network with successive blocks of multi-scale filters to discriminate and extract overlapping classes with different spectral characteristics. We also implement an output format and permutation invariant training loss that enable the network to detect, classify, and localize multiple instances of the same class simultaneously. Experiments show that the proposed network outperforms the CRNN baseline networks in classification and localization metrics.

***Index Terms—*** sound event localization and detection, direction of arrival estimation, permutation-invariant training

## 1. INTRODUCTION

The goal of sound event localization and detection (SELD) is to detect, classify, and localize acoustic sources in time and direction of arrival (DOA) using recordings from a multi-channel microphone array. SELD can provide important perceptual input for a number of automation tasks including robotics, home assistant systems, and security applications. However, SELD is a challenging task in acoustic environments with varying impulse responses and reverberation, high background noise, and interfering acoustic sources.

The 2021 DCASE SELD task includes many aspects that make the task challenging. It requires systems to classify and localize 12 different target classes in a wide variety of room environments with different room impulse responses and background noise characteristics. It also contains many time periods where up to three target classes are present simultaneously along with directional interference sources that don't belong to any of the target classes. The target classes can be stationary or mobile, and occur at varying SNR levels from 6dB to 30dB. The recordings are provided as 4-channel microphone or first order ambisonics (FOA) data as measured by a spherical microphone array.

Historically, sound event detection and localization have been accomplished using signal processing and tracking algorithms such as the TRAMP algorithm that utilizes a voice-activity detector and pseudo-intensity measurements in a particle filter to detect and localize acoustic sources [1]. Recently, however, end-to-end trained convolutional recurrent neural networks (CRNN) have been shown to exceed the performance of traditional algorithms for sound event detection and localization [2, 3]. Additionally, end-to-end trained networks have been the top performing approaches to previous DCASE SELD tasks including an ensemble of CRNNs in 2019 [4] and an ensemble of multiple deep neural networks in 2020 [5].

This report describes an end-to-end trained deep network for jointly detecting, classifying, and localizing the acoustic target classes using the FOA data. The network takes as input log magnitude spectral representations of the acoustic time-series, along with corresponding intensity vector representations, and outputs class confidence and DOA estimates at 100 ms intervals. The network is capable of detecting, classifying, and localizing up to two instances of each class simultaneously. Due to the high degree of polyphony (multiple sound sources transmitting jointly) in the data, we design the network to operate at multiple time/frequency scales throughout and to carry this multi-scale operation through the network. Our intuition in doing this is that to detect, classify, and localize multiple sound sources simultaneously, the model must recognize spectral content at different scales for different classes, and then filter the corresponding frequencies in the intensity vector representation to estimate DOA.

In the following sections we describe the network architecture, loss function, and training procedure. Experimental results show that the multi-scale network achieves SELD metrics that outperform the baseline CRNN model on the DCASE 2021 test set, Fold 6, when trained on Folds 1 - 5.

## 2. MODEL DESCRIPTION

Figure 1 shows the model architecture. The log-spectral and intensity vector inputs are processed through several successive layers of neural architecture search (NAS) [6] and pyramid scene parsing blocks [7] before being processed by a multi-headed self-attention layer (MHSA) [8] that outputs to three parallel dense layers to predict output detections and DOA coordinates for each class in cartesian coordinates. The output tensors have dimension 40x12x2, representing 40 time samples (for 4 seconds of data with 100 ms resolution), twelve classes, and up to two instances of the same class simultaneously. The output structure performs detection, classification, and DOA estimation jointly by using the activity-coupled cartesian DOA (ACCDOA) output vector proposed by Shimada, et al. [9]. ACCDOA uses the magnitude of the

