Development of scream detection system with large-scale scream dataset

YoungJun Kim
Information & Media Research Center
Korea Electronics Technology Institute
Seoul, Korea
youngjunss@keti.re.kr

Dalwon Jang
Information & Media Research Center
Korea Electronics Technology Institute
Seoul, Korea
dalwon@keti.re.kr

JongSeol Lee
Information & Media Research Center
Korea Electronics Technology Institute
Seoul, Korea
leejs@keti.re.kr

***Abstract*— In this paper, the process of developing scream detection system is written. This research provides a basic technology Scream detection that can detect dangerous situations in public and private spaces can be helpful in public and personal safety, but it has not been studied with large scream data. The previous researches are based on small datasets in which there are less than 1,000 scream samples. To develop a scream detection system, we collected 11,921 scream samples. By utilizing the dataset, we evaluated and compared five different detection models that can accurately detect and classify screams. In our experiments, the model that combines CNN and Transformer outperformed in scream classification accuracy and F1 Score, demonstrating its effectiveness in processing and classifying speech data in real-world situations. Future research directions include improving the performance of scream detection in various environments, building additional datasets and optimizing the model, and developing the ability to recognize various kinds of dangerous situations.**

Keywords— Scream, CNN, Transformer, Scream Detection, CCTV, black spot

I. Introduction

With the recent technological advancement, detection and recognition have been used in various fields such as smart cities, smart homes, public safety monitoring, and emergency response systems. Particularly, the use of deep learning leads to improvement of such research areas, thus it is possible to extract valuable information from acoustic data, which had not been easily used. Among acoustic data, detection of some sound signals indicating urgent situations like screams can be highly utilized [1,2]. These technologies are expected to greatly contribute to improving safety and welfare across society.

However, existing scream detection models suffer from degraded detection performance due to various background noise and environmental factors, and balancing accuracy with real-time processing power remains a technical challenge. To address these issues, the development of more advanced algorithm-based systems is needed, which starts with more precise analysis and understanding of the characteristics of speech data such as screams [3-5].

Furthermore, given the lack of recent research on scream detection and emergency response, and the fact that existing research is not based on extensive datasets. In this thesis, we present the performance of our scream detection system on a large dataset of screams. While existing studies are based on small datasets of less than 1,000, we build a dataset of about 11,000 screams in this thesis. We experimented with scream/noise classification and scream time detection on this dataset. The experimental results from Mel-spectrogram and five different deep learning-based models are presented. This will serve as a basis for further scream detection research.

This paper describes the dataset construction in Section 2, followed by the system and experiments in Section 3, where we describe the data preprocessing, model structure, and learning process. In Section 4, we present the evaluation results and performance metrics. Finally, Section 5 concludes the paper by discussing the results, implications, limitations, and future research directions.

II. Scream detection dataset

A. Collecting scream data

Previous scream detection researches had involved system design and development, experimentation, and validation based on very small data sets of a few hundred units. Table 1. shows the utilization of screaming data in previous studies. A large quantity of data is essential for deep learning-based model design and training. But, previous studies could not be performed with a large scale of data.

TABLE I. Prior research utilization of scream data

	Number of scream data utilizations
[2]	180
[7]	63
[6]	431
[8]	3,078 (based on 342 original recordings)

Our objective is to collect more than 10,000 scream data, and with the objective our dataset is made with three different datasets. We utilized data published on AIHub [9], the MIVIA dataset [10], and data recorded in an anechoic chamber. The data published on AIHub is called "AI_Hub_Emergency Speech and Sound" and consists of wav files and corresponding metadata. The metadata contains information about the file organization and information about the file content, and we collected the data by categorizing the data that is labeled as 'scream' in the category. MIVIA dataset is a dataset that collects audio of people or objects performing activities and actions, and it consists of metadata and wav, and we collected screams based on the value of CLASS_info in the metadata. Scream data in the above two datasets are thought to be recorded by professional actor or actress. We thought scream data of unprofessional people are necessary. Thus, the data is recorded in an anechoic room with 8 people whose age are between 20-37.

