ICASSP 2022 L3DAS22 CHALLENGE: ENSEMBLE OF RESNET-CONFORMERS WITH AMBISONICS DATA AUGMENTATION FOR SOUND EVENT LOCALIZATION AND DETECTION

Yongjian Mao, Ying Zeng, Hongqing Liu, Wenbin Zhu, and Yi Zhou

School of Communication and Information Engineering,
Chongqing University of Posts and Telecommunications, Chongqing, China

ABSTRACT

It remains a tough challenge to tackle sound event localization and detection (SELD) problem, especially when sound scene complexity increases and overlapping acoustic sources appear. To improve the SELD performance, we propose an ensemble system, which consists of a ResNet and Conformer backbone network (SELD-RCnet) and its two variants, SED-RCnet and SSL-RCnet. For SELD-RCnet and SSL-RCnet, we use short time Fourier transform (STFT) magnitude spectrogram, phase spectrogram, and active acoustic intensity vectors (IVs) as input features. For SSL-RCnet, an innovative predictive target is also developed and the performance is thus improved. For SED-RCnet, we use Log-Mel spectrogram as input features. To overcome the lack of training data, we adopt two novel approaches to first order Ambisonic (FOA) format dataset augmentation, namely audio channel swapping (ACS) and time-frequency masking (TFM). Finally, in the L3DAS22 Challenge, our submitted system achieves significant improvements over the baseline and ranks the second place for the Task2. Therefore, according to the competition rules, we submit this work to describe our system in details.

Index Terms— Sound event detection, sound source localization, ambisonics, conformer

1. INTRODUCTION

Sound event localization and detection (SELD) is a task to detect the occurrences of sound events and localize them even when multiple events overlap both temporally and spatially. This joint mission is different from classical sound source localization (SSL) which only focuses on the spatial information of sources. It also differs from sound event detection (SED) task which only considers the types of events. SELD also has a wide range of applications in machine listening, acoustic scene analysis, and audio surveillance in intelligent home and cities [1, 2].

The previous SED methods [3] were mostly based on the Gaussian mixture models-hidden Markov models (GMM-HMM) and the non-negative matrix factorization (NMF) [4, 5] technique. However, these methods are not robust and satisfactory when sound scene complexity increases and overlapping sound sources appear. For the last decade, deep neural networks (DNNs) have been employed in this domain and the convolutional recurrent neural network (CRNN) [6] achieved the state-of-the-art results for SED problem.

The purpose of SSL is to estimate direction-of-arrival (DOA) for every sound source. In terms of DOA estimation for multiple sources, several methods have been developed using traditional array signal processing techniques, including the multiple signal classification (MUSIC) [7], the estimation of signal parameters via rotational invariance technique (ESPRIT) [8], and the steered response power phase transform (SRP-PHAT) [9]. In the presence of noise and reverberation, the accuracy of all the aforementioned methods decreases dramatically [10]. Recently, a series of DNNs-based methods [11, 12] have improved the robustness of DOA estimation in adverse scenes and achieved better results than traditional methods.

Ambisonics, as a multichannel spatial audio format, is also popular in DOA estimation domain. It is a method of codifying a sound field by considering its directional properties and spatial information. In real-life applications, we tend to use a microphone array with a limited amount of channels. For example, in B-format Ambisonics signal, 4 channels are used. In addition, it has been shown that the first-order Ambisonics (FOA) signal works better than high-order Ambisonics (HOA) signal in DOA estimation task [13]. The property of Ambisonics directly codifying sound field and spatial information makes it popular in DOA estimation task.

However, the aforementioned studies treated the tasks of SED and SSL separately, not trying to jointly tackle them even though it is very meaningful to do so. The recent studies show that DNNs-based SELD methods are popular by considering both tasks simultaneously. The SELD was presented for the first time in the DCASE2019 Challenge [14] and the SELDnet [15] was the baseline. As the first attempt, the SELDnet reached the state-of-the-art result, where CNNs extract spatial information from inputs and RNNs estimate the outputs. Besides, acoustic intensity vectors (IVs), which are elements of middle and high frequency positioning, are also

extracted as input features in [10] and further improvement was achieved.

