MLP-MIXER ENHANCED CRNN FOR SOUND EVENT LOCALIZATION AND DETECTION IN DCASE 2022 TASK 3

Technical Report

Shichao Wu[1,2,3]*, Shouwang Huang*[1,2,3]*, Zicheng Liu*[1,2,3]*, Jingtai Liu*[1,2,3]

[1] College of Artificial Intelligence, Nankai University, Tianjin, China.
[2] Institute of Robotics and Automatic Information System, Nankai University, Tianjin, China.
[3] Tianjin Key Laboratory of Intelligent Robotics, Nankai University, TianJin, China.
{wusc, 2013128, 2012178}@mail.nankai.edu.cn, liujt@nankai.edu.cn

ABSTRACT

In this technical report, we propose to give the system details about our MLP-Mixer enhanced convolutional recurrent neural networks (CRNN) submitted to the sound event localization and detection challenge in DCASE 2022. Specifically, we present two improvements concerning the input features and the model structures compared to the baseline methods. For the input feature design, we propose to involve the variable-Q transform (VQT) audio feature both for Ambisonic (FOA) and microphone array (MIC) audio representations. For deep neural network design, we improved the original CRNN by inserting a shallow MLP-Mixer module between the convolution filters and the recurrent layers to elaborately model the interchannel audio patterns, which we thought are extremely conducive to the sound directions of arrival (DOA) estimation. Experiments on the Sony-TAu Realistic Spatial Soundscapes 2022 (STARS22) benchmark dataset showed our system outperformed the DCASE baseline method.

Index Terms— Sound event localization and detection, DCASE, MLP-Mixer, convolutional recurrent neural networks (CRNN)

1. INTRODUCTION

Sound event localization and detection (SELD) [1] unifies the multiple channels of audio-based computational analysis tasks of sound event detection (SED) and sound source localization (SSL), which is also known as sound directions-of-arrival (DOA) estimation, with one method performed simultaneously. It means the sound event types and their locations in the spatial space are recognized with the sample audio inputs, even under the complex polyphonic [2] challenge circumstances. Both sound event detection and sound source localization has great promotion effects on the evolution of speaker identification and localization [3], acoustic information-based robot navigation [4, 5], and human-robot interaction with audio [6].

The SELD research has received a lot of study interest during the past years, owing to the yearly consecutive held Detection and Classification of Acoustic Scenes and Events (DCASE 13', 16'-22'). The sound event localization and detection challenge has been the subtask of the DCASE challenge held for the audio computational analysis community for many years [7]. The latest SELD challenge is the fourth iteration of the task in the DCASE challenge [8]. For DCASE2019 to DCASE2021, the SELD challenge has evolved from the static sound sources recognization to the moving sound sources recognization, which was based on the simulated multichannel recordings. The simulated recordings were generated with some selected sound event samples from the banks that were then convolved with the spatial room impulse responses collected in different rooms. To bring the task closer to more challenging real-world conditions, the simulated recordings were further added with spatial ambient noise, as well as the directional interferences in DCASE2021 [9]. In addition to the single sound event detection and localization, multiple sound events of the same type or different kinds overlapping in time and positions were also considered, since the polyphonic challenges are very common in the daily scenarios. Different from the simulated datasets used in DCASE2019 to DCASE2021, the dataset built for the SELD task in DCASE2022 was one real spatial sound scene recordings formatted dataset.