Finally, 11,921 scream data are collected, and these are used to develop scream detection system. The lengths of files are within 10s. Onset and offset of all data are annotated.

979-8-3503-6463-7/24/$31.00 ©2024 IEEE

B. *Noise addition*

It is assumed that scream detection system will be performed in noisy road condition, thus dataset of both scream with road noise and road noise are necessary. The assumption makes the detection system be helpful for detecting dangerous situations in real CCTV or outside. For this, road noise is also recorded and used in our dataset. The recorded road noise data is divided into two groups: one is used for positive scream data, and the other is used for negative noise data. From the first group of noise data, 10-second data were randomly selected and extracted, and duplicates were excluded by performing a duplicate check. They are added to scream data after adjusting signal magnitude with SNR of 0dB. From the second group of noise data, 10-second data were selected without duplication. There are 11,921 scream and 11,921 noise data in our dataset.

III. Scream Detection System

A. *Features*

Among the features of speech, the most representative features are MFCC, Mel-Spectrogram, and Chroma, but Mel-Spectrogram has been used the most in recent research, so Mel-Spectrogram is used as a feature of speech in this paper. Mel-Spectrogram is a speech feature that expresses speech as two-dimensional image information, and it contains frequency information, time information, and amplitude information. In addition, since the input data can be standardized, it contains enough information of speech and is easy to use as input data for training [11-12].

B. *Scream Detection models*

To build scream detection system, 5 detection models are tested. They have not been used for scream detection, but they were applied for detection of acoustic/voice information.

First, the Sound Event Detection (SED) model structure of Ebbers_UPB [13], which participated in the Sound Event Detection And Separation in Domestic Environments of DCASE 2023 [14], was taken and modified to detect screams. Ebbers_UPB's SED model is composed of a Forward Backward Convolution- Recurrent Neural Network (FBCRNN) and a Bi-direction Convolution- Recurrent Neural Network (bi-CRNN). The FBCRNN detects whether a sound occurs or not, and the Bi-CRNN detects the label and segment of the sound that occurs. The model in [13] was a multi-class classification model that classified 10 classes, but BCE Loss was used for binary classification to distinguish between scream and non-scream.

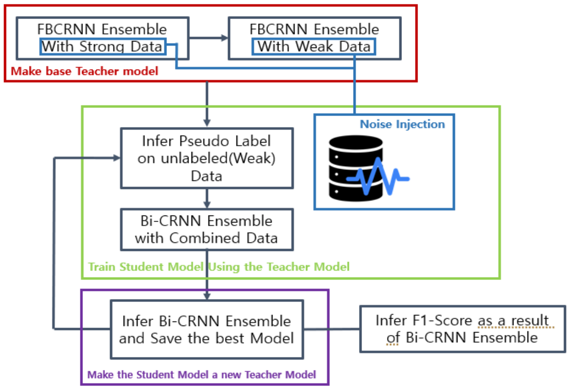
FBCRNN Ensemble With Strong Data
FBCRNN Ensemble With Weak Data
Make base Teacher model
Noise Injection
Infer Pseudo Label on unlabeled(Weak) Data
Bi-CRNN Ensemble with Combined Data
Train Student Model Using the Teacher Model
Infer Bi-CRNN Ensemble and Save the best Model
Infer F1-Score as a result of Bi-CRNN Ensemble
Make the Student Model a new Teacher Model

Fig. 1. Scream detection model based on a variant of Ebbers_UPB's SED model

Second, we conducted experiments on Convolution-Recurrent Neural Network (CRNN), which is a combination of CNN and RNN. Since we used Mel-Spectrogram as the input feature of scream data, we designed a Convolution Neural Network Layer (CNN) to capture feature-based patterns and a Recurrent Neural Network (RNN) series to consider the time information of time series data. There are two representative RNN models, Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), and we conducted experiments with each of them.

