www.czasopisma.pan.pl PAN www.journals.pan.pl

jet INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2021, VOL. 67, NO. 4, PP. 735-741
Manuscript received May 13, 2021; revised November, 2021 DOI:10.24425/ijet.2021.139771

Detection and Localization of Audio Event for Home Surveillance Using CRNN

Suruthhi V S, V Smita, Rolant Gini J and K I Ramachandran

***Abstract*—Safety and security have been a prime priority in people's lives, and having a surveillance system at home keeps people and their property more secured. In this paper, an audio surveillance system has been proposed that does both the detection and localization of the audio or sound events. The combined task of detecting and localizing the audio events is known as Sound Event Localization and Detection (SELD). The SELD in this work is executed through Convolutional Recurrent Neural Network (CRNN) architecture. CRNN is a stacked layer of convolutional neural network (CNN), recurrent neural network (RNN) and fully connected neural network (FNN). The CRNN takes multichannel audio as input, extracts features and does the detection and localization of the input audio events in parallel. The SELD results obtained by CRNN with the gated recurrent unit (GRU) and with long short-term memory (LSTM) unit are compared and discussed in this paper. The SELD results of CRNN with LSTM unit gives 75% F1 score and 82.8% frame recall for one overlapping sound. Therefore, the proposed audio surveillance system that uses LSTM unit produces better detection and overall performance for one overlapping sound.**

***Keywords*—convolutional recurrent neural network (CRNN), gated recurrent unit (GRU), long short-term memory (LSTM), sound event localization and detection (SELD)**

I. Introduction

THE report by the United Nations Office on Drugs and Crime in 2017 indicates that the burglary rate over the years has increased across different countries [1]. Increase in crime rate leads to the requirement of home surveillance. Presently, two main home surveillance systems are in use. The first one is visual surveillance [2,3] that uses cameras, which cover the visible light range of the electromagnetic spectrum or other ranges like IR to observe the surroundings [4]. The second one, a growing research field that is redefining the surveillance system, is audio surveillance that uses microphones for observing the sound changes happening in the surrounding environment. When it comes to visual surveillance, the area uncovered by the camera and amount of memory required to store the information becomes an issue [3]. However, audio surveillance can overcome these limitations and can do surveillance under low computation power [5]. The audio input from the microphone has to be detected, classified and localized for audio surveillance. Detection and classification of audio events have been done as Sound Event Detection (SED) task, which is the prediction of temporal information; that is the onset and offset time of the audio event. Localization of audio events has been done as the estimation of direction of arrival (DOA), which is the prediction of Spherical or Cartesian coordinates of the audio event.

Various traditional neural network classifiers for SED task are implemented for acoustic audio surveillance [6, 7]. The supervised classification approaches like Support Vector Machine (SVM) [6], Gaussian Mixture Model (GMM) [7] and Hidden Markov Model (HMM) [7] and the unsupervised classification approaches like K-means clustering [6], GMM clustering [7] brings out the reverberation problem in traditional neural network classifiers. Persistence of sound than its actual duration caused by reverberation leads to unwanted overlapping of sounds. SED task leads to different classifier approaches as discussed in [8] evinces deep learning models like Deep Neural Network (DNN), Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) performs better than traditional classification methods. The recent developments [9, 10] in the SED task to detect multiple sound sources that are active at the same time can be an advantage for surveillance tasks.

Since the CNN [11] can learn the spatial information from the input features, and RNN [12] can learn the long temporal information, they both became a prominent approach for SED task. Recently, a combination of CNN and RNN, known as Convolutional Recurrent Neural Network (CRNN) has been proposed for the SED task [9, 10], which is found to overcome the reverberation problem. CRNN is a stacked layer of convolutional neural network (CNN), recurrent neural network (RNN) and fully connected neural network (FNN). The method proposed in [9] used a high-level feature known as Mel-frequency cepstral coefficients (MFCC) to learn different sound classes' common features. In contrast, the MSEDnet (multichannel SED network) and SEDnet (single channel SED network) method proposed in [10] used low-level (basic) features like log Mel-band energy, autocorrelation, generalized cross-correlation with phase transform (GCC-PHAT) to prove that the low-level features are capable of learning powerful representations. Despite the fact that the chosen features and CRNN architecture (given by [9] and [10]) contradict one another; both prove that stacking CNN and RNN produces better SED results compared to the individual implementation of CNN and RNN.