For a more challenging acoustic scene in the ICASSP 2022 L3DAS22 Challenge [16], where multiple overlapping acoustic sound sources appear, we propose a novel SELD system, called SELD-RCnet, to improve the performance of SELD. In SELD-RCnet, the ResNet and Conformer as backbone model are constructed. For the proposed SELD-RCnet, a deep residual network replaces CNNs and the Conformer module is utilized instead of conventional bidirectional GRUs in SELDnet. Moreover, to further improve the performance, we ensemble SELD-RCnet and its two variants, SED-RCnet and SSL-RCne. Additionally, with the use of audio channel swapping (ACS) and time-frequency masking (TFM), we are able to augmentate the FOA audio data and increase DOA labels significantly.

The remainder of the paper is structured as follows. In Section 2, the whole system is introduced in details. In Section 3, the experimental settings are given. The final results of our submitted system are described in Section 4. Finally, the conclusions are drawn in Section 5.

2. PROPOSED METHOD

2.1. Input features

There are totally two combinations of input features in our system. For SSL target , we use 11 channels of input features with first eight for STFT magnitude and phase spectrograms and last three for acoustic IVs. According to psycho-acoustic based spatial audio theory [17], the sound DOA can be estimated as the opposite direction of the IVs, which is widely used in sound field localization [18]. For SED target, we use 4 channels of Log-Mel spectrograms since SED mainly relies on magnitude information. The Log-Mel spectrogram retains the feature information similar to that in human's listening, and also reduces the number of parameters.

2.2. ResNet-Conformer network

Inspired by [6], we utilize the ResNet-Conformer network (SELD-RCnet) as the backbone model, which integrates ResNet and Conformer. An overview of the system is depicted in Fig. 1(a). For SELD problem, we focus on both time and frequency dimensions. We take the similar ResNet block in the framework of speaker encoder in [19], but make a use of attention mechanism instead of CNNs block in baseline for catching local fine-grained features, extracting high dimension information, and improving the performance of system.

Given the input with size (C, 256, 2400), the output of ResNet Layers is reshaped to size (300, 256), where C is the number of channels of input features. After that, two consecutive Conformer blocks learn both the local features and temporal context information and output a feature with the shape of (300, 256). The Conformer used depth-wise separable convolution to improve the performance of transformer [20]. The convolution module is able to exploit local features compared to the traditional transformer. Due to the fact that Conformer simultaneously models both local and long-range global context dependencies for audio sequence, it outperforms RNNs or transformer in some tasks. After Conformer blocks, two adjacent fully connected (FC) layers map the features into final SED and SSL representations.

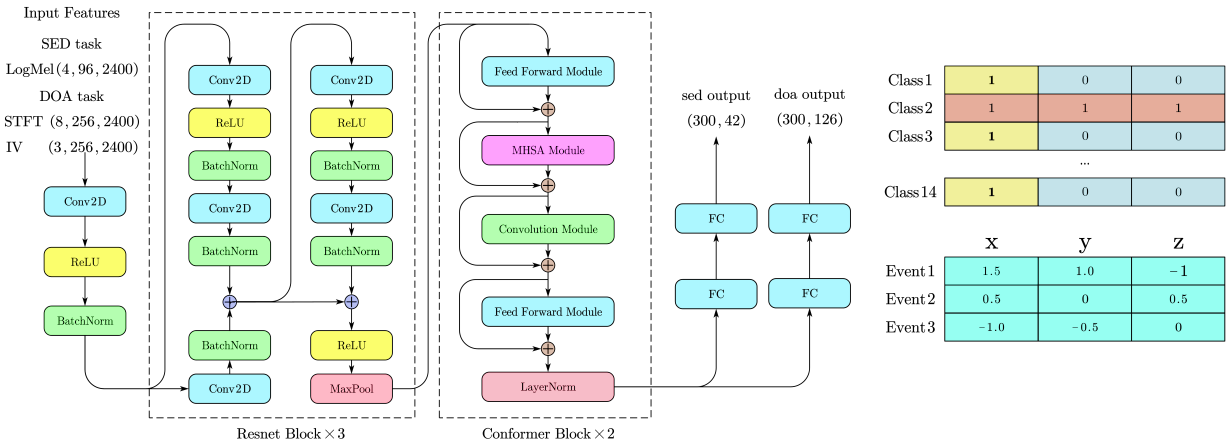
For SED target, the model predicts a matrix of the shape (300, 42) with the value in the range of [0,1], which represents the status of 14 different events in 300 time frames. There values are thresholded to map the SED output to true or false values, the same as labels, which indicates the sound event is active or inactive at the frame, respectively. For example, as shown in Fig. 1(b), in the vector of second class, the outputs with three true values denote there are three similar events happening at the same time. The other case is shown in the first column, where three true values in boldface denote there are three different types of events at the same time. Notice that there will only be up to three values hold true at a time.