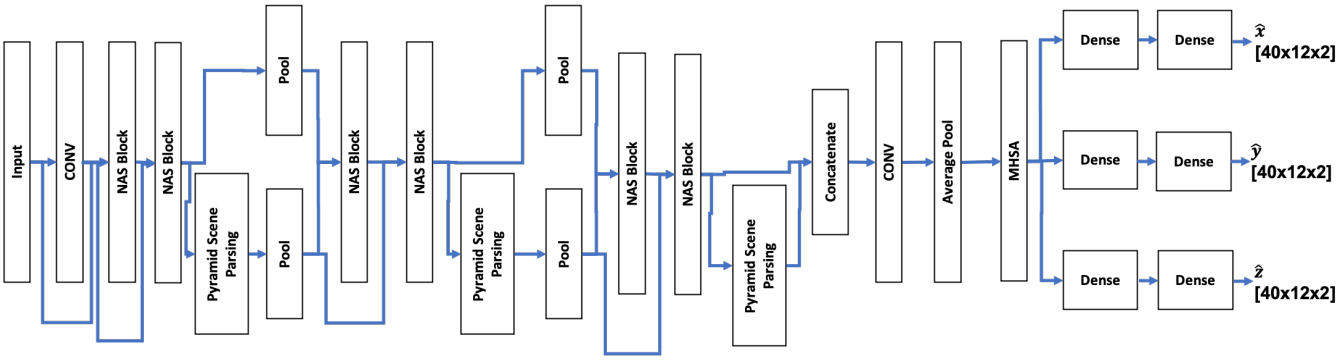


Figure 1: SELD Model Architecture.

$[\hat{x}(t,c,i), \hat{y}(t,c,i), \hat{z}(t,c,i)]$ vector to indicate the presence or absence of an acoustic source at a particular time ($t$), class ($c$), index ($i$) tuple, and the vector direction gives the DOA.

In the following subsections, we provide additional details on the input representation and data augmentation, on the network architecture, and finally on the loss function and training optimization.

### 2.1. Input Representation

The network input is derived from the four-channel FOA format data sampled at 24 kHz. We include log-magnitude spectral representations for each FOA channel and accompanying $x, y, z$ components of the corresponding intensity vector. We train networks using either constant-Q or logmel spectral representation with associated intensity vectors.

Constant-Q is time-frequency representation where the ratio of center frequency to bandwidth of the bins is kept constant. We utilize the constant-Q implementation in Librosa [10] with a minimum frequency of 50 Hz, 48 bins-per-octave, a hop length of 576, and 379 bins (corresponding to a maximum frequency of 11.74 kHz). Our network processes 4 second chunks of data, resulting in 167 time scans in the input representation. We also compute an intensity vector for each bin in the constant-Q input surface. Define the complex constant-Q data for the omni, x, y, and z channels of the FOA data at time $t$ and frequency $f$ as $o_{cq}(t,f), x_{cq}(t,f), y_{cq}(t,f), z_{cq}(t,f)$. We can compute the corresponding $x, y, z$ components of the intensity vector as [3]:

$$
\begin{aligned}
i_x(t,f) &= \text{Real}\left\{o^*_{cq}(t,f)x_{cq}(t,f)\right\}, \\
i_y(t,f) &= \text{Real}\left\{o^*_{cq}(t,f)y_{cq}(t,f)\right\}, \\
i_z(t,f) &= \text{Real}\left\{o^*_{cq}(t,f)z_{cq}(t,f)\right\},
\end{aligned}
$$

where $o^*_{cq}(t,f)$ denotes complex conjugate of the omni-channel data.

We normalize the log-magnitude of the constant-Q surfaces by clipping to a range of $[-25, -3]$ and then scale the values to lie between 0 and 1. We normalize the intensity data by computing the log-magnitude, clipping it to a range of $[-50, 6]$, scaling it to lie between 0 and 1, and then re-applying the sign of the original intensity bin. Examples of the resulting normalized omni-channel and intensity vector representations are shown in Figure 2.

For the logmel input, we use the same processing parameters and intensity vector representation as Cao, et al. [11]. This includes segmenting the data into 4-second segments and computing a logmel spectral and intensity representation with 256 bins and 160 time scans.

### 2.2. Data Augmentation

Previous DCASE SELD competitions have shown that data augmentation is critical for achieving good generalization performance [12]. We perform wav mixing and rotation on FOA time-series data prior to generating logmel or normalized constant-Q spectral products, where we then apply frequency and time masking [13] before passing the augmented inputs to our models.

We perform rotation augmentation by first transforming the coordinates of the truth data from azimuth, elevation to cartesian x, y, z coordinates. We then apply a random rotation matrix to the truth and the x, y, z channels of the FOA time-series. This is comparable to the labels first rotation method proposed in [14].

