International Journal of Innovative Computing, Information and Control
Volume **20**, Number **3**, June **2024**

ICIC International ©2024 ISSN 1349-4198
pp. **709–723**

SEPARATION SOUND EVENT LOCALIZATION AND DETECTION USING NEURAL NETWORK AND TIME FREQUENCY MASKING

Ranny[1,4,*], Dessi Puji Lestari[2] and Tati Latifah Erawati Rajab Mengko[3]

[1]Doctoral Program in School of Electrical Engineering and Informatics
[2]School of Electrical Engineering and Informatics
[3]Department of Biomedical Engineering, School of Electrical Engineering and Informatics
Bandung Institute of Technology
Jl. Ganesa No. 10, Lb. Siliwangi, Kecamatan Coblong, Kota Bandung, Jawa Barat 40132, Indonesia
{ dessipuji; tati }@stei.itb.ac.id

[4]Computer Science Department
School of Computer Science
Bina Nusantara University
Jl. Kebon Jeruk Raya No. 27, Kebon Jeruk, West Jakarta, Jakarta 11530, Indonesia
[*]Corresponding author: ranny@binus.ac.id

Received August 2023; revised November 2023

Abstract. *Sound processing frameworks have many kinds of approaches. These approaches are distinguished based on the data and purpose of the sound processing research area. One of the research purposes is the localization and detection of sound events. The research also uses non-overlapping data and overlapping data to test the sound processing algorithms. Previous research has shown good accuracy with non-overlapping sound, but the accuracy decreased when overlapping data sound was used, which is more representative of real-life conditions. Therefore, this research is developing a framework that can handle both non-overlapping and overlapping data. The focus of this research is on overlapping sound processing. The use of overlapping sound in sound recognition systems decreases recognition accuracy, and this research aims to improve the sound recognition system without eliminating the use of overlapping sound. This research uses a different approach compared with any other common research. It uses non-overlapping data and separated sound data for sound event localization and detection (SELD). The separation of overlapping sound generates the separated sound, which is used for detecting the overlapping data sound. This new framework is called Separation SELDnet. The Separation SELDnet framework achieves 81% localization accuracy and 70% detection accuracy, while the SELDnet alone provides lower accuracy for overlapping sound. The Separation SELDnet offers a great opportunity to increase sound recognition using both non-overlapping and overlapping sound. This new method can be more adaptive to real-life conditions.*

Keywords: Overlapping sound, Sound event localization and detection, Sound separation

1. **Introduction.** Environment sound processing is used in many tasks, such as in elder care monitoring systems to detect distress activities [1-3]. It is also utilized to help visualize surroundings through sound in hearing aid systems [4,5]. In the field of robotics, environment sound processing is developed for artificial human hearing processes [6,7]. Most of the task problems consist of localization and detection. Sound event localization and detection (SELD) combine two areas of research in sound processing. The research

DOI: 10.24507/ijicic.20.03.709

purpose of sound event localization is to process spatial sound data information, generating sound source directional processing from this research area. Additionally, the purpose of event sound detection is to label sound source data, such as human sound, animal sound, or environmental sound [8-11]. The results of this research are affected not only by the approach and technique used to generate the learning model but also by the real data from a real environment.

There are many challenges when implementing a system in a real situation, the unexpected background noise, overlapping sound, and low-quality challenges during the implementation process. This research addresses the issue of overlapping sounds. The deep neural network (DNN), one of the most common techniques, is believed to be used for handling the limitation of the data variant [12-15]. Some research focuses on developing DNN architectures to increase accuracy [12], but the accuracy of the system decreases when using overlapping data [16]. DNN for SELD with overlapping data yields various results depending on the relation between the architecture and the data [13,17]. Besides, the model learning technique like DNN is not the only approach to solve the overlapping problem.

The combination of data training between non-overlapping and overlapping data is used to increase accuracy, but it has limitations when using a different dataset [18]. The overlapping data condition is one of the common problems in the implementation step in a real environment. Some research gives better results on non-overlapping data, but it decreases when using overlapping data [18,19]. This research develops a method based on the separation of overlapping data on SELDnet. The separation method occurs before the model learning step. The result of the separation is used for training data on SELDnet. The separation data used as training data is a contribution of this research. To be clearer, the difference between the common technique and the purpose technique can be seen in Figure 1.

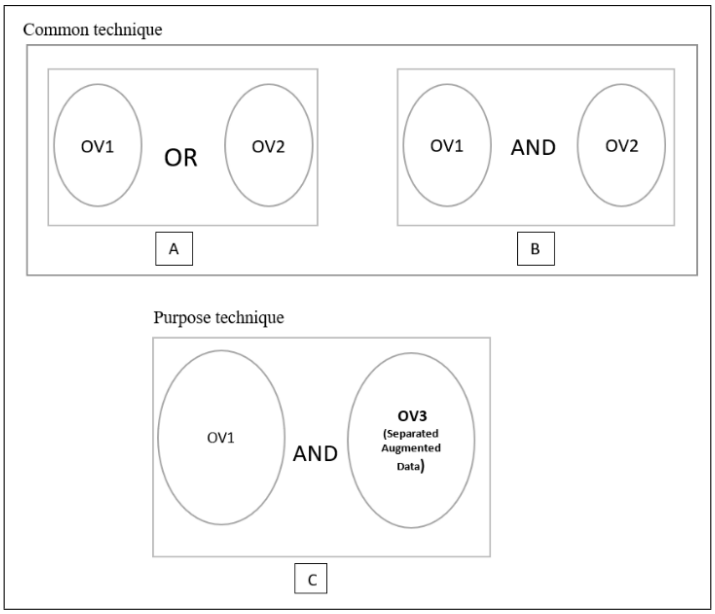

FIGURE 1. Comparison technique

The common technique has two approaches. The first uses non-overlapping data for the recognition process or for separating the overlapping data, as shown in Figure 1 part A. This approach provides separated sound characteristics between the non-overlapping (OV1) and overlapping data (OV2). The non-overlapping data is used when the purpose of

