IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 6 June 2022, accepted 16 June 2022, date of publication 22 June 2022, date of current version 13 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3185224

Acoustic-Based Train Arrival Detection Using Convolutional Neural Networks With Attention

VAN-THUAN TRAN AND WEI-HO TSAI, (Member, IEEE)
Department of Electronic Engineering, National Taipei University of Technology, Taipei 10608, Taiwan
Corresponding author: Wei-Ho Tsai (whtsai@ntut.edu.tw)

This work was supported in part by the Ministry of Science and Technology, Taiwan, under Grant MOST 108-2221-E-027-064.

ABSTRACT In the places of railroad crossings, audible warning signals such as train whistles and railway alarms are utilized to warn the road users of paying attention and giving priority to the approaching train(s). However, road users may sometimes be unaware of warning signals due to various reasons, resulting in inappropriate cooperation or even traffic collision between railway vehicles and non-railway vehicles. This work studies deep learning-based approaches to develop systems for acoustic-based train arrival detection (A-TAD). Firstly, we develop a novel audio dataset of train horns, railway alarms, railway noise, and other urban noises to conduct A-TAD experiments. We then examine the efficiency of handcrafted acoustic features (i.e. MFCC and Mel-spectrogram) in building A-TAD's audio classifier, the MSNet, which is based on two-dimensional convolutional neural networks (2D-CNN). Next, we propose to apply the attention mechanism and utilize MFCC and spectrogram simultaneously to enhance the classification accuracy, in which the combined use of acoustic features is considered at the input level (with InCom-TADNet), high-level feature level (with FCCom-TADNet), and decision level (with DLCom-TADNet). Our experiments have shown the efficiency of MSNet and attention mechanism as the MSNet trained with the single feature is more performant than the baseline models and applying attention modules results in better accuracies. Also, the combined use of MFCC and spectrogram significantly improve the system's accuracy and robustness. A-TAD systems can be utilized to extend the safety function of the railway crossing systems, private cars, and self-driving cars, and particularly be useful for hearing-impaired road users.

INDEX TERMS Audio classification, attention mechanism, convolutional neural networks, feature aggregation, railway audible warning signals, railway safety, train arrival detection.

I. INTRODUCTION

Train arrival detection (TAD) is an essential problem for railway safety, involving railway passengers, road users, and railway employees such as field operators and maintenance personnel. For traffic safety in general, the early detection of train arrival is used to warn road users of approaching trains, so they can pay attention and cooperate appropriately, especially at the level crossings. From the side of the road users, they could recognize train presence directly by warning signals from the trains like train horn sounds, or indirectly by other warning signals such as audible alarms, flashlights, traffic lights, or public address systems, which are generated by safety systems. However, sometimes, road users may not catch the warning signals. For example, the visual warnings may occasionally be out of drivers' vision, and drivers may be unaware of the audible warnings due to the noisy environment, interference of the in-car audio signal, soundproofing of modern cars, or the distraction of drivers themselves. The unawareness of train arrival can cause potential traffic collisions between trains and other vehicles. The early detection of oncoming trains also contributes to the safety in the railroad working environment since the railway employees, like maintenance workers, are almost always exposed to the danger of injury or crash when their activities are required to conduct in parallel with the transit of trains to ensure smooth traffic.

TAD can be conducted using human operations or automatic systems, in which the former approach has a low level of flexibility and could be suffered from the operator's

The associate editor coordinating the review of this manuscript and approving it for publication was Jesus Felez.

distraction or absence. The latter approach is applying the sensing techniques and detection methods that are more popular in modern train arrival detection systems. In sensing techniques [1], the traditional methods include the use of treadle mechanisms, inductive sensors, and infrared beam sensors, while more modern systems utilize radar technologies [2], acoustic sensing [3], time-domain reflectometry [4], anisotropic magneto-resistive magnetometer [5], and rail vibrations with accelerometers [6]. This work focuses on the use of the acoustic sensing technique for TAD based on the detection of audible train warning signals, which can be deployed flexibly in real-world applications. Acoustic-based TAD (A-TAD) can be not only set up along the fixed locations of railway systems but can also be integrated into moving objects like private cars or railway maintenance cars. By contrast, other sensing-based systems such as treadle mechanisms, inductive sensors, or the measurement of rail vibration are only suitable for specific installation locations like on the rail tracks.