Suruthhi V S, V Smita and Rolant Gini J are with Department of Electronics and Communication Engineering, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, India. (e-mail: suruthhinkl@gmail.com, vsmita.98@gmail.com, j_rolantgini@cb.amrita.edu, rolantgini@gmail.com).

K I Ramachandran is with Centre for Computational Engineering & Networking (CEN), Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, India. (e-mail: ki_ram@cb.amrita.edu).

DOAs of the classified sounds are estimated using two main approaches; parametric based [13-15] approach and Deep Neural Network (DNN) based [16, 17] approach. The estimation of DOA using multiple signal classification (MUSIC) method [13], steered response power (SRP) method [14] and time difference of arrival (TDOA) method [15] are few of the most applied parametric based methods for localization of the audio events. The basic limitations of the parametric based method are their inability to associate the detected audio event with their respective DOAs for polyphonic sounds, their sensitivity towards low signal to noise (SNR) ratio and their reverberation conditions. These limitations of the parametric based methods are overcome by the DNN based methods [16, 17]; where DNN can learn the connections between the input features and their estimated DOAs. These DNN based approaches have proved that they perform equivalent to the parametric methods and also being robust to reverberant conditions [16,17]. Classification [16] and regression [17] are two supervised learning approaches in DNN for DOA estimation. The classification approach has limitations like the requirement of a larger dataset for training, estimation of only limited discrete angles, the unpredictability of unseen DOAs. Whereas the regression-based approach (DOAnet [17]), has proved to perform efficiently with comparatively smaller datasets, estimating the DOAs in continuous range and producing a seamless output for the unseen DOAs. DOAnet is the first method to estimate DOA for two sounds overlapping in the time window in the audio signal.

A surveillance system has to identify audio events (similar or different) while receiving from more than one direction. To accomplish this task, it is important to combine the detection and localization of the audio events; which can be achieved either by joint localization and detection [18] approach or by combining the results of detection and localization [19] approach. The latter approach has the difficulty of associating the detected sound events with their respective estimated DOAs. This data association problem is overcome by Sound Event Localisation and Detection network (SELDnet), a joint approach, which has been first developed and has been explained in [18]. Therefore, employing SELDnet for audio surveillance system is more reasonable for better processing. The neural network architecture employed in SELDnet is CRNN; which is a stacked layer of CNN, RNN and FNN. The SELDnet is able to overcome the reverberation problems faced in [6, 7]. To make the surveillance system generic towards different input array structures, phase and magnitude features are preferred over the other method specific features as used in the SELD approach [19]; since method-specific features are dependent on the nature of input array structure.

Based on the factors analysed in the above discussions, the SELDnet method is preferred for audio surveillance. The bidirectional RNN in SELDnet [18] uses a memory unit called gated recurrent unit (GRU) for learning sequential information from the input. The GRU update and reset gate to keep in the necessary information to learn and discard the unnecessary information. Another memory unit called long short-term memory (LSTM) available in bidirectional RNN of SELDnet does the same work but using three gates: input, output and forget gates. As LSTM has simpler and defined structure than GRU (GRU does the work done by three gates of LSTM with just two gates of complex structures), it learns the long sequential information better [10]. So, in this paper, detection and localization of audio events, which are the two main aspects of audio surveillance, is implemented for both SELDnet with GRU and SELDnet with LSTM unit to compare their SELD performance. The paper is organized as follows. In Section II, the SELDnet method employed is discussed. In section III, results and discussions are consolidated. In section IV, the conclusion of the proposed work is summarized.

II. METHODOLOGY

This section explains the SELDnet method shown by Fig. 1, which is used in the proposed work. SELDnet takes multichannel audio as input. From each input audio channel, phase and magnitude spectrograms are extracted, which are used as distinct features by neural network architecture (CRNN) for detection and localization. The CRNN predicts the active sound classes and their respective spatial locations as 3D cartesian coordinates in each input spectrogram frame, where the spectrogram frames are taken in sequence. The multiple sound classes that are predicted active by CRNN from the input spectrogram are classified to their respective sound classes using multi-label classification model. In parallel to the classification process, the multi-output regression model is used to obtain the 3D cartesian coordinates of predicted active sound classes. Thus, the SELDnet detects the active sound classes and estimates their direction of arrival in parallel. The SELDnet's feature extraction process and neural network architecture (CRNN) used in the proposed work is elucidated in the following discussion.