the system is to recognize or classify the non-overlapping data, and the same applies to the overlapping data. Many research studies use this first approach in their sound processing systems [20-22]. The second approach uses both non-overlapping data (OV1) and the overlapping data (OV2) as the training set data, resulting in both sound characteristics being mixed in the training set data, as shown in Figure 1 part B. The second approach is believed to improve the recognition result on overlapping data, which is a common challenge in most of the similar research topics [23-25]. The third part, C, shows a different method using non-overlapping data (OV1) and separated overlapping data (OV3) as the dataset for training and testing. This third approach is established in this research as one of the research novelties. The method is developed based on this approach and is named Separation SELDnet. Basically, the Separation SELDnet method is developed based on the previous method SELDnet. The previous SELDnet framework uses CNN, RNN, and FNN as the methods for sound event and localization processes [19]. The SELDnet training process uses both single and overlapping sound, making it similar to the common technique B in Figure 1. The single SELDnet has low accuracy when tested using overlapping sound, reaching below 60% accuracy, indicating a chance for improvement. One of the goals of this research is to increase the accuracy of the overlapping sound condition. We believe that adding a separate process to the SELDnet could improve accuracy.

Many kinds of separation methods are used to separate overlapping sound. NUSSL is one of the research projects that focus on music separation problems. It develops various separation methods and generates a basic sound processing that can be used in other sound research areas [26]. Time frequency masking (T-F masking) is one of the separation methods in the NUSSL project and is used in this research. T-F masking uses masking data to separate overlapping data. The accuracy of the separation is better than the non-masking method, which is why T-F masking is used in this research [8,27-29]. The separation result of overlapping data sound is used in the sound event localization and detection system (SELDnet). The SELDnet framework has two purposes: detection and localization of sound events. It uses neural network techniques as its approach. The method is divided into two main purposes: sound event detection (SED) and sound event localization. SED is a common purpose in sound processing, where it processes sound to recognize the sound label during one-time sequence. Sound localization is used to get the tracking position of the source sound, using Azimuth and Elevation degree as the data. Both identification and localization of sound events are tested using their augmented dataset based on the NIGENS dataset, the augmented dataset called TAU-NIGENS Spatial Sound Events 2020 dataset. The results of the experiment are measured using a matrix score based on F-score and error rate, and SELDnet score. Details about SELDnet and NUSSL separation method are discussed in Section 2. Section 3 discusses the augmented dataset. The results and analysis are discussed in Section 4. Finally, Section 5 concludes the paper.

2. **Methods.** This research uses sound event detection and localization based on neural network technique called SELDnet, which was developed by a community that focuses on the sound event detection and localization problem known as the DCASE Community. The community regularly publishes challenges, and one of these challenges is used in this research to develop the proposed method. The new method, Separation SELDnet, aims to improve the previous methods by utilizing a separation method called NUSSL (pronounced 'nuzzle') [26]. NUSSL is used to separate overlapping sounds and has shown significant accuracy in the separation process. Combining the separation process with SELDnet can increase the accuracy of SELDnet without requiring NUSSL.

2.1. **SELDnet.** Sound event detection and localization are the two main purposes of the experiment. The active sound estimates the spatial location of the sound source, a process known as sound event detection and localization. These two main goals are used in this research, which focuses on sound detection and recognition based on the activity called sound event detection (SED). Most SED techniques use supervised classification to recognize each frame with overlapping sound conditions. Some research also utilizes different data formats based on various recording tools to improve accuracy. The commonly used data format includes single microphone, binaural, and first-order Ambisonics (FOA) [30]. In this research, FOA is used for the experimental step. After the sound event detection, the next process is sound localization development within the system.

Previous research focused on sound localization based on the simultaneous arrival of sound sources, known as direction-of-arrival (DOA), which is referenced in this research [31-34]. Generally, the DOA technique is divided into two approaches: parametric and deep neural network (DNN) [27,35-37]. The parametric approach is limited in handling data with noise in the form of reverberation and low signal-to-noise (SNR) values, while DNN performs better in handling reverberation noise. Additionally, DOA research deals with overlapping sound data and estimates the number of active sound sources using regression or classification techniques and DNN [10,11]. Apart from developing techniques and recording tools, the use of geographic elements on recording tools, such as full azimuth and linear arrays, is also explored to enhance the performance of detection and localization. The full azimuth angle value covers 180 degrees. There is also a combination of full azimuth and linear arrays using FOA or Ambisonics signals [38].

The framework of SELDnet begins with feature extraction, similar to other common techniques. This process accepts the audio input and extracts the spectrogram for each C channel from multichannel audio using discrete Fourier transform (DFT). It also utilizes Hamming window with the M window size and fifty percent overlapping. The output provides spectrogram and magnitude values as the feature extraction. Only the positive values are used for the next step in two-dimensional convolutional neural network (2D CNN). The CNN layer consists of three layers as a process after the feature extraction [19,30,39,40]. Each CNN layer has P filters with dimensions of $3 \times 3 \times 2C$ and uses the ReLU (rectified linear unit) activation function. After that, the activation output is normalized using batch normalization, and dimension reduction is performed using max-pooling (M, Pi) along the frequency axis. The output dimension $T \times 2 \times P$ is the dimension of the last CNN layer with P filters. The activation output from the CNN is then resized into a sequence of T-sized frames with a feature length of 2P to be input to the RNN layer.

The RNN uses the gated recurrent unit (GRU) layer with tanh activation and is bidirectional. RNN is a robust neural network for sequence data or time series data, making it suitable for use in SELDnet. GRU is a type of RNN that carries out the training process faster than other types, such as long short-term memory (LSTM) [41]. After obtaining the RNN output, it is divided into two parts: SED and DOA, both using fully connected (FC) layers with a value of nodes R. In SED, the sigmoid activation function is used with N nodes, while DOA has three nodes (3N) representing the sound direction data (x, y, z). The SED output will be a classification result with multi-labels in the form of continuous values $[0, 1]$, while the DOA output will be a regression estimate with multi-outputs in the form of continuous values $[-1, 1]$.