For SSL target, same as baseline [16], the model's output matrix is three times the size of SED target in order to predict separate location and detection information for all possible simultaneous sounds of the same class.

2.3. Model variants and ensemble

The SELD-RCnet, which is similar to the baseline model SELDnet, has two FC branches at the end and needs a joint optimization. To avoid the difficulties brought by joint task, and to improve the robustness, model ensemble technique is utilized. To that aim, we adjust the model structure to construct two model variants, SED-RCnet and SSL-RCnet. As depicted in Fig. 2, the difference is mainly reflected in the input features and prediction targets used. For SED-RCnet, it only predicts SED representation and takes Log-Mel spectrograms as input features. For SSL-RCnet, it only predicts SSL representation and takes STFT magnitude, phase spectrograms, and IVs as input features.

For SSL-RCnet, the innovation is that the model only predicts three sets of Cartesian coordinates with the shape of (300,9) in the order of active events based on the prior knowledge that there are at most three events at a time. If only single event happens in a time frame, then the last two coordinate vectors are all zeros. This design makes the training easier and improves the performance. During the experiments, for SELD-RCnet, we find that baseline's outputs usually produce too many small and DOA-invalid predictions which confuse the DOA estimation. We also have a try to mask invalid DOA predictions with zeros based on SED predictions, but the accuracy is dependent on SED results strongly. In addition, the SED performance degrades dramatically if we adopt the similar strategy of SSL-RCnet in SELD-RCnet. The reason is that



(a) ResNet-Conformer architecture (b) Output representation

Fig. 1. An overview of the ResNet-Conformer model.

those DOA-invalid values close to zeros can guide estimation of SED in each position of event in joint SELD network, though introducing the difficulties to estimate DOA.

Three models are trained separately and the ensemble is finally utilized for predicting the final target.

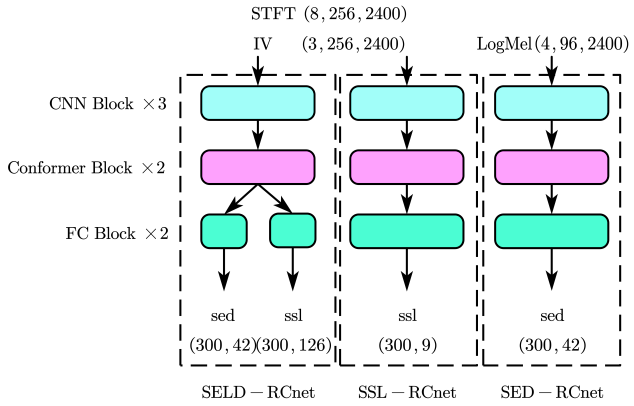


Fig. 2. The architectures of SELD-RCnet, SSL-RCnet, and SED-RCnet.

2.4. Loss function

The following strategy is applied to train the network. We train the SED-RCnet with binary cross entropy (BCE) loss criterion and SSL-RCnet with mean square error (MSE) loss criterion. For SELD-RCnet, the weighted loss functions are used. For the sake of simplicity of instructions, if weight is 0, it means that the corresponding branch does not exist, that is, it is a single-objective SED-RCnet or SSL-RCnet.

3. EXPERIMENTS

3.1. Dataset and augmentation

The whole system is evaluated using the L3DAS22 dataset [16] with 7.5 hours of FOA audio, where totally 14 transient classes are to be detected. When multiple acoustic events are active at the same time, with an approximate probability of 12% at least 2 sounds may belong to the same class. Although the official dual microphone recording data are available, we only use single microphone recording data, because it is a more realistic scenario in practice.

However, it is still challenging to train our system with such a small data size. Note that it is also difficult to obtain the similar sound event data and synthesize more data. To overcome the lack of training data and to alleviate potential overfitting, we take two approaches proved to be suitable and effective for FOA dataset augmentation, namely ACS and TFM [6].