### 2.3. Network Architecture

The network consists of two primary modules. A Convolutional network feature extractor and a Multi-Head-Self-Attention (MHSA) block that operates on the output of the CNN feature extractor. To facilitate learning from signals with potentially varying spectral characteristics we use NASnet-like [6] convolution modules, termed NAS blocks in Figure 1, along with convolution modules inspired by PSPNet [7], termed pyramid scene parsing blocks in Figure 1. We anticipate that these multi-scale modules will enable the CNN to extract features at varying scales in the data, allowing for potentially better generalization to acoustic targets with varying spectral characteristics and bandwidths.

The DCASE 2021 SELD task contains many time periods where multiple instances of the same class are present simultaneously at different DOAs. Therefore, we implement an output representation that allows the network to classify up to two instances of the same class in each time scan. The network is made to produce outputs for cartesian x, y, and z coordinates for up to two instances of the same class simultaneously. We use the simplifying ACCDOA representation to predict classification and DOA with a single model [9]. Due to the ACCDOA representation we use only a single network to jointly estimate cartesian DOA and class labels without an explicit classification loss function.

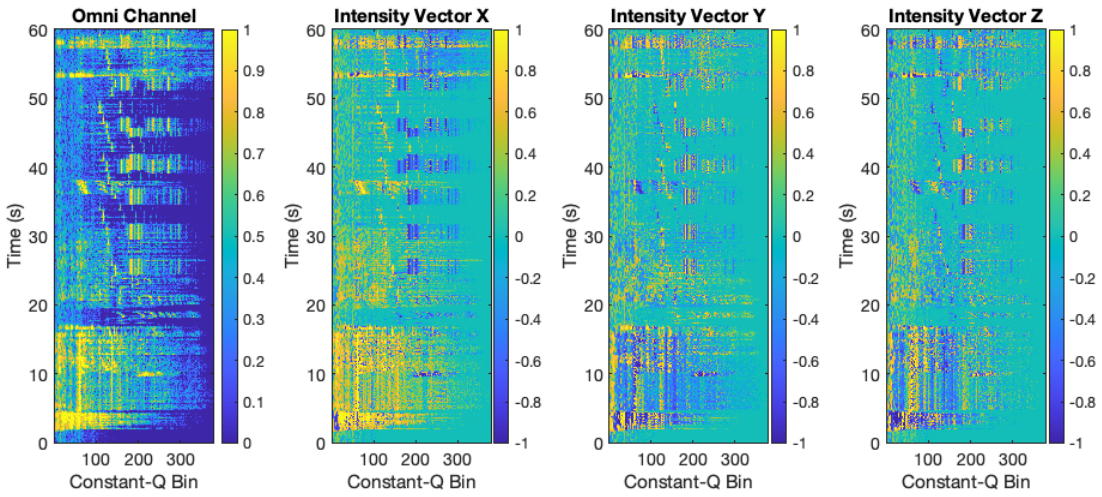


Figure 2: A 60-second sample of the normalized constant-Q representation of the FOA Omni channel and normalized intensity vector surfaces for one of the DCASE 2021 training files.

### 2.4. Network Training

We train models with varying sized MHSA heads and with constant-Q or logmel spectral representations as input. We use the AdamW optimizer [15] with a warmup-cooldown schedule as in [8]. The learning rate is warmed up for the first 20K/30K steps (depending on the model) of training before cooling down for the remainder of the training steps. In total we train models for about 187,500 steps. We perform random combinations of the augmentations given in Section 2.2 and generate the resulting logmel and constant-Q spectra on-the-fly to facilitate training with a maximum variety of data. We optimize all of our models with a permutation-invariant version of mean squared error similar to that first proposed in [16]. The permutation invariance is applied along the instance dimension of the $\hat{x}, \hat{y}, \hat{z}$ outputs jointly to account for the ambiguity in assigning a detection to an instance.

## 3. EXPERIMENTAL SETUP AND RESULTS

We utilize the DCASE 2021 FOA data to train and evaluate several networks for comparison to the DCASE 2021 baseline. We train on Folds 1 - 5 and evaluate on Fold 6. Each fold contains 100 one-minute recordings with multiple overlapping sound events. Each fold contains contains data from room environments with different impulse response and background noise characteristics. The time-series recordings also contain random instances of interferers that do not belong to any class. More details of the data can be found on the DCASE 2021 SELD challenge page [12].