The training process determines the initial target of DOA and SED. In DOA, the initial targets are x, y, and z, and are used when the sound event is active, while all initial target values are set to zero when the event is not active. The SED target value is set to one if the event is active and zero when the event is not active. The SED output provides

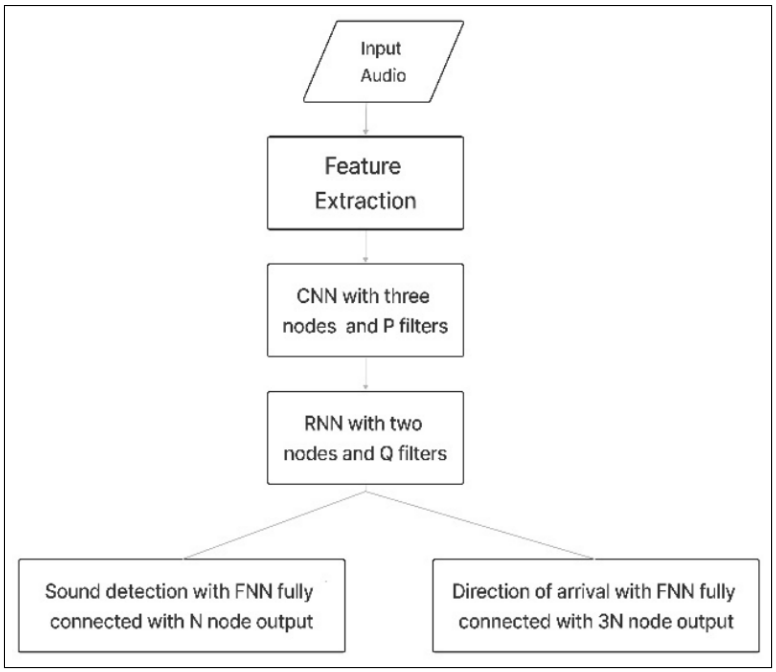

FIGURE 2. SELDnet framework

an event sound class label that appears on the timeline, indicating when the event sound is active. The DOA output provides Cartesian values x, y, z with values in the range $[-1, 1]$, representing the positions of active sound sources at each time frame. Figure 2 shows the framework of SELDnet.

2.2. **Separation method.** One of the main purposes of this research is to increase the accuracy of overlapping sound recognition by using the separation of overlapping sound. The separation technique is combined with the SELDnet method for sound recognition. There are several methods of sound separation, which can be categorized into two approaches: with masking data and without masking data (blind separation). An example of blind separation is non-negative matrix factorization (NMF) [42]. Blind separation can separate overlapping sound without training the data, but its accuracy is lower compared to the masking data approach, especially when dealing with abstract sounds like environmental sound [8]. The masking separation method uses Time-Frequency (T-F) masking, where short time Fourier transform (STFT) is used to generate magnitude spectrogram data. The data masking is calculated based on the spectrogram data and is used for the separation process. The result of separation is obtained from the STFT value, and the inverse of STFT converts it back into audio data format [8,28].

In previous research, sound separation problems have focused on separating musical instruments and separated background music from vocal sound. T-F masking separation methods have been employed to separate overlapping environmental sounds in various conditions. Northwestern University Source Separation Library (NUSSL) is one of the libraries that develops separation techniques using Python as the programming language. The separation process starts with basic audio signal processing, including reading audio data, padding, adding, and cutting audio data, audio transformations, and writing audio. These initial processes are bundled into an AudioSignal object. The input and output audio data are represented using a two-dimensional array of pulse-code modulated (PCM) values. The dimensions represent the time and channels. In this research, there are four channels, referring to the first-order Ambisonics format used in the TAU NIGENS 2020 dataset.

2.3. **Design framework of Separation SELDnet.** This research aims to develop a sound recognition system to improve the accuracy of overlapping sound conditions. Based on previous research, there is a significant difference in accuracy between non-overlapping and overlapping sound. The accuracy of overlapping sound is lower than that of non-overlapping sound. However, in real-world conditions, overlapping sound frequently occurs and becomes a constraint on the system. Therefore, this research focuses on solving this problem. The recognition system is based on the SELDnet research. The results of the SELDnet method indicate that the use of overlapping sound datasets results in lower accuracy compared to non-overlapping sound datasets. To improve the accuracy of overlapping sound conditions, the separation method is implemented in the SELDnet method, resulting in the Separation SELDnet method. The separation technique is based on time-frequency masking. The short time Fourier transform (STFT) values are converted into magnitude spectrogram values. The magnitude spectrogram is processed to obtain masking values, which are then used in the separation step. The process of separation with time-frequency masking requires sample data from the non-overlapping sound to generate the overlapping sound. The non-overlapping sound is used as masking data to separate the overlapping sound. The first step of the separation process involves masking data processing, resulting in an output called masked_stft. Masked_stft is obtained by combining masked_abs and phase values. Masked_abs is generated by multiplying the magnitude and masked_data. The absolute value of STFT dataMix becomes the magnitude value. The phases value comes from the angle of STFT's mix data sound. The mask_data value is obtained by dividing the single data sound divided with the maximum number between STFT's single sound and STFT's mix data sound. Masked_stft, as the output of masking, is used in the separation process and then the result of separation is inverted using STFT inverse and converted back into audio data format. Figure 3 shows the steps of Time-Frequency masking.