The REAL dataset [18], which uses real-life recordings of the urbansound8K to generate its audio records, are more pertinent for training home audio surveillance systems. The REAL dataset consists of separate one, two and three overlapping sounds (ov1, ov2 and ov3) datasets. Each dataset has 240 recordings for training and 60 recordings for testing. Thus, the proposed work has been tested and evaluated on the REAL dataset for its performance measures.

A. Feature extractor

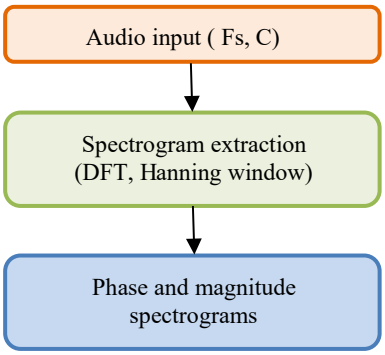

Fig.1. Feature extractor for home audio surveillance

The feature extractor as shown in Fig. 1, takes in the multichannel audio as input. The main objective of the feature extractor is to extract the magnitude and phase components of the audio signal for each audio channel. The input audio signal is sampled at a sampling rate, Fs which is 44.1 kHz. Then the

sampled audio signal is approximated to a standard length of L by truncating the signal if they exceed the length of L or padded with zeros if shorter than L. After standardizing the sampled audio signal's length, an N-point Discrete Fourier Transform (DFT) with an N-point Hanning window of N/2 hop length is applied to extract the spectrogram for each audio channel. Thus, the resulting spectrograms have a dimension of T x N/2 x C, where T is the frame sequence taking only positive frequency bins (N/2) excluding the zeroth bin, for all C audio channels. Obtained spectrograms are normalized and then have been used to extract each channel's respective magnitude and phase spectrograms. The magnitude and phase spectrogram extraction alters the dimension of the feature extraction output as T x N/2 x 2C. The extracted features are then taken as input by the CRNN to predict the active sounds and their DOAs.

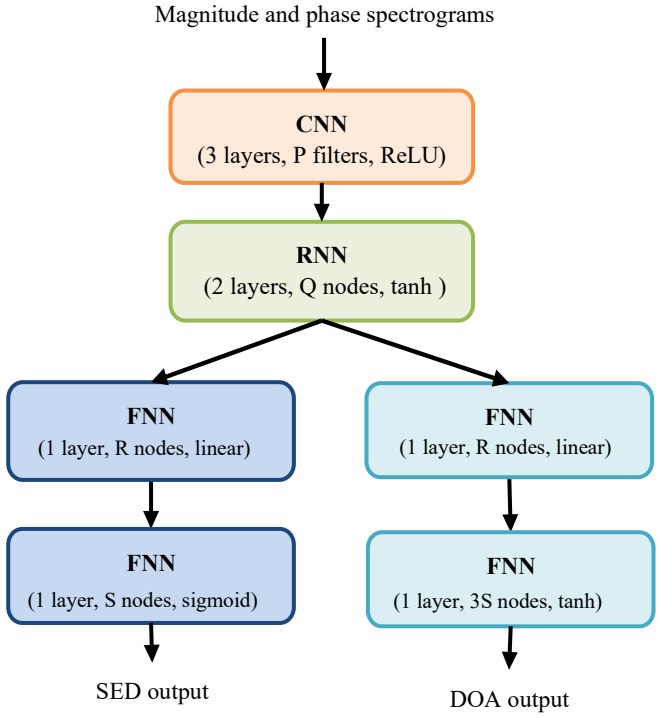

Fig. 2. SELD implementation for home audio surveillance using CRNN architecture

B. *CRNN architecture*

1) *Convolutional Neural Network (CNN)*

The obtained magnitude and phase spectrograms (T x N/2 x 2C) (which are feature extractor output) are divided into several segments where each segment have T number of frames. Each segment is taken as input by the three-layer 2D CNN which has been visualized in Fig. 2. The three layers of CNN learn the shift-invariant features from the magnitude and phase spectrogram. In each layer of CNN, P number of kernels (with a dimension of 3 x 3 x 2C) acts along the time-frequency-channel axis of the spectrogram to get the convolution output. As kernels are spanned across the channels of the spectrogram, CNN is able to learn relevant features that correlate the channels, which are important for the localization process. Whereas the features within each channel that are required for both detection and localization process are learned by spanning its kernel across the time and frequency axis of the spectrogram. The convolved output given by the kernels in each layer of CNN is normalized using batch normalization. The normalized output is optimized using Rectified linear unit (ReLU) activation as it eludes the vanishing gradients. The output dimension of the CNN in each layer is reduced to ease the computation in the forthcoming layers using max-pooling that acts along the frequency axis to maintain the frame sequence length (T) constant. The max-pooling size in each layer is determined accordingly to obtain a frequency bin dimension of two at the last layer of CNN. Thus, the final layer output of CNN contains all the learned shift-invariant features with a dimension of T x 2 x P.