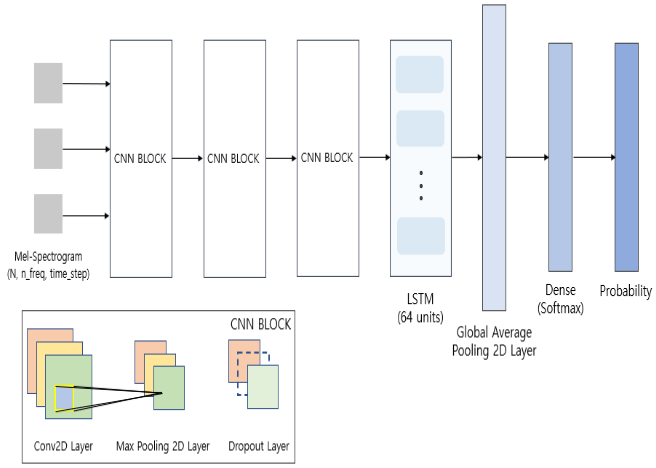

Fig. 2. CRNN model Structure (CNN + LSTM)

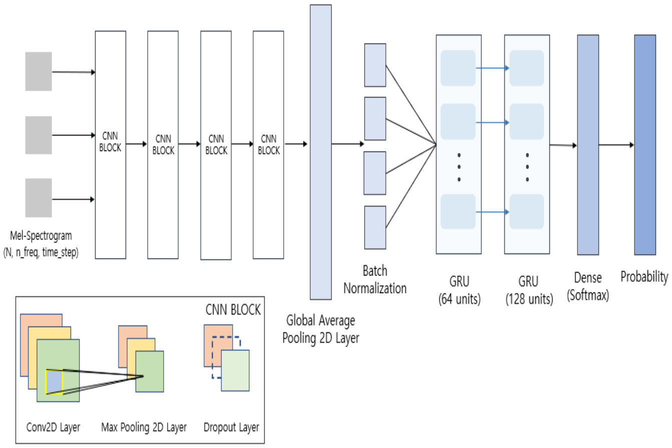

Fig. 3. CRNN model Structure (CNN + GRU)

The fourth model was a combination of CNN and Transformer models [15]. The Transformer model is also a model related to time series data, and it uses CNN to recognize local patterns and perform feature-based learning, while the Transformer model learns long-range dependencies between elements in the input sequence, which is advantageous for grasping complex relationships between sequences, and can process sequences at once, enabling faster learning and inference than conventional RNNs. For the above reasons, we conducted experiments by combining CNN and Transformer models to design a model.

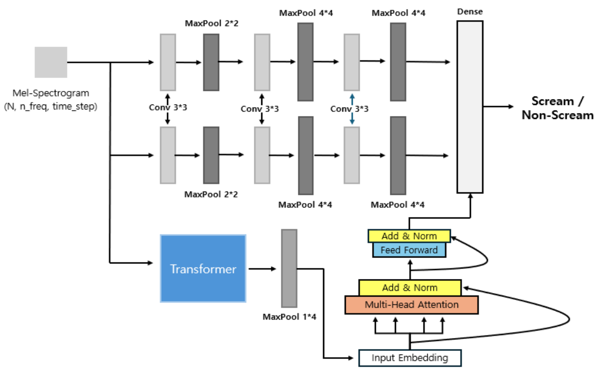

Fig. 4. CNN + Transformer model Structure

Finally, we designed the model using Residual Networks (ResNet) developed by Microsoft [16]. ResNet is a residual-based learning that combines the output of the previous layer with the input of the next layer by skipping several layers. It can solve the problem of gradient loss and explosion due to the effect of residual connection, and it is efficient in learning complex features because it can reuse features in various layers. In the case of screaming, it is important to capture the feature of screaming in a specific section, and we thought that the ResNet model is suitable for the above reasons, and it can strongly recognize and analyze visual patterns when the input data is a Mel-spectrogram, and it is efficient in recognizing visual changes in the Mel-spectrogram in the screaming section, so we designed and experimented with one of the scream detection models proposed in this paper.