The following steps outline the separation process using Time-Frequency masking with STFT and STFT inverse:
1. Make data masking:
a. mask_data = dataSingle/(max(dataMix,dataSingle)) + nussl.constants.EPSILON
b. masked_stft = masked_abs * np.exp(1j * phase)
where:
masked_abs = magnitude * mask_data
maginitude = np.abs(dataMix.stft_data)
phase = np.angle(dataMix.stft_data)
2. Separation step:
a. estSeparation = dataMix.make_copy_with_stft_data (masked_stft)
b. inverse STFT = estSeparation.istft()
3. Write audio:
estSeparation.write_audio_to_file (folderOutput + "HasilSeparated.wav", sample_rate = 44100)

FIGURE 3. Time-Frequency masking process

The result of the separation process is used as input data for the sound event detection and localization SELDnet. As discussed in Section 2, the detection and localization of sound events start with feature extraction, following the method of the original SELDnet. The new Separation SELDnet is developed to increase the accuracy of overlapping sound detection and localization that generates from the separation process. Therefore, the SELDnet has already been tested and provides good accuracy on non-overlapping sound, and as a result, the new Separation SELDnet also achieves better accuracy on

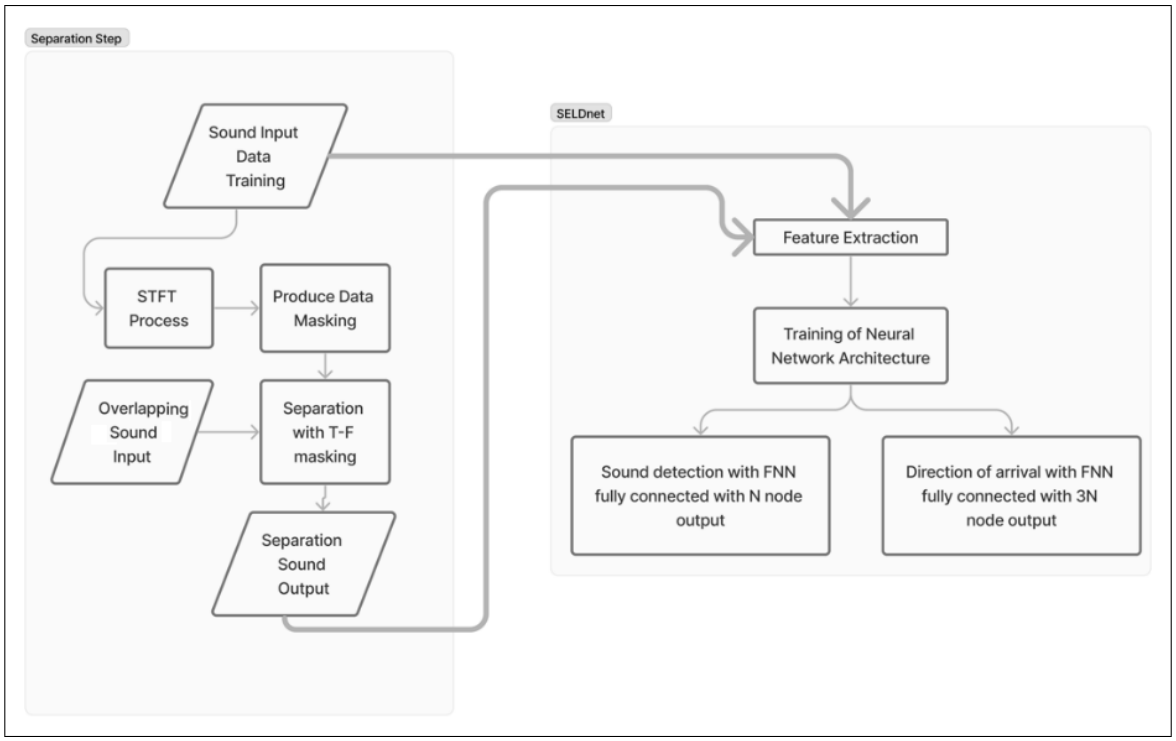

FIGURE 4. Framework separation sound event detection and localization

non-overlapping sound that generates from overlapping sound. Figure 4 shows the framework of Separation SELDnet.

3. **Data Augmentation and Experiment.** The Separation SELDnet technique is tested using multiple datasets, specifically the TAU-NIGENS Spatial Sound Events 2020 dataset. This dataset was published and used in the 2020 DCASE Challenge for the experiments on the SELDnet-only technique. The experiment results in the challenge were measured by error rate and F-score to evaluate the accuracy and performance of the Separation SELDnet technique. This section discusses the data and experiments conducted in this research.

3.1. **TAU-NIGENS Spatial Sound Events 2020.** The TAU-NIGENS Spatial Sound Events 2020 dataset was used to measure the performance of SELDnet in DCASE 2020. The data is generated from NIGENS, containing sound sources in WAV format with various types of sound. The NIGENS dataset is played in different types of rooms, at different times and directions, with overlapping conditions. The recorded sound is in first-order Ambisonics format, resulting in four-channel sound. The dataset includes both static and moving sound sources, resulting in the creation of the TAU-NIGENS Spatial Sound Events 2020 dataset, which provides direction information and overlapping conditions. This dataset is suitable for localization and identification experiments. More details about this dataset can be found in [43]. The dataset consists of various recording conditions, including non-overlapping and overlapping data. Each WAV file has a duration of one minute and contains more than two types of sound events. In the case of overlapping data, certain parts of the WAV file contain two overlapping sound events, while non-overlapping data have no overlapping sounds. This grouping of data is used in this research. The original dataset is used to find suitable parameters for the experiments.

3.2. **Data augmentation.** The TAU-NIGENS Spatial Sound Events 2020 dataset has two groups of datasets based on overlapping conditions: non-overlapping (OV1) and overlapping (OV2) sounds. All the datasets within both groups consist of 14 classes, each with a duration of one minute. These classes are similar to the NIGENS dataset classes but without a general label class. The specific label classes include alarm, baby, crash, dog, engine, female scream, female speech, fire, footsteps, knock, male scream, phone piano. The existing overlapping dataset has a limited number of variants, so this research produced an augmented dataset to increase the number of variants for overlapping datasets. The data augmentation process uses the non-overlapping dataset as the material for augmentation. The augmentation process involves framing the sound based on the class label, then combining two different classes with the overlapping condition. Subsequently, this single overlapping sound is combined again with other class sounds but without overlapping conditions, resulting in a final sound file that has a similar overlapping condition compared to the original overlapping dataset. A total of 1200 WAV files were generated as the augmented dataset, including both overlapping and non-overlapping conditions. This augmented dataset is used to test the Separation SELDnet method in the experiments. Figure 5 shows the process of augmenting the data. Labels A and B represent the non-overlapping data used as input for the augmentation process, and label C represents the result of the augmentation process, providing the overlapping data sound used in the experiments for both Separation SELDnet and SELDnet-only methods.