Given an audio segment captured from the surrounding environment, the role of the objective A-TAD system is to determine whether that audio segment is an audible train warning (ATW) signal or not, which can be viewed as a kind of prediction about which of two classes, the ATW class and not ATW (or NonATW) class, the input audio belongs to. Thus, we can formulate A-TAD as a binary audio classification problem with ATW and NonATW classes, in which the first class involves ATW signals like train horn sounds and railway alarm sounds, while the second class covers any types of noises rather than ATW, such as traffic noises, environmental sounds, and other sources of urban noises. Figure 1 illustrates the overall structure of the A-TAD system, in which a microphone is used to continuously capture equal-length audio chunks from the surrounding environment, and the features of each chunk are extracted to feed into the binary audio classifier which predicts probabilities for ATW and NonATW classes. The system determines a positive detection of train arrival if the decision rule applied on the classifier's outputs assigns the input signal to ATW class.

The contributions of this work can be listed as follows. First, for experiment data, since there is no public dataset related to the A-TAD problem, we develop a dataset for experiments and evaluation. We collect data from sound sources reflecting real soundscapes nearby the railway system and urban traffic, so proposed networks trained on our dataset could meet the requirement for practical applications. Second, we propose a two-dimensional CNN-based model (i.e. MSNet) to use as the classifier of the A-TAD system, in which the model can be trained with Mel-frequency cepstral coefficients (MFCCs) or spectrogram features. We further examine the use of attention mechanisms and prove that using attention blocks in MSNet brings about significant improvement in classification accuracies. Equally important, all variants of MSNet configured with and without attention blocks are more performant than the baseline models. Next, we propose the combined use of acoustic features, in which we utilize MFCC and spectrogram features together in three styles, including input combination with InCom-TADNet, the combination of high-level features at fully connected layers using FCCom-TADNet, and decision level fusion with DLCom-TADNet. All InCom-TADNet, FCCom-TADNet, and DLCom-TADNet achieve better accuracies compared to those of networks trained on a single feature set, and the FCCom-TADNet obtains the highest accuracy and reaches a high level of robustness. Last but not least, the results of this work can lay a good foundation for the further development of A-TAD systems and can be applied for useful applications. We can improve the safety functions of modern cars by using the A-TAD system to alert drivers who are unintentionally unaware of the approaching trains. This application is especially useful for drivers and road users with hearing impairment. For the safety of railway service, the A-TAD system can be employed individually or combined with other detection approaches to provide early accurate warning of train arrival to employees and passengers. A-TAD can be also integrated into smart railway crossing solutions like automated railway gate control.

We organize the rest of this paper as follows. The related works are introduced in Section II. Section III analyzes the methods we use for classifying the audible train warning signals and noises. Then, we present the experimental results in Section IV and provide a conclusion in Section V.

II. RELATED WORKS

A-TAD broadly belongs to the audio classification problem which is also known as sound classification (SC), so prior works on SC can lay a good foundation for the development of A-TAD systems. Sound classification has recently received increasing research attention and applied in a wide variety of applications, such as surveillance [7], predictive maintenance [8], smart home security systems [9], emergency vehicle detection [10], [11], environmental sound classification (ESC) [12]–[14], and speech recognition [15]. The conventional solutions for sound classification include signal processing techniques and traditional machine learning methods, such as the Gaussian mixture model (GMM) [16], [17], support vector machine (SVM) [18], and hidden Markov model (HMM) [19]. Recently, deep learning has been applied and shown outstanding performance in many applications, and sound classification is not an exception. Deep learning models can extract useful discriminative features from a large amount of training data, process the input of high dimensions, and have excellent ability of generalization. In many works of sound classification, such as in [10], [11], and [13], deep learning-based systems have shown much better performances compared to those based on conventional machine learning techniques.