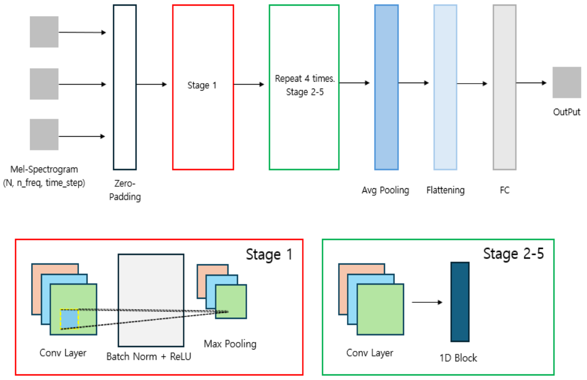

Fig. 5. ResNet model Structure

IV. Experimental Results

A. Experimental Setup

In this paper, various scream detection models were used in experiments to verify the accuracy of scream detection and scream interval detection. The number of screaming data and non-screaming data were matched to avoid data imbalance, and for accurate detection of scream interval, only the difference between the actual screaming interval and the predicted screaming interval was within 0.5 seconds among the correctly recognized screaming data.

In our experiments, the dataset is divided into training, validation, and test datasets, and the ratio was 7:1.5:1.5. During the partitioning process, we considered that the classes of the partitioned data were evenly divided, and the batch size was set to 16 in the mini-batch learning method, considering the problem of computation and computing resource usage in the learning process [17].

B. Results

In the experiments in this paper, we considered two metrics: scream classification accuracy and F1-Score, which is based on the accuracy of scream segments. Scream classification accuracy is measured by the agreement between ground truth and prediction. Since the purpose of scream detection is to detect screams quickly and accurately, we considered the prediction value of the validation data to be a correct prediction if the difference between the scream section of the ground truth and the scream section of the prediction was within 0.5 seconds among the data with correct accuracy.

In addition, since the model tested in this paper is a classification model, the objective performance of the model was evaluated by checking the classification model evaluation metrics.

TABLE 2. below shows the scream classification accuracy and F1 Score of the scream detection model.

TABLE II. Scream Classification Accuracy and F1 Score by Model

Scream Detection Models	*Accuracy*	*Mean F1 Score*
Based on Ebbers_UPB	79.35%	0.7993
CNN + LSTM	80.7%	0.8211
CNN + GRU	84.03%	0.8507
ResNet	96.89%	0.9654
CNN + Transformer	**97.64%**	**0.9822**

The accuracy in Table 2 is the class prediction accuracy that compares the predicted class with the ground truth of the test set and determines whether it matches or not. Therefore, after judging whether a person is screaming or not, we proceed to validate the screaming interval. The judgment of screaming was recognized as correct when the difference between the ground truth and the predicted screaming was within 0.5 seconds. Based on these criteria, the model with the highest accuracy is the combined CNN and transformer model with 97.64% accuracy.

Furthermore, the F1 score is calculated for each class as the average of the sum of the recall and precision scores, and in this paper, we averaged the F1 scores of the scream and non-scream classes. The process of measuring the F1 scores of the scream class and non-scream class was similar to that of the scream classification, with the F1 score being calculated when the predicted scream interval differs from the ground truth by less than 0.5 seconds. Since both screams and non-screams are important for scream detection, and if either of them has a low value, the average F1 score will also be low, averaging them will help to evenly evaluate the two classes.

The average F1 score was 0.9822, with the model combining CNN and transformer having the highest score.

V. Conclusion

Development of scream detection system is explained in this paper. To develop the system, scream data is first collected. Using two open datasets and recorded scream samples, we got 11,921 scream samples. After making negative samples (road noise) and annotating onset/offset, we studied baseline performances. Five detection models are tested with Mel-spectrogram feature. Among the models, the model combining CNN and Transformer showed the best performance. Therefore, future research should explore ways to further extend its applicability to a variety of environments beyond road noise. This can be done by building datasets based on data collected in more diverse situations and optimizing the model to broaden its applicability.