We evaluate the models using the metrics specified for the DCASE 2021 challenge. These include the error and F-Score at 20 degrees denoted $\mathrm{ER}_{20^\circ}$ and $\mathrm{F}_{20^\circ}$ and the classification dependent localization error and localization recall denoted $\mathrm{LE}_{\mathrm{CD}}$ and $\mathrm{LR}_{\mathrm{CD}}$. $\mathrm{ER}_{20^\circ}$ and $\mathrm{F}_{20^\circ}$ compute classification error rate and F-Score on classifications localized to within $20^\circ$ of the true DOA. $\mathrm{LE}_{\mathrm{CD}}$ computes the localization error in degrees between truth and estimates of the same class, and $\mathrm{LR}_{\mathrm{CD}}$ computes class-based recall. These metrics are described in further detail in [17].

Table 1 gives metrics for the DCASE 2021 FOA Baseline network [12] and the proposed multi-scale network trained with the logmel and constant-Q inputs. Results show that the proposed approach outperforms the baseline network in all metrics, with significantly better $\mathrm{F}_{20^\circ}$ and $\mathrm{LR}_{\mathrm{CD}}$.

| **Network** | $\mathrm{ER}_{20^\circ}$ | $\mathrm{F}_{20^\circ}$ | $\mathrm{LE}_{\mathrm{CD}}$ | $\mathrm{LR}_{\mathrm{CD}}$ |
|---|---|---|---|---|
| DCASE 2021 FOA Baseline | 0.69 | 33.9 % | $24.1^\circ$ | 43.9 % |
| DCASE 2021 MIC Baseline | 0.74 | 24.7 % | $30.9^\circ$ | 38.2 % |
| Multi-Scale logmel | 0.65 | 57.1 % | $17.5^\circ$ | 62.8 % |
| Multi-Scale constant-Q | 0.63 | 54.3 % | $18.2^\circ$ | 55.3 % |

Table 1: SELD metrics for four networks when trained on Folds 1-5 and evaluated on Fold 6.

## 4. CONCLUSION

We have presented a multi-scale network for detecting, classifying, and localizing acoustic targets and have applied it to the DCASE 2021 SELD task. The output structure enables the network to classify multiply instances of the same target class simultaneously. Experiments show that the proposed network achieves improved performance in all metric categories when compared to the baseline model for the 2021 DCASE SELD task.

## 5. REFERENCES

[1] S. Kitić and A. Guérin, "Tramp: Tracking by a´ real-time ambisonic-based particle filter," *arXiv preprint arXiv:1810.04080*, 2018.

[2] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *2018 26th European Signal Pro-*

*cessing Conference (EUSIPCO)*. IEEE, 2018, pp. 1462–1466.

[3] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "Crnn-based multiple doa estimation using acoustic intensity features for ambisonics recordings," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 22–33, 2019.

[4] S. Kapka and M. Lewandowski, "Sound source detection, localization and classification using consecutive ensemble of crnn models," *arXiv preprint arXiv:1908.00766*, 2019.

[5] Q. Wang, H. Wu, Z. Jing, F. Ma, Y. Fang, Y. Wang, T. Chen, J. Pan, J. Du, and C.-H. Lee, "The ustc-iflytek system for sound event localization and detection of dcase2020 challenge," DCASE2020 Challenge, Tech. Rep, Tech. Rep., 2020.

[6] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," *arXiv preprint arXiv:1611.01578*, 2016.

[7] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[9] K. Shimada, N. Takahashi, S. Takahashi, and Y. Mitsufuji, "Sound event localization and detection using activity-coupled cartesian doa vector and rd3net," *arXiv preprint arXiv:2006.12014*, 2020.

[10] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8. Citeseer, 2015, pp. 18–25.

[11] Y. Cao, T. Iqbal, Q. Kong, Z. Yue, W. Wang, and M. D. Plumbley, "Event-independent network for polyphonic sound event localization and detection," DCASE2020 Challenge, Tech. Rep., July 2020.

[12] http://dcase.community/challenge2021/.

[13] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[14] L. Mazzon, Y. Koizumi, M. Yasuda, and N. Harada, "First order ambisonics domain spatial augmentation for dnn-based direction of arrival estimation," *arXiv preprint arXiv:1910.04388*, 2019.

[15] I. Loshchilov and F. Hutter, "Fixing weight decay regularization in adam," 2018.

[16] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 241–245.

[17] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen, "Overview and evaluation of sound event localization and detection in dcase 2019," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020.