Research on SELD based on deep learning methods could be classified into three main aspects. One is to design audio features from the originally collected multiple channels of waveforms, and they were used as inputs for the deep neural network. Since the sound event detection recognization task is achieved mainly based on the time-frequency pattern difference between different types of sound events [10]. And the sound source localization recognization task is realized mainly based on the time difference of arrival of the sound impulse between microphones. Therefore, there is a tradeoff balance to designing optimized audio features that were both suitable for the sound event detection and localization tasks. The Mel-spectrogram, acoustic intensity vectors, generalize cross-correlation (GCC) sequences, SALSA spatial features [10] that are composed of multichannel log-spectrograms stacked channel-wise with the normalized principal eigenvectors of the spectrotemporal corresponding spatial covariance matrices, as well as the lite version of SALSA-lite [11] are commonly used experimentally demonstrated audio features for the SELD task. Deep learning methods used for SELD are the second research topic. The deep neural network that combined the convolutional and recurrent network (abbreviated as CRNN) presented by Adavanne et al. [1] was esteemed as the baseline approach for SELD in the audio research community. Other models from the computer vision area, i.e., transformer [12], ResNet [11], ensemble approaches [13] were also adopted for this task. Pre-processing and post-processing are the third research topic. Research on pre-processing mainly concentrates on how to augment the input audio features to increase its diversities, and a variety of data augmentation approaches have been demonstrated in the SELD task [14, 15]. Post-processing approaches have been studied to tackle the sound event detection and localization tasks

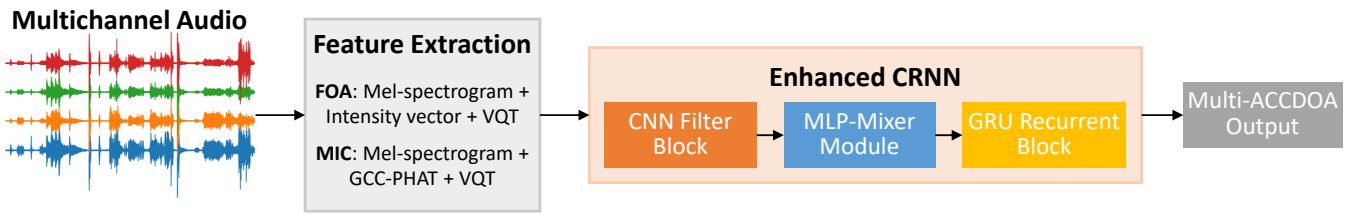

Figure 1: The overall framework of the enhanced CRNN network proposed in this work for the SELD task. For the audio feature extraction, we supplemented the VQT features from multichannel audio inputs. For the network design, we insert one MLP-Mixer module between the CNN filter block and the GRU recurrent bock.

simultaneously by converting the two branch outputs into a single one with the activity-coupled Cartesian DOA (ACCDOA) representation [13], as well as the improved version of Multi-ACCDOA [16] to detect the same event class from multiple locations.

In this report, we propose to adopt one new audio feature that is commonly used in speech recognization (which hasn't been studied for SELD as far as we know) and build an MLP-Mixer enhanced convolutional recurrent neural network for the sound event localization and detection challenge in DCASE2022. The rest of this report is organized as follows. In section 2, we give details on the proposed MLP-Mixer enhanced convolutional recurrent neural network. After that, we conduct experiments and discuss the results in section 3. Finally, we conclude this report in section 4.

2. THE PROPOSED METHOD

The proposed systems mainly contributed to two improvements compared to the baseline approach [8]. For the input features, we propose to use the variable-Q transform (VQT) audio feature in addition to the basic audio features, used in the baseline approach, both for Ambisonic (FOA) and microphone array (MIC) audio representations. For the deep neural network design, we propose to integrate one MLP-Mixer module between the convolutional and recurrent block in the CRNN model to model the interchannel audio patterns.

2.1. Input features

As the baseline method stated [8], for the FOA audio representation, the audio features of the Mel-spectrogram combined with the intensity vector are optimal for the SELD task. And for the MIC audio representation, the audio features of the Mel-spectrogram combined with the generalized cross-correlation (GCC) sequences are optimal. In this report, we propose to involve the variable-Q transform (VQT) audio feature. The VQT feature was computed based on every channel of waveforms. Thus, four-channel VQT features are computed with 1024-point FFTs using a Hanning window and 512-point hop length. It needs to be noted that the hop length set to compute VQT should be the power of 2, and it's not the same as that to compute the Mel-spectrogram, the intensity vector, as well as the GCC features. To combine the multiple types of audio features as an integrity input for the neural network, we padded/truncated the VQT feature to the same length as that of the basic features. Therefore, the input feature of each frame stacked across the channel dimension corresponding to FOA and MIC representations are $(4+3+4)\times 64$ features, and $(4+6+4)\times 64$ features, respectively.