This paper develops an advanced AI-based scream detection model, which is an important step forward in ensuring public and personal safety. Through experiments using 11,921 scream data mixed with various environmental noises, we evaluated and compared various models that balance real-time processing and accuracy. In particular, the model combining CNN and Transformer showed excellent performance in processing and classifying speech data in complex real-world situations, which strongly suggests the practical applicability of scream detection technology.

In particular, it shows that it can directly contribute to CCTV or systems for detecting dangerous situations outside. Detecting screams in public or private spaces is crucial for rapid response and prevention of dangerous situations. These technologies provide the basis for detecting dangers in real time and taking necessary actions in large areas where human surveillance is difficult, or where people are vulnerable to danger.

Acknowledgment

This research was supported by Development of Social Complex Disaster Response Technology through the Korea Planning & Evaluation Institute of Industrial Technology funded by Ministry of the Interior and Safety in 2023. (Project Name : Development of risk analysis and evaluation technology for high reliability stampede accidents using CCTV and Drone imaging, Project Number : 20024403, Contribution Rate : 100 %)

References

[1] Lee, So-Min, et al. "Screaming data analysis for security system with audio capability." Proceedings of the Korean Society of Broadcast Engineers Conference. The Korean Institute of Broadcast and Media Engineers, 2013.

[2] Aki Harma, Martin F. McKinney, and Janto Skowronek, “Automatic surveillance of the acoustic activity in our living environment”, in IEEE International Conference on Multimedia and Expo, Amsterdam, July 2005.

[3] Juhyun Park, Jihoon Seo, and Seokpil Lee. "Analysis of environmental noise and screams for an audio acquisition-based crime prevention system." Journal of the Institute of Electrical Engineers 63.6 (2014): 804-809.

[4] D. Jang, J. Lee, and J.-S. Lee, “Acoustic Feedback Detection for Online Video Conferencing,” in Proc. of IEEE ICTC, 2021.

[5] D. Jang, J. Lee, and J.-S. Lee, “Development of Sound Event Detection for Home with Limited Computation Power,” in Proc. of KIBME conference, 2019

[6] P. Laffitte, Y. Wang, D. Sodoyer, and L. Girin, “Assessing the performances of different neural network architectures for the detection of screams and shouts in public transportation,“ Expert Syst. Appl. Vol 117, pp. 29-41, 2019

[7] S.-H. Chung, Y.-J. Chung, “A Comparison between Methods for Scream Detection Based on SVM and GMM,” Journal of KIIT. Vol. 15, No. 3, pp. 65-71, Mar. 31, 2017

[8] F. S. Saeed, A. A. Bashit, V. Viswanathan, and D. Valles, ‘‘An initial machine learning-based victim’s scream detection analysis for burning sites,’’ Appl. Sci., vol. 11, no. 18, pp. 1–22, Sep. 2021.

[9] AI-Hub[online] available: https://aihub.or.kr/

[10] MIVIA Dataset. [online] Available: https://mivia.unisa.it/datasets/audio-analysis/mivia-audio-events/

[11] Mathur, Rohan, Tejas Chintala, and D. Rajeswari. "Identification of Illicit Activities & Scream Detection using Computer Vision & Deep Learning." 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE, 2022.

[12] Huang, Yo-Ping, and Richard Mushi. "Deep Convolutional Neural Networks for the Classification and Detection of Human Vocal Exclamations of Panic in Subway Systems." IEEE Access (2023).

[13] Ebbers, Janek, and Reinhold Haeb-Umbach. "Pre-training and self-training for sound event detection in domestic environments." Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge (2022).

[14] Detection and Classification of Acoustic Scenes and Events [online] available: https://dcase.community/

[15] Zenkov, Ilia, Parallel is All You Want: Combining Spatial and Temporal Feature Representions of Speech Emotion by Parallelizing CNNs and Transformer-Encoders, (2020), GitHub repository, https://github.com/IliaZenkov/sklearn-audio-classification?tab=readme-ov-file

[16] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[17] Li, Mu, et al. "Efficient mini-batch training for stochastic optimization." Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 2014.