2) *Recurrent Neural Network (RNN)*

The RNN layers have different types of inbuilt memory units. The first RNN layer takes in the previous CNN layer's output as its input, after reshaping its dimension to T x 2P as shown in Fig. 2. The temporal features from the input are learned using two layers of bidirectional RNN. The SELDnet performance is tested using two different memory units present in the RNN in obtaining the sound classes and DOA. The tested memory units of RNN are gated recurrent unit (GRU) and long short-term memory (LSTM) unit. The two layers of bidirectional RNN have Q number of considered memory units in each layer followed by tanh activation function for optimization of the output. Thus, the final output of the RNN contains the learned temporal information with a dimension of TxQ.

3) *Fully Connected Neural Network (FNN)*

The optimized output from the RNN is fed as input to two separate branches of FNN, each branch with two layers of FNN. One FNN branch is for SED output and the other for DOA estimation. The first layer in both the FNN branches has R number of nodes with linear activation, producing output with a dimension of T x R. The second layer in the SED branch has S number of nodes with sigmoid activation; each node's output represents one of the S sound classes of the input signal. The simultaneous activation of multiple classes is possible through sigmoid activation. The second layer in the DOA branch has 3S nodes with tanh activation, where the 3 in 3S nodes represents the x, y and z of the 3D cartesian coordinates of the respective S sound classes. The tanh activation used in DOA estimation optimizes the DOA output ranges to [-1,1] as in 3D cartesian coordinates. The final outputs of the two FNN branches are the estimated SED and DOA values for each sound classes.

The SED output for each sound classes will be in the continuous range of [0, 1] and DOA estimates in the continuous range of [-1, 1]. A threshold value of 0.5 is applied to the SED output to detect the active sound classes and its DOA. The detected active sound classes' respective DOA estimates are taken into consideration for evaluation. Thus the sound events are detected and localized for home audios. The evaluation metrics for SED and DOA are mentioned in the following section. The results obtained are evaluated and the performance has been discussed.

III. Results and Discussion

A. Evaluation metrics

The detection and localization results are evaluated using separate metrics. F1-score and error rate(ER) as proposed in [20] are used for SED evaluation. DOA error and frame recall (FR) as in [18] are used for DOA evaluation. Higher F1-score and lower ER indicates better performance of the detection task. DOA error comprises of DOA-got (doa_gt) and DOA-predicted (doa_pred). The doa_pred is calculated for sound classes that are estimated active in the SELDnet whereas doa_gt is calculated for sound classes that are mentioned active in the dataset. Lower DOA error and higher FR indicate better performance of the localization task. Lower the doa_pred value and the closer it gets to the doa_gt value indicate the improvement in the performance of the detection task. To evaluate the overall performance of both detection and localization, SELD-score is used [18]. Lower the SELD score, better the performance of SELDnet.

B. Results and discussions of the proposed method on different parameters

The optimized neural network parameters specified in [18] are of 64 filters for each CNN layer, 128 GRU nodes for each RNN layer and 128 nodes for first FNN layer of both branches, and has no dropout layer. Using these optimized parameters, and to find the optimum window length for the proposed method, the neural network architecture with GRU memory unit is evaluated with two different windowing lengths: 512 and 1024 points for 10 and 20 iterations or epochs, on one overlapping sound (ov1). For the window length of 1024, a max-pooling size of (8, 8, 4) and for the window length of 512, a max-pooling size of (8, 8, 2) for respective CNN layers are used in the proposed method such that the final frequency bin length is two. The frame sequence length of 512 is used for both window lengths. Compared to ten iterations, twenty iterations improve both the SED and DOA evaluation metrics values, for both the window lengths as given by Fig. 3 and Fig. 4. The neural network architecture with 512-point window length outmatches the SELD performance of the 1024-point window length, which has been shown in Fig.3. Therefore, for further evaluations, 512-point window length and 20 iterations are used in the proposed method.