The ACS utilizes the sound field rotational property of Ambisonics. For example, assuming there is a sound field rotation (ϕ=ϕ+π/2, θ=-θ), the spherical harmonics vector $\mathbf{Y}(\theta, \phi)$ in encoding matrix can be recalculated as

$$\boldsymbol{Y}_{new}(\theta, \phi) = \frac{1}{\sqrt{4\pi}} \begin{bmatrix} 1 \\ \sqrt{3}\cos\phi\cos\theta \\ -\sqrt{3}\sin\theta \\ -\sqrt{3}\sin\phi\cos\theta \end{bmatrix}. \quad (1)$$

It can be found that 90-degree rotation and mirroring in azimuth and elevation correspond to sign inversion, channel swapping, or both. Accordingly, we can generate a larger and more diverse dataset with more different DOA labels by transforming original DOA labels and limiting them in the domain of azimuth $\phi \in [-180°, 180°]$ and elevation $\theta \in$

$[-90°, 90°]$. Note that the transformation between spherical and Cartesian coordinates is needed for original target data.

In addition to that, we also adopt TFM, which randomly masks consecutive time frames or frequency bands of input features [6] to enhance model robustness. In this experiment, the lengths of mask are 200 and 50 in the time and the frequency dimensions, respectively.

3.2. Implementations

For each FOA audio file with 4 channels, 32 kHz sampling rate, and 30 seconds duration, we extract 256 STFT magnitude and phase spectrogram, and 256 active acoustic IVs, using 512-point FFT and Hamming window with 112-point overlap. After that, 96 log-Mel magnitude spectrograms are obtained with the use of 96 Mel-filter bank. Finally, we obtain input features with the shapes of (4, 96, 2400) for SED-RCnet and (11, 256, 2400) for SELD-RCnet and SSL-RCnet.

The filter sizes in the first bottleneck Conv2D layer and three ResNet blocks are {64,64,128,128}. Due to the different shapes of the input features, for SELD-RCnet and SSL-RCnet, the pooling sizes in ResNet blocks are {{8,2},{8,2},{2,2}}, and for SED-RCnet, the pooling sizes are {{6,2},{4,2},{2,2}}.

Two consecutive Conformer blocks with the configurations of 8 heads and 256 attention dimension are utilized. The remaining parameters in Conformer follows the suggestions in [20]. The system is trained by the Adam optimizer with a learning rate of 0.0001. The batch size is 20, and dropout and early stopping are employed, where the experienced dropout rate is 0.1 and training is stopped if there are no improvements on evaluation set after 70 epochs.

3.3. Model ensemble

The details of ensemble are as follows. First, for SSL-RCnet, ACS and TFM efficiently increase DOA representations, and the model predicts three sets of coordinates in the order of events. Second, we only take SED outputs from SELD-RCnet trained with and without ACS, because the precise DOA is difficult to estimate based on true positive SED predictions. The third SED output is obtained from SELD-RCnet. Finally, we identified the simple average of the outputs predicted by different models produces the best performance after detailed experiments. The number of parameters of three models are 9.6M, 9.32M, and 9.28M, respectively. The remaining detail information is shown in Table 1.

Table 1. Ensemble of Models

Model	*Output*	*LossWeights*[1]	*DataAug*
SELD-RCnet	SED1	(1,5)	ACS;TFM
SELD-RCnet	SED2	(1,5)	TFM
SED-RCnet	SED3	(1,0)	TFM
SSL-RCnet	SSL	(0,1)	ACS;TFM

[1]SED and SSL loss weights in training.

4. RESULTS AND ANALYSIS

4.1. Evaluation metrics

We evaluate our system using the T2 metric specified for the L3DAS22 Challenge [16]. It is a location-sensitive detection error and computed on each time frame. It measures the Cartesian distance between the predicted and true events with the same label, and counts a true positive only when its label is correct and its location is within a threshold from its reference location. In this Challenge, the spatial error threshold is 2 meters. The T2 metric is in the range [0,1], where the higher the value is, the better results the system produces.

Table 2. Subjective evaluation on the L3DAS22 Challenge

Methods	*T2 Metric*	*Precision*	*Recall*
SELDnet(baseline)	0.343	0.423	0.289
Proposed system	**0.592**	**0.600**	**0.584**

4.2. Subjective results in the L3DAS22 Challenge

In Table 2, the subjective results of the submissions are presented, which is provided by the Challenge organizer. Our proposed system reaches an T2 metric of 0.592 for Task2, with 0.600 precision and 0.584 recall, which outperforms the baseline model by overall 0.25 T2 metric.