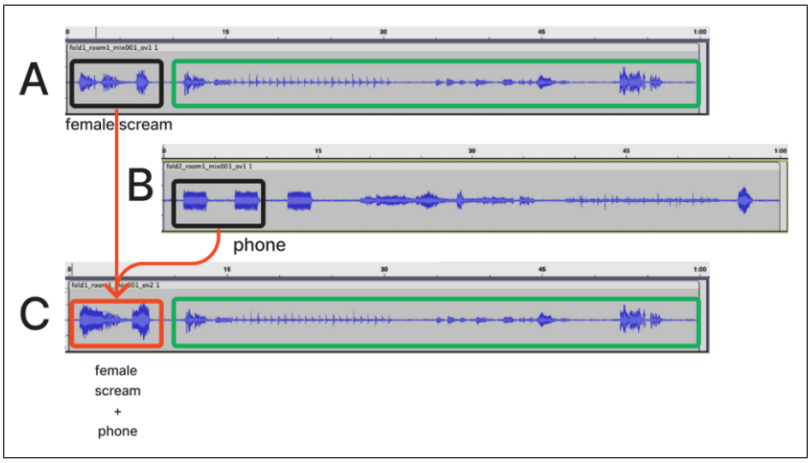

FIGURE 5. Simulation of augmentation data process

3.3. **Experimental setup.** The experiments are divided into several steps, each having different purposes. The first experiment aims to find the best parameters for SELDnet using the original dataset. Additionally, this experiment aims to determine which parameters can be handled by the machine. There are various parameter combinations to be adjusted in this experiment, as shown in Table 1. The index in front of the number represents the combination number, while the "er" preceding the index indicates that the experiment resulted in an error, and the machine could not handle the parameter, leading to an "out of memory" status. The "Length of label sequence" column contains values of 64, 128, 256 and 512 variants, where 256 is the maximum value for the machine. The next parameter to be adjusted is the "Number of CNN2D filter", which controls the number of CNN nodes and remains constant for each layer. The "RNN size" and "FNN size" parameters are both in list data structures containing the number of nodes for RNN and FNN, and the length of the list corresponds to the number of layers. The number of RNN and FNN nodes does not significantly impact the machine's performance. The number of

Table 1. Combination of parameters

Index	Length of label sequence	Batch size	Number of CNN2D filter	RNN size	Number of layers	FNN size	Number of epochs	Status
er0	64	256	64	128	2	128	2	Out of memory
1	64	**128**	64	128	2	128	2	
2	64	128	64	**64**	2	**64**	2	
er1	64	128	**128**	64	2	64	2	Out of memory
3	64	128	**64**	**256**	2	**256**	2	
4	64	128	64	256	2	**128**	2	
5	64	128	64	**512**	2	**512**	2	
er2	**128**	128	64	**64**	2	**64**	2	Out of memory
6	128	**32**	64	64	2	64	2	
7	**256**	32	64	64	2	64	2	
8	256	**64**	64	64	2	64	2	
er3	**512**	64	64	64	2	64	2	Out of memory
9	**256**	64	64	**128**	2	**128**	2	
er4	256	**128**	64	128	2	128	2	Out of memory
er5	256	128	64	**64**	2	**64**	2	Out of memory
10	256	**64**	**32**	**128**	2	**128**	2	
11	256	64	32	**64**	**3**	**64**	2	
12	256	64	**64**	64	3	64	2	
13	256	64	64	64	3	64	2	

Table 2. The best result of combination parameter

Index	Length of label sequence	Batch size	Number of CNN2D filter	RNN size	Number of layers	FNN size	Number of epochs
A	64	128	64	256	2	256	50
B	128	64	64	128	3	128	50
C	64	64	128	128	2	128	50

RNN layers is indicated in the "Number of layers" column, while the FNN layers remain constant at one layer.

This experiment uses quick mode, which only trains the network for two epochs, resulting in accuracy results that do not reflect the true performance of the method. Instead, the aim is to determine which parameters the machine can handle. Based on the experiment, the selected parameter combinations are used in the full mode experiment, as shown in Table 2. In the full mode experiment, fifty epochs are used in the training model architecture. The Separation SELDnet method is tested using the C index parameter combination in the full mode experiment. The dataset component leads to the experiment being divided into two parts. The first part uses a combination of augmented data and original data from the TAU NIGENS 2020 dataset. The second part uses 100% augmented data. This division of the dataset helps test the distribution of data labels.

4. **Results and Analysis.** The full mode experiment results of the three parameter combinations are shown in Table 3. The best values are highlighted in the grey background field. The matrix score is divided into several parts. The first part is based on the

TABLE 3. The result of experiment with full mode

	Class-aware localization scores									
Score	DOA_error					Frame_recall				
Split mode / Index (best epoch)	Validation	Test	OV1	OV2	IR1	Validation	Test	OV1	OV2	IR1
A (36)	23	24.1	17.8	27.9	23.5	64.3	60.3	**70.5**	54.3	60.4
B (43)	22.1	**22.2**	**16.3**	25.3	**21.5**	63.5	**69.6**	68.6	54.4	59.6
C (43)	**21**	22.4	17.1	**25.1**	21.7	**65.5**	61.7	70.1	**56.9**	**61.7**
	Localization-aware detection scores									
Score	Error rate					F_score				
Split mode / Index (best epoch)	Validation	Test	OV1	OV2	IR1	Validation	Test	OV1	OV2	IR1
A (36)	0.69	0.72	0.61	0.78	0.72	39.7	36.7	49.8	29.5	37.1
B (43)	**0.51**	**0.71**	0.6	**0.75**	**0.7**	40.1	38.9	49.8	**33.3**	39.4
C (43)	0.66	**0.71**	**0.59**	0.76	**0.7**	**44.4**	**39.4**	**52.7**	32.6	**40**
Score	seld_score									
Split mode / Index (best epoch)	Validation		Test		OV1		OV2		IR1	
A (36)	0.44		0.47		0.38		0.52		0.47	
B (43)	0.44		0.46		0.38		**0.5**		0.46	
C (43)	**0.42**		**0.45**		**0.37**		**0.5**		**0.45**	

localization and detection focus. The class-aware localization score measures the localization process but with true label detection, and similarly, the localization-aware detection score shows the score of the detection process but with true localization only. The next part of the matrix score is the split mode, which has five fields: Validation, Test, OV1, OV2, and IR mode. The Validation mode uses the same class label data for both training and testing, while the Test mode uses different label data for training and testing. The OV1 field shows the performance using the non-overlapping data, and the OV2 field is for the overlapping dataset. The last field, IR, represents the accuracy for various types of reverberation datasets.