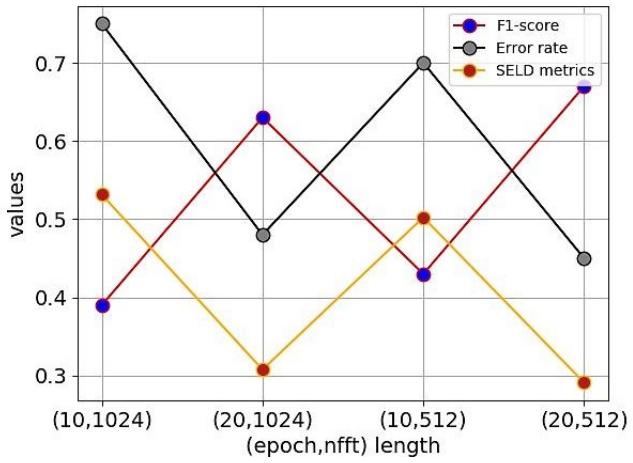

Fig. 3. SED and SELD evaluation metrics for different epochs and nfft points of GRU of the proposed method

With the finalized parameters of SELDnet, the proposed SELDnet with LSTM is evaluated for different dropout rates: 0.2, 0.3, 0.5 and 0.7, for one overlapping sound (ov1) to test the performance. LSTM with a dropout rate of 0.3 produces higher F1-score, lower ER and lower SELD-score whereas a dropout rate of 0.2 produces lower doa_pred and doa_gt error as seen in Fig. 5 and Fig. 6. Further experimenting with 0.0 dropout rate, the SED and DOA performance of SELDnet improved as shown in Fig. 5 and Fig. 6. Therefore, LSTM with no dropout rate is preferred for further evaluation of ov2 and ov3 datasets in this proposed method.

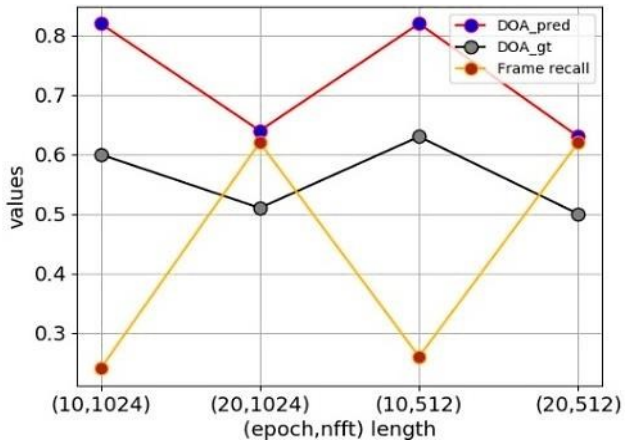

Fig. 4. DOA evaluation metrics for different epochs and nfft points of GRU of the proposed method

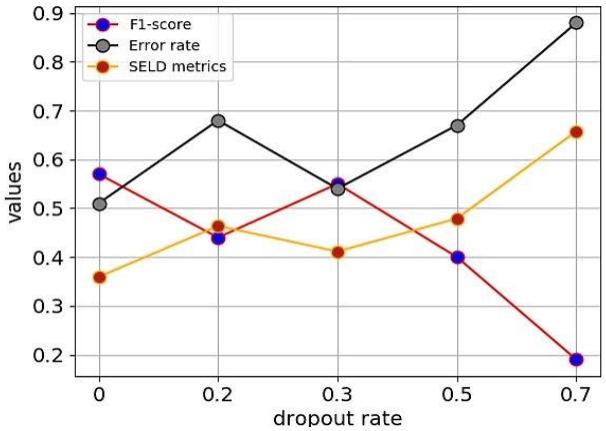

Fig.5. SED and SELD evaluation metrics for different dropout rates of LSTM of the proposed method

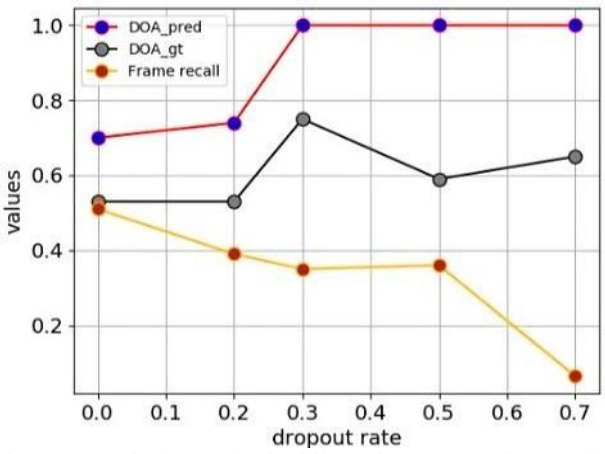

Fig.6. DOA evaluation metrics for different dropout rates of LSTM of the proposed method