2.2. Network architecture

The overall framework of the proposed enhanced CRNN network was shown in Fig. 1. In addition to the basic audio features, i.e. Mel-spectrogram, intensity vector for FOA, Mel-spectrogram, GCC-PHAT for MIC, we supplemented the VQT audio feature both for these two audio representations. The enhanced CRNN network was modified based on the baseline CRNN model [1, 8] with one MLP-Mixer module [17] inserted between the CNN filter block and the GRU recurrent block to capture the interchannel audio patterns. Since we believe these interchannel audio patterns play a vital role in sound source localization. The network details of the CNN filter block and the GRU recurrent block have remained the same as the baseline network. Therefore, the output feature shape of the CNN filter block is $(batch\_size \times 64 \times 50 \times 2)$. Consequently, we set the time and frequency size of the MLP-Mixer module as 1, and 2 to output the required formatted features for GRU recurrent block. Moreover, the designed MLP-Mixer module contains 8 MLP sublayers. At last, one dense layer was appended to output the Multi-ACCDOA representation.

3. EXPERIMENTAL RESULTS

3.1. Dataset and training settings

We train on the training set of the Sony-TAu Realistic Spatial Soundscapes 2022 (STARS22) [8] benchmark dataset and the external synthesized simulation data, that was used in DCASE2021 [9], as the baseline work did. Then, evaluate the network with the validation set of the STARS22 dataset, and report the corresponding detection and localization scores. We performed all experiments involved in this report on a single NVIDIA GeForce RTX3090 GPU. And the training data augmentation strategy of mixup [14] was used to increase the audio diversity. Other training details, such as the optimizer, and the learning rate schedule strategy have remained the same as the baseline work.

3.2. Results

We evaluate the models using the metrics specified for the DCASE2022 challenge, including the localization-dependent classification error and F-Score at 20 degrees, and the classification-dependent localization error and recall.

Table 1 shows the evaluation metrics of the baseline network [8] and the proposed enhanced CRNN (*E_CRNN*) with the VQT audio feature involved. It can be seen that the proposed systems outperformed the baseline approach both on the FOA and MIC audio representations concerning the sound event detection evaluation

Model		$ER_{20^\circ} \downarrow$	$F_{20^\circ} \uparrow$	$LE_{CD} \downarrow$	$LR_{CD} \uparrow$
Baseline	FOA	0.71	0.21	29.3°	0.46
	MIC	0.71	0.18	32.2°	0.47
E_CRNN	FOA	0.69	0.34	18.68°	0.45
	MIC	0.65	0.30	17.98°	0.44

Table 1: Evaluation results of the baseline approach and the submitted systems (denoted as *E_CRNN* in the table) on the *dev-set-test* split of the development set.

metrics of ER_{20° and F_{20°. For sound event localization evaluation, the proposed systems are superior to the baseline with a lower localization error score of the LE_{CD} metric. While the *E_CRNN* model is a little inferior to the baseline model on the evaluation metric of localization recall score of LR_{CD}. Moreover, we can see that the proposed method is extremely effective in reducing the localization errors, which decreases the localization errors from 29.3°to 18.68°on FOA, and from 32.2°to 17.98°on MIC, respectively. We think the main reason for this improvement, could be owing to the adoption of the VQT audio feature, and the MLP-Mixer module inserted in the baseline CRNN model to capture the interchannel audio patterns. On the whole, the proposed system achieved a higher improvement on the MIC audio representation, compared to the baseline approach, than that on FOA audio representation.