Finally, the accuracy of all systems is measured using SELD for each split mode. This SELD score is calculated based on the scores from the class-aware localization and localization-aware detection. The best accuracy has a value as close to zero as possible. The SELD score formula is shown in Formula (1).

$$SELD_{score} = (SED\ score + DOA\ score)/2 \tag{1}$$

where

$$SED\ score = (Error\ Rate + (1 - F_{score}))/2$$
$$DOA\ score = (DOA\_error/180 + (1 - Frame\_recall))/2$$

The SED score is used to calculate the performance of detection part of SELDnet and Separation SELDnet. The SED score consists of two elements, F-score and error rate. The F-score is generated by Formula (2) and error rate using Formula (3). The error rate is total value of *substitution* $S_{(k)}$, *deletion* $D_{(k)}$ and *insertion* $I_{(k)}$ divided by all class

number $N_{(k)}$.

$$F_{score} = \frac{2 \cdot \sum_{k=1}^{K} TP_{(k)}}{2 \cdot \sum_{k=1}^{K} TP_{(k)} + \sum_{k=1}^{K} FP_{(k)} + \sum_{k=1}^{K} FN_{(k)}} \tag{2}$$

where $TP_{(k)}$ or *true positive*, the prediction and reference class give the same result. $FP_{(k)}$ or *false positive*, the prediction shows the class is active, but the reference is not active. $FN_{(k)}$ or *false negative*, the prediction shows the class is not active, but the reference is active.

$$Error\ Rate = \frac{\sum_{k=1}^{K} S_{(k)} + \sum_{k=1}^{K} D_{(k)} + \sum_{k=1}^{K} I_{(k)}}{\sum_{k=1}^{K} N_{(k)}} \tag{3}$$

where

$$S_{(k)} = \min\left(FN_{(k)}, FP_{(k)}\right)$$
$$D_{(k)} = \max\left(0, FN_{(k)} - FP_{(k)}\right)$$
$$I_{(k)} = \max\left(0, FP_{(k)} - FN_{(k)}\right)$$

The class-aware localization uses DOA error and Frame recall to measure accuracy. The DOA error value is generated by comparing the target class (x_G, y_G, z_G) with the prediction result of localization class (x_E, y_E, z_E). This DOA error value is calculated by Formulas (4) and (5). The Frame recall is calculated by Formula (6).

$$\sigma = 2 \cdot \arcsin\left(\frac{\sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2}}{2}\right) \cdot \frac{180}{\pi} \tag{4}$$

where $\Delta x = x_G - x_E$; $\Delta y = y_G - y_E$; $\Delta z = z_G - z_E$.

$$DOA_{error} = \frac{1}{D} \cdot \sum_{d=1}^{D} \sigma\left(\left(x_G^d, y_G^d, z_G^d\right), \left(x_E^d, y_E^d, z_E^d\right)\right) \tag{5}$$

where D is total number of all dataset localization class; $\left(\left(x_G^d, y_G^d, z_G^d\right), \left(x_E^d, y_E^d, z_E^d\right)\right)$ is estimated degree and reference DOA degree on the dth data.

$$Frame\_recall = TP/(TP + FN) \tag{6}$$

In the class-aware localization part, the DOA error score indicates that the index B combination parameter has the lowest score, 16.3 on OV1, showing that the index B provides the minimum error on the non-overlapping condition. However, for the F-score, the index C parameter combination gives the best result. The C index parameter combination yields the best experimental result, and thus, these parameter combinations are used for the Separation SELDnet experiment, which is the main purpose of this research. The last experiment serves the main purpose of the research and uses the same matrix score as the first and second experiments. The performance of the Separation SELDnet method is then compared with the SELDnet only. The focus of the comparison result is on the OV2 split mode, as mentioned in the experimental setup. The Separation SELDnet experiment utilizes both the augmentation data and original data.

The component data of the first experiment includes augmentation data and original data, leading to a significant increase in accuracy on the OV2 split mode. The comparison results of the experiment are presented in Figure 6. Figure 6 shows that the performance of Separation SELDnet results in a lower direction of arrival error on class-aware localization than the SELDnet only method. For frame recall, the Separation SELDnet with augmentation dataset achieves the best frame recall score among the experimental setups. Furthermore, in terms of localization-aware detection scores, the F-score, error rate, and

FIGURE 6. The experiment result on comparison graphic

SELD score of Separation SELDnet with augmentation dataset demonstrate the best performance and the lowest score. These results indicate that Separation SELDnet yields the best overall results in the experiments.

5. **Conclusions.** Based on the experiment, the Separation SELDnet method demonstrates an increase in accuracy for sound event localization and detection in overlapping sound problems. Moreover, it does not affect the performance of non-overlapping sound processing, unlike other common methods. The performance of the Separation SELDnet method is adaptable for both non-overlapping and overlapping sound conditions. The results of this research can serve as a fundamental basis for addressing other problems using this novel approach to sound processing. Additionally, it can be applied in scenarios where adaptability to real conditions with varying sound data is essential. There are still opportunities to further develop the Separation SELDnet with different separation methods to enhance the accuracy of localization and detection. Various datasets can also be employed to test the method's performance, and a comparison with different localization and detection methods can be conducted against SELDnet. Therefore, numerous opportunities arise from this research, presenting potential advancements and applications in the field of sound processing.

