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
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 data sampled at 24 kHz. We include log-magnitude spectral representation.

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 ACC-DOA 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 1: SELD Model Architecture.
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 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 \( \text{ER}_{20} \) and \( \text{F}_{20} \), and the classification dependent localization error and localization recall denoted \( \text{LE}_{CD} \) and \( \text{LR}_{CD} \). \( \text{ER}_{20} \) and \( \text{F}_{20} \) compute classification error rate and F-Score on classifications localized to within 20° of the true DOA. \( \text{LE}_{CD} \) computes the localization error in degrees between truth and estimates of the same class, and \( \text{LR}_{CD} \) 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 \( \text{F}_{20} \) and \( \text{LR}_{CD} \).

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