C. *Performance comparison of GRU and LSTM for 1, 2 and 3 overlapping sounds*

The SED and DOA performance results obtained for both GRU and LSTM of the proposed method for three different overlapping sounds (ov1, ov2 and ov3) are tabulated in Table I. Using GRU, the increase in ER and DOA error along with the reduction in F1 score and ER with the increase in the number of overlapping sound indicates the increase in complexity to detect and localize the sounds. The difference between doa_gt and doa_pred increases for an increase in the number of overlapping sounds, which further indicates the reduction in SED performance. For LSTM, the performance of ov1 is better than ov2 and ov3.

Between ov2 and ov3, SED performance of ov3 is better than ov2 for LSTM. The SED and DOA performance of GRU and LSTM for ov3 is more comparable than ov1 and ov2, which can be concluded from Table I. Further, the LSTM is able to perform without much degradation in overall SELD performance for the increase in overlapping sounds as compared to GRU, whose performance reduces constantly with the increase in overlapping sounds as seen in Fig. 7.

Table I
SED and DOA metrics for ov1, ov2 and ov3 of the REAL dataset using GRU and LSTM units of the proposed method for 20 epochs

METRICS		GRU			LSTM		
		ov1	ov2	ov3	ov1	ov2	ov3
SED metrics:	F1	0.67	0.56	0.51	0.57	0.46	0.51
	ER	0.45	0.57	0.61	0.51	0.67	0.62
DOA metrics:	doa_pred	0.63	0.79	0.93	0.7	0.85	0.94
	doa_gt	0.5	0.57	0.65	0.53	0.59	0.69
	FR	0.62	0.20	0.05	0.51	0.1	0.01

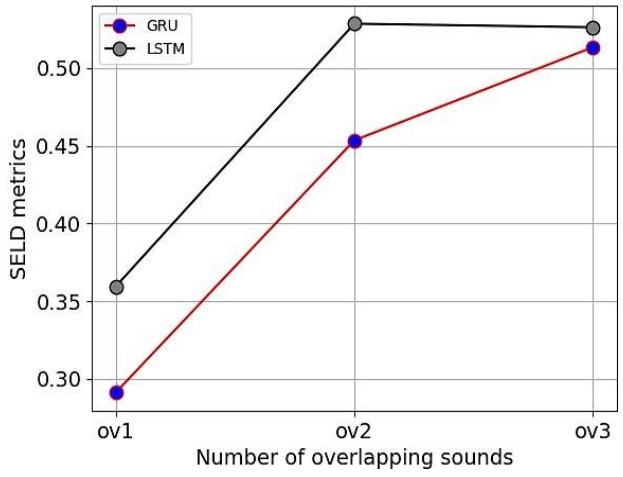

Fig.7. SELD score with respect to GRU and LSTM for ov1, ov2 and ov3

The confusion matrix for the predicted sound classes using GRU and LSTM is given in Fig. 8 and Fig. 9. It has been observed that true positive values are reducing with an increase in overlapping sounds. Comparing Fig. 8 and Fig. 9 of GRU and LSTM for ov1, it has been observed that LSTM doesn't misclassify one overlapping sound as two as GRU does. Thus indicates that LSTM learns sequences better than GRU. The FR and DOA error for both GRU and LSTM still need to be enhanced. So to compare the SED and DOA performance of GRU and LSTM, both are evaluated on ov1 dataset for 1000 epochs by the proposed method.

r\p	0	1	2
0	0.99	0.00	0.00
1	0.24	0.75	0.01

(a)

r\p	0	1	2
0	0.99	0.00	0.00
1	0.27	0.68	0.04
2	0.17	0.68	0.14

(b)

r\p	0	1	2	3
0	0.99	0.00	0.00	0.00
1	0.41	0.46	0.16	0.00
2	0.26	0.44	0.26	0.03
3	0.17	0.45	0.34	0.04

(c)

Fig.8. Normalized confusion matrix of SELDnet of the proposed method with GRU of the REAL dataset for 20 epochs: (a) ov1; (b) ov2; (c) ov3. The 'p' represents the predicted and the 'r' represents the reference number of sounds in the figure.