REFERENCES

[1] H. M. Do, K. C. Welch and W. Sheng, SoHAM: A sound-based human activity monitoring framework for home service robots, *IEEE Trans. Autom. Sci. Eng.*, pp.1-15, DOI: 10.1109/TASE.2021.3081406, 2021.

[2] C. N. Doukas and I. Maglogiannis, Emergency fall incidents detection in assisted living environments utilizing motion, sound, and visual perceptual components, *IEEE Trans. Inf. Technol. Biomed.*, vol.15, no.2, pp.277-289, DOI: 10.1109/TITB.2010.2091140, 2011.

[3] D. Istrate, M. Vacher and J. F. Serignat, Generic implementation of a distress sound extraction system for elder care, *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, pp.3309-3312, DOI: 10.1109/IEMBS.2006.259469, 2006.

[4] C. Y. Chen, P. Y. Kuo, Y. H. Chiang, J. Y. Liang, K. W. Liang and P. C. Chang, Audio-based early warning system of sound events on the road for improving the safety of hearing-impaired people, *2019 IEEE 8th Glob. Conf. Consum. Electron. (GCCE 2019)*, pp.933-936, DOI: 10.1109/GCCE 46687.2019.9015516, 2019.

[5] M. Mielke and R. Brueck, Design and evaluation of a smartphone application for non-speech sound awareness for people with hearing loss, *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp.5008-5011, DOI: 10.1109/EMBC.2015.7319516, 2015.

[6] B. Kwon, Y. Park and Y. S. Park, Sound source localization for robot auditory system using the summed GCC method, *2008 Int. Conf. Control. Autom. Syst. (ICCAS 2008)*, vol.1, no.1, pp.241-245, DOI: 10.1109/ICCAS.2008.4694557, 2008.

[7] Y. Geng and J. Jung, Sound-source localization system for robotics and industrial automatic control systems based on neural network, *International Conference on Smart Manufacturing Application*, pp.311-315, 2008.

[8] Y. Luo and N. Mesgarani, Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol.27, no.8, pp.1256-1266, DOI: 10.1109/TASLP.2019.2915167, 2019.

[9] Y. Li and Z. Wu, Animal sound recognition based on double feature of spectrogram in real environment, *2015 Int. Conf. Wirel. Commun. Signal Process. (WCSP 2015)*, DOI: 10.1109/WCSP.2015.7341003, 2015.

[10] E. L. Ferguson, S. B. Williams and C. T. Jin, Sound source localization in a multipath environment using convolutional neural networks, *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.2386-2390, DOI: 10.1109/ICASSP.2018.8462024, 2018.

[11] F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini and F. Piazza, A neural network based algorithm for speaker localization in a multi-room environment, *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp.1-6, DOI: 10.1109/MLSP.2016.7738817, 2016.

[12] Y. Zhang, W. Wang and H. Zhang, Neural cryptography based on quaternion-valued neural network, *International Journal of Innovative Computing, Information and Control*, vol.18, no.6, pp.1871-1883, DOI: 10.24507/ijicic.18.06.1871, 2022.

[13] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S. Y. Chang and T. Sainath, Deep learning for audio signal processing, *IEEE J. Sel. Top. Signal Process.*, vol.13, no.2, DOI: 10.1109/JSTSP.2019.2908700, 2019.

[14] T.-E. Chen et al., S1 and S2 heart sound recognition using deep neural networks, *IEEE Trans. Biomed. Eng.*, vol.64, no.2, pp.372-380, DOI: 10.1109/TBME.2016.2559800, 2017.

[15] S. Uhlich et al., Improving music source separation based on deep neural networks through data augmentation and network blending, *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.261-265, DOI: 10.1109/ICASSP.2017.7952158, 2017.

[16] D. Krause, A. Politis and K. Kowalczyk, Comparison of convolution types in CNN-based feature extraction for sound source localization, *Eur. Signal Process. Conf.*, vol.2021-Janua, pp.820-824, DOI: 10.23919/Eusipco47968.2020.9287344, 2021.

[17] S. Sigtia, A. M. Stark, S. Krstulović and M. D. Plumbley, Automatic environmental sound recognition: Performance versus computational cost, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol.24, no.11, pp.2096-2107, DOI: 10.1109/TASLP.2016.2592698, 2016.

[18] I. Trowitzsch, *Robust Sound Event Detection in Binaural Computational Auditory Scene Analysis*, https://depositonce.tu-berlin.de/handle/11303/10967, 2020.

[19] 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, DOI: 10.1109/JSTSP.2018.2885636, 2019.

[20] T. J. Han, K. J. Kim and H. Park, Location estimation of predominant sound source with embedded source separation in amplitude-panned stereo signal, *IEEE Signal Process. Lett.*, vol.22, no.10, pp.1685-1688, DOI: 10.1109/LSP.2015.2424991, 2015.

[21] B. Gao, W. L. Woo and S. S. Dlay, Adaptive sparsity non-negative matrix factorization for single-channel source separation, *IEEE J. Sel. Top. Signal Process.*, vol.5, no.5, pp.989-1001, DOI: 10.1109/JSTSP.2011.2160840, 2011.

[22] C. C. Took, S. Sanei, S. Rickard, J. Chambers and S. Dunne, Fractional delay estimation for blind source separation and localization of temporomandibular joint sounds, *IEEE Trans. Biomed. Eng.*, vol.55, no.3, pp.949-956, DOI: 10.1109/TBME.2007.909534, 2008.

[23] D. Tian, X. Xu, Y. Tao and X. Wang, An improved activity recognition method based on smart watch data, *2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, pp.756-759, DOI: 10.1109/CSE-EUC.2017.148, 2017.

[24] Y. Li, J. Woodruff and D. Wang, Monaural musical sound separation based on pitch and common amplitude modulation, *IEEE Transactions on Audio, Speech, and Language Processing*, vol.17, no.7, pp.1361-1371, DOI: 10.1109/TASL.2009.2020886, 2009.