4. CONCLUSION

In this technical report, we present an MLP-Mixer enhanced convolutional recurrent neural network (CRNN) for the sound event localization and detection challenge in DCASE 2022, with the combination of the basic audio features and the VQT feature as inputs. The experimental results on the Sony-TAu Realistic Spatial Soundscapes 2022 (STARS22) benchmark dataset showed our system outperformed the baseline method, especially on the localization error evaluation metric.

5. REFERENCES

[1] 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, 2018.

[2] Y. Cao, Q. Kong, T. Iqbal, F. An, W. Wang, and M. D. Plumbley, "Polyphonic sound event detection and localization using a two-stage strategy," *arXiv preprint arXiv:1905.00268*, 2019.

[3] D. Salvati, C. Drioli, and G. L. Foresti, "Two-microphone end-to-end speaker joint identification and localization via convolutional neural networks," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–6.

[4] C. Evers and P. A. Naylor, "Acoustic slam," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1484–1498, 2018.

[5] C. Chen, U. Jain, C. Schissler, S. V. A. Gari, Z. Al-Halah, V. K. Ithapu, P. Robinson, and K. Grauman, "Soundspaces: Audio-visual navigation in 3d environments," in *European Conference on Computer Vision*. Springer, 2020, pp. 17–36.

[6] S. Lathuilière, B. Massé, P. Mesejo, and R. Horaud, "Neural network based reinforcement learning for audio–visual gaze control in human–robot interaction," *Pattern Recognition Letters*, vol. 118, pp. 61–71, 2019.

[7] 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*, vol. 29, pp. 684–698, 2020.

[8] A. Politis, K. Shimada, P. Sudarsanam, S. Adavanne, D. Krause, Y. Koyama, N. Takahashi, S. Takahashi, Y. Mitsufuji, and T. Virtanen, "Starss22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," *arXiv preprint arXiv:2206.01948*, 2022.

[9] A. Politis, S. Adavanne, D. Krause, A. Deleforge, P. Srivastava, and T. Virtanen, "A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection," *arXiv preprint arXiv:2106.06999*, 2021.

[10] T. N. T. Nguyen, K. Watcharasupat, N. K. Nguyen, D. L. Jones, and W. S. Gan, "Dcase 2021 task 3: Spectrotemporally-aligned features for polyphonic sound event localization and detection," *arXiv preprint arXiv:2106.15190*, 2021.

[11] T. N. T. Nguyen, D. L. Jones, K. N. Watcharasupat, H. Phan, and W.-S. Gan, "Salsa-lite: A fast and effective feature for polyphonic sound event localization and detection with microphone arrays," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 716–720.

[12] S. Park, Y. Jeong, and T. Lee, "Many-to-many audio spectrogram tansformer: Transformer for sound event localization and detection." in *DCASE*, 2021, pp. 105–109.

[13] K. Shimada, Y. Koyama, N. Takahashi, S. Takahashi, and Y. Mitsufuji, "Accdoa: Activity-coupled cartesian direction of arrival representation for sound event localization and detection," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 915–919.

[14] K. Xu, D. Feng, H. Mi, B. Zhu, D. Wang, L. Zhang, H. Cai, and S. Liu, "Mixup-based acoustic scene classification using multi-channel convolutional neural network," in *Pacific Rim conference on multimedia*. Springer, 2018, pp. 14–23.

[15] Q. Wang, J. Du, H.-X. Wu, J. Pan, F. Ma, and C.-H. Lee, "A four-stage data augmentation approach to resnet-conformer based acoustic modeling for sound event localization and detection," *arXiv preprint arXiv:2101.02919*, 2021.

[16] K. Shimada, Y. Koyama, S. Takahashi, N. Takahashi, E. Tsunoo, and Y. Mitsufuji, "Multi-accdoa: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 316–320.

[17] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, *et al.*, "Mlp-mixer: An all-mlp architecture for vision," *Advances in Neural Information Processing Systems*, vol. 34, pp. 24 261–24 272, 2021.