5. CONCLUSIONS

In this Challenge, we propose a system for SELD problem, which consists of a SELD-RCnet backbone network and its two variants, SED-RCnet and SSL-RCnet. For SED target, we take Log-Mel spectrograms as input features, and the output is active status of each class of possible events. For SSL target, we take magnitude, phase spectrograms, and IVs as input features, and in SSL-RCnet, the outputs are Cartesian coordinates in the order of three possible events. The ensemble of three models greatly improves the performance of the system and overcomes the disadvantages of a single model. To avoid the potential risk of overfitting, we utilize ACS and TFM to increase DOA representations. The subjective results show that the proposed system outperforms the baseline and ranks the second place for the Task2 of the L3DAS22 Challenge.

6. REFERENCES

[1] Weipeng He, Petr Motlicek, and Jean-Marc Odobez, "Joint Localization and Classification of Multiple Sound Sources Using a Multi-task Neural Network," in *Proc. Interspeech 2018*, 2018, pp. 312–316.

[2] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*, 2007, pp. 21–26.

[3] J. Schröder, B. Cauchi, M.R. Schädler, N. Moritz, K. Adiloglu, J. Anemüller, S. Doclo, B. Kollmeier, and S. Goetze, "Acoustic event detection using signal enhancement and spectro-temporal feature extraction," in *Proc. IEEE AASP Challenge: Detection and Classification of Acoustic Scenes and Events*, 2013.

[4] Toni Heittola, Annamaria Mesaros, Tuomas Virtanen, and Antti Eronen, "Sound event detection in multisource environments using source separation," in *Machine Listening in Multisource Environments*. Citeseer, 2011.

[5] Satoshi Innami and Hiroyuki Kasai, "Nmf-based environmental sound source separation using time-variant gain features," *Computers and Mathematics with Applications*, vol. 64, no. 5, pp. 1333–1342, 2012, Advanced Technologies in Computer, Consumer and Control.

[6] Qing Wang, Jun Du, Hua-Xin Wu, Jia Pan, Feng Ma, and Chin-Hui Lee, "A four-stage data augmentation approach to resnet-conformer based acoustic modeling for sound event localization and detection," 2021.

[7] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[8] R. Roy and T. Kailath, "Esprit-estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 7, pp. 984–995, 1989.

[9] Hoang Do, Harvey F. Silverman, and Ying Yu, "A real-time srp-phat source location implementation using stochastic region contraction(src) on a large-aperture microphone array," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, 2007, vol. 1, pp. I–121–I–124.

[10] Lauréline Perotin, Romain Serizel, Emmanuel Vincent, and Alexandre 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.

[11] 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 Processing Conference (EUSIPCO)*, 2018, pp. 1462–1466.

[12] Daniele Salvati, Carlo Drioli, and Gian Luca Foresti, "Exploiting cnns for improving acoustic source localization in noisy and reverberant conditions," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 103–116, 2018.

[13] N. Poschadel, R. Hupke, S. Preihs, and J. Peissig, "Direction of arrival estimation of noisy speech using convolutional recurrent neural networks with higher-order ambisonics signals," *2021 29th European Signal Processing Conference (EUSIPCO)*, pp. 211–215, 2021.

[14] Archontis Politis, Annamaria Mesaros, Sharath Adavanne, Toni Heittola, and Tuomas Virtanen, "Overview and evaluation of sound event localization and detection in dcase 2019," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 684–698, 2021.

[15] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, 2019.

[16] E. Guizzo, C. Marinoni, M. Pennese, X. Ren, X. Zheng, C. Zhang, B. Masiero, and D. Comminiello, "L3DAS22 challenge: Learning 3D Audio Sources in a Real Office Environment," in *2022 IEEE International Conference on Acoucstic, Speech, and Signal Processing (ICASSP)*, 2022, pp. 1–6.

[17] Ville Pulkki, "Spatial sound reproduction with directional audio coding," *J. Audio Eng. Soc*, vol. 55, no. 6, pp. 503–516, 2007.

[18] Michela Ricciardi Celsi, Simone Scardapane, and Danilo Comminiello, "Quaternion neural networks for 3d sound source localization in reverberant environments," in *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2020, pp. 1–6.

[19] Hongqiang Du and Lei Xie, "Improving Robustness of One-Shot Voice Conversion with Deep Discriminative Speaker Encoder," in *Proc. Interspeech 2021*, 2021, pp. 1379–1383.

[20] A. Gulati, J. Qin, C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, et al., "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.