[25] R. Mogi and H. Kasai, Noise-robust environmental sound classification method based on combination of ICA and MP features, *Artif. Intell. Res.*, vol.2, no.1, DOI: 10.5430/air.v2n1p107, 2012.

[26] E. Manilow, P. Seetharaman and B. Pardo, The Northwestern University Source Separation Library, *Proc. of the 19th International Society of Music Information Retrieval Conference*, 2018.
[27] R. Roy and T. Kailath, ESPRIT-estimation of signal parameters via rotational invariance techniques, *IEEE Trans. Acoust.*, vol.37, no.7, pp.984-995, DOI: 10.1109/29.32276, 1989.
[28] L.-C. Yang and A. Lerch, Remixing music with visual conditioning, *2020 IEEE International Symposium on Multimedia (ISM)*, pp.181-188, DOI: 10.1109/ISM.2020.00039, 2020.
[29] Ö. Yilmaz and S. Rickard, Blind separation of speech mixtures via time-frequency masking, *IEEE Trans. Signal Process.*, vol.52, no.7, pp.1830-1846, DOI: 10.1109/TSP.2004.828896, 2004.
[30] S. Adavanne, A. Politis and T. Virtanen, Multichannel sound event detection using 3D convolutional neural networks for learning inter-channel features, *2018 International Joint Conference on Neural Networks (IJCNN)*, pp.1-7, DOI: 10.1109/IJCNN.2018.8489542, 2018.
[31] 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)*, pp.1462-1466, 2018.
[32] S. Chakrabarty and E. A. P. Habets, Multi-speaker localization using convolutional neural network trained with noise, *Proc. of Neural Inf. Process. Syst.*, http://arxiv.org/abs/1712.04276, 2017.
[33] W. He, P. Motlicek and J.-M. Odobez, Deep neural networks for multiple speaker detection and localization, *Proc. of Int. Conf. Robot. Autom.*, pp.74-79, http://arxiv.org/abs/1711.11565, 2017.
[34] T. Hirvonen, Classification of spatial audio location and content using convolutional neural networks, *Proc. of Audio Eng. Soc.*, no.5, 2015.
[35] Y. Huang, J. Benesty, G. W. Elko and R. M. Mersereati, Real-time passive source localization: A practical linear-correction least-squares approach, *IEEE Trans. Speech Audio Process.*, vol.9, pp.943-956, 2001.
[36] J. H. Dibiase, *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays*, Ph.D. Thesis, Brown University, 2000.
[37] R. Schmidt, Multiple emitter location and signal parameter estimation, *IEEE Trans. Antennas Propag.*, vol.34, no.3, pp.276-280, DOI: 10.1109/TAP.1986.1143830, 1986.
[38] H. Teutsch, *Modal Array Signal Processing: Principles and Applications of Acoustic Wavefield Decomposition*, 1st Edition, Springer, Berlin, Heidelberg, DOI: 10.1007/978-3-540-40896-3, 2007.
[39] H. Lim, J. Park, K. Lee and Y. Han, Rare sound event detection using 1D convolutional recurrent neural networks, *Workshop on Detection and Classification of Acoustic Scenes and Events*, Music and Audio Research Group, Seoul National University, Seoul, Korea, 2017.
[40] A. Scenes, *A Report on Sound Event Detection with Different Binaural Features Sharath Adavanne*, Ph.D. Thesis, Tuomas Virtanen Department of Signal Processing, Tampere University of Technology, 2017.
[41] J. Chung, C. Gulçehre, K. Cho and Y. Bengio, Gated recurrent neural networks on sequence modeling, *arXiv Preprint*, arXiv: 1412.3555, 2014.
[42] R. Ranny, D. P. Lestari, T. L. E. Rajab and I. S. Suwardi, Separation of overlapping sound using nonnegative matrix factorization, *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, pp.424-429, DOI: 10.1109/ISRITI48646.2019.9034580, 2019.
[43] A. Politis, S. Adavanne and T. Virtanen, A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection, *arXiv Preprint*, arXiv: 2006.01919, 2020.

Author Biography

Ranny received bachelor's degree in Computer Science from Tarumanagara University, Jakarta, Indonesia and the master's degree from Universitas Indonesia in 2013. Now, she joins as Faculty Member in Bina Nusantara University also a candidate doctoral in Information Technology, Bandung Institute of Technology. Her research lines are sound processing, sound pattern recognition, classification of fruit tapping sound, baby crying sound classification and other non-speech sound processing.

Dessi Puji Lestari received the B.E. degree in Informatics Engineering from the Bandung Institute of Technology, Bandung, Indonesia, in 2002, and the M.Eng. and Ph.D. degrees in Computer Science from the Tokyo Institute of Technology (Titech) Tokyo, Japan, in 2007 and 2011, respectively, where she joined Furui Speech and Language Processing Laboratory from 2004 until 2011. Her latest position is Senior Lecturer at School of Electrical Engineering and Informatics, ITB. Her research interests include speech signal processing, speech recognition, speech synthesizer, speaker recognition, emotional recognition, and broad domain of human computer interaction and machine learning. She has been actively engaged as a country representative of the Asian Spoken Language Research and Evaluation (CASLRE) and the Oriental Chapter of the International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques (O-Cocosda). In industry, she develops and assists industries in developing speech-based applications and is currently co-founding Prosa Solusi Cerdas where she leads the speech products research and development.

Tati Latifah Erawati Rajab Mengko received the bachelor's degree in Electrical Engineering from the Bandung Institute of Technology, Bandung, Indonesia, in 1977, and the Ph.D. degree from the École Nationale Supérieure d'Électronique et de Radioélectricité de Grenoble (ENSERG), Institut National Polytechnique de Grenoble, France, in 1985, where she studied texture-based image processing. Since 2005, she has been a Professor with the School of Electrical Engineering and Informatics, ITB, where she is currently the Head of the Biomedical Engineering Research Group. Her research interests include biomedical signal and image processing.