r\p	0	1
0	0.99	0.00
1	0.37	0.63

(a)

r\p	0	1	2
0	0.99	0.00	0.00
1	0.37	0.60	0.03
2	0.29	0.65	0.05

(b)

r\p	0	1	2	3
0	0.99	0.00	0.00	0.00
1	0.34	0.54	0.11	0.00
2	0.27	0.63	0.09	0.00
3	0.16	0.64	0.09	0.05

(c)

Fig.9. Normalized confusion matrix of SELDnet of the proposed method with LSTM of REAL dataset for 20 epochs: (a) ov1; (b) ov2; (c) ov3

D. *Performance comparison of SELDnet (LSTM) with other published methods*

Evaluating the SELDnet with LSTM unit on ov1 of the REAL dataset for 1000 epochs has given better SED and DOA results compared to the 20 epochs which can be observed from Table I and Table II. The obtained results of SELDnet with LSTM are compared with other SED and DOA baseline methods wherever possible and is as listed in Table II. The results of the proposed methods are highlighted in the table. The high F1 score and low ER of SELDnet with LSTM by the proposed method prove that the LSTM does the detection of REAL dataset better than MSEDnet [10], SEDnet and SELDnet (GRU) [18]. The FR of SELDnet with LSTM by the proposed method is very high compared to the SELDnet with

GRU [1] and DOAnet [17]. The 95% true positive rate of SELDnet with LSTM as seen in Fig. 10 supports the high FR value.

Table II
SED and DOA metrics for the REAL dataset for the ov1 of the proposed method (for 1000 epochs) compared with the published results

	Method	Metrics values	
SED metrics:		**F1**	**ER**
	SELDnet (LSTM)	**0.75**	**0.34**
	SELDnet (GRU) **[18]**	0.603	0.40
	SEDnet **[10]**	0.646	0.38
	MSEDnet **[10]**	0.662	0.35
DOA metrics:		**doa_pred**	**FR**
	SELDnet (LSTM)	**0.40**	**0.828**
	SELDnet (GRU) **[18]**	0.266	0.649
	DOAnet **[17]**	0.063	0.465
SELD metrics:		**SELD score**	
	SELDnet (LSTM)	**0.191**	
	SELDnet (GRU) **[18]**	0.287	

r\p	0	1
0	0.99	0.00
1	0.04	0.95

Fig.10. Normalized confusion matrix of SELDnet for ov1 for 1000 epochs of the proposed method

The DOAnet of the proposed method has the lowest DOA error compared to other methods. The SELDnet with LSTM gives the best SELD performance with low SELD score compared to SELDnet with GRU [18]. Even though SELDnet with LSTM gives very high FR value, the DOA error value is high compared to the DOA error value of DOAnet [17]. Thus considering both FR and DOA error values for the best DOA method, it is concluded that DOAnet gives better DOA performance. Therefore, the proposed SELDnet with LSTM gives the best SED and SELD performance with highest F1 score and lowest SELD score and ER than the other published methods.

IV. Conclusion

In this paper, a CRNN architecture known as SELDnet is used for detection and localization of audio events for home audio surveillance. The SELDnet detects and localizes the audio events in parallel. The detection is carried out as a multi-label classification approach and localization as a multi-output regression approach. SELDnet performance by CRNN with two different memory units: GRU and LSTM are compared and discussed in this paper. From the results, it is concluded that the SELDnet with LSTM outperforms the SELDnet with GRU and other published methods in SED and overall SELD task. Therefore, the high performance given by LSTM on REAL dataset indicates that LSTM is able to learn the real-life sounds better than other methods, which is essential for surveillance systems.

The SELDnet with LSTM has high FR value but their DOA error value has to be reduced to improve the DOA estimation. Therefore, the future work is focused on improving the DOA performance of the SELDnet with LSTM and also to evaluate their performance on more than one overlapping sounds for 1000 epochs.

References

[1] UNODC: United Nations Office on Drugs and Crimes, "Burglary | Statistics and data," 2017. https://dataunodc.un.org/crime/burglary.

[2] K. Lashmi and A. S. Pillai, "Ambient Intelligence and IoT Based Decision Support System for Intruder Detection," *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, Coimbatore, India, 2019, pp. 1-4. https://doi.org/10.1109/ICECCT.2019.8869327

[3] Dr. P. Prakash, R. Suresh and P.N. Kumar Dhinesh, "Smart City Video Surveillance using Fog Computing," in *International Journal of Enterprise Network Management*, vol. 10, no. 3/4, pp.389 – 399, 2019. https://doi.org/10.1504/IJENM.2019.103165

[4] Caught on camera, "Different Types of CCTV-CCTV Camera Types and Uses," 2020. [Online]. Available: https://www.caughtoncamera.net/news/ different-types-of-cctv/ .

[5] S. Ntalampiras, "Audio Surveillance," 2012. [pdf]. Available: https://www.itpress.com/Secure/elibrary/papers/9781845645625/9781845645625012FU1.pdf

[6] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio and M. Vento, "Audio Surveillance of Roads: A System for Detecting Anomalous Sounds," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 1, pp. 279-288, Jan. 2016. https://doi.org/10.1109/TITS.2015.2470216

[7] S. Ntalampiras, I. Potamitis and N. Fakotakis, "Probabilistic Novelty Detection for Acoustic Surveillance Under Real-World Conditions," in *IEEE Transactions on Multimedia*, vol. 13, no. 4, pp. 713-719, Aug. 2011. https://doi.org/10.1109/TMM.2011.2122247

[8] A. Mesaros et al., "Detection and Classification of Acoustic Scenes and Events: Outcome of the DCASE 2016 Challenge," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379-393, Feb. 2018. https://doi.org/10.1109/TASLP.2017.2778423

[9] E. Çakır, G. Parascandolo, T. Heittola, H. Huttunen and T. Virtanen, "Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1291-1303, June 2017. https://doi.org/10.1109/T ASLP.2017.2690575

[10] S. Adavanne, P. Pertilä and T. Virtanen, "Sound event detection using spatial features and convolutional recurrent neural network," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 2017, pp. 771-775. https://doi.org/10.1109/ICA SSP.2017.7952260

[11] P. Zinemanas, P. Cancela and M. Rocamora, "End-to-end Convolutional Neural Networks for Sound Event Detection in Urban Environments," *2019 24th Conference of Open Innovations Association (FRUCT)*, Moscow, Russia, 2019, pp. 533-539. https://doi.org/10.23919/FRUCT.2019.8711906

[12] G. Parascandolo, H. Huttunen and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real-life recordings," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, 2016, pp. 6440-6444. https://doi.org/10.11 09/ICASSP.2016.7472917

[13] L. Birnie, T. D. Abhayapala, H. Chen and P. N. Samarasinghe, "Sound Source Localization in a Reverberant Room Using Harmonic Based Music," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, 2019, pp. 651-655. https://doi.org/10.1109/ICASSP.2019.8683098

[14] L. O. Nunes et al., "A Steered-Response Power Algorithm Employing Hierarchical Search for Acoustic Source Localization Using Microphone Arrays," in *IEEE Transactions on Signal Processing*, vol. 62, no. 19, pp. 5171-5183, Oct.1, 2014. https://doi.org/10.1109/TSP.2014.2336636

[15] M. W. Hansen, J. R. Jensen and M. G. Christensen, "Pitch and TDOA-based localization of acoustic sources with distributed arrays," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, QLD, 2015, pp. 2664-2668. https://doi.org/10.1109/ICASSP.2015.7178454

[16] J. Pak and J. W. Shin, "Sound Localization Based on Phase Difference Enhancement Using Deep Neural Networks," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1335-1345, Aug. 2019. https://doi.org/10.1109/TASLP.2019.2919378

[17] S. Adavanne, A. Politis and T. Virtanen, "Direction of Arrival Estimation for Multiple Sound Sources Using Convolutional Recurrent Neural Network," *2018 26th European Signal Processing Conference (EUSIPCO)*, Rome, 2018, pp. 1462-1466, https://doi.org/10.23919/EUSIP CO.2018.8553182

[18] S. Adavanne, A. Politis, J. Nikunen and T. Virtanen, "Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Networks," in *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34-48, March 2019. https://doi.org/10.110 9/JSTSP.2018.2885636

[19] T. Butko, F. G. Pla, C. Segura, C. Nadeu and J. Hernando, "Two-source acoustic event detection and localization: Online implementation in a Smart-room," *2011 19th European Signal Processing Conference*, Barcelona, 2011, pp.1317-1321.

[20] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, pp. 162–178, 2016. https://doi.org/10.3390/app6060162