Generally, existing works employ deep learning methods for sound classification in two directions. In the first direction, handcrafted acoustic features are extracted from

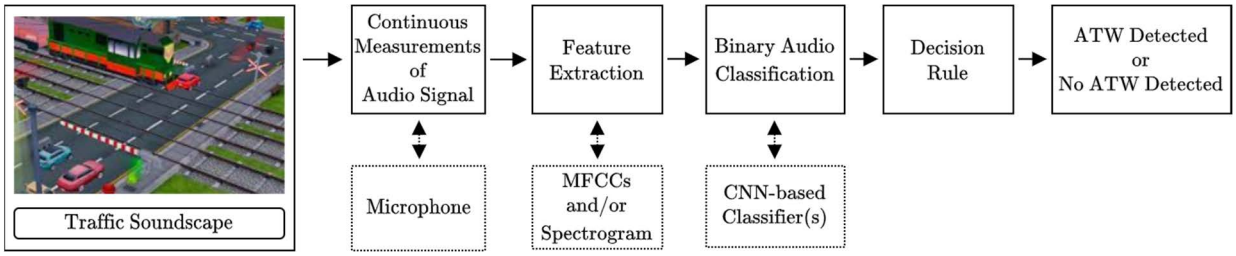


FIGURE 1. The overall structure of the A-TAD System. ATW stands for audible train warning.

the audio signal and fed into neural networks, in which the widely used approach is to transform the audio into time-frequency representations, such as log-mel spectrogram or log-gammatone spectrogram, which are utilized as inputs of two-dimensional CNN (2D-CNN) classifiers. One of the early works in the first direction is [13] which proposed a 2D-CNN model with log-mel spectrogram input and outperformed the conventional machine learning-based classifiers in different environmental sound datasets. Since sound classification with the use of image-like features and 2D-CNN is similar to the image classification problem, several networks for image classification like GoogLeNet and AlexNet have been examined in sound classification [14], resulting in promising results. Besides 2D-CNN, recurrent neural networks (RNN) such as long short-term memory neural networks (LSTM) [20], which can efficiently learn the temporal dependencies in the spectrogram, have been also adopted for sound classification.

The second direction is to use the raw audio waveform as the input of the networks, in other words, deep networks directly learn discriminative representations from 1D raw data rather than from handcrafted features. [21] proposed very deep fully convolutional networks for sound classification with raw waveform input. [21]'s networks achieved comparable performance compared to that of the baseline model [13] which was based on log-mel spectrogram input, but utilized much deeper architectures, with up to 34 1D convolutional layers. More recently, the end-to-end approach using 1D-CNN proposed in [22] obtained competitive accuracy with most of the state-of-the-art approaches in environmental sound classification. Models based on the combined use of 1D-CNN and 2D-CNN have been also proposed and showed promising results, such as in [10] and [23], in which 1D-Conv layers and max-pooling layers are used to convert raw data to 2D representation which is classified by 2D Conv layers. In [10], the combined use of handcrafted features and raw data for training an audio classifier has been also proposed, showing significant improvement in classification accuracy.

Recently, more and more techniques have been proposed and incorporated into neural networks to improve the performance in sound classification tasks. Audio data augmentation is used to increase the amount of training data and mitigate the problem of overfitting. [12], [24], and [25] have applied different augmentation techniques, including both time-domain and frequency-domain deformations such as time-stretching, noise adding, and pitch-shifting, which are efficient at improving the performances of sound classifiers. Being inspired by the efficiency of the attention mechanism in a wide variety of applications like machine translation, automatic speech recognition, and document classification, some works have investigated the use of attention in audio classification problems and obtained favorable results. In [26], the temporal attention mechanism is applied to the LSTM layers of the convolutional recurrent neural network to predict the importance of each time step, in which a shallow neural network with a single fully connected layer and a linear output layer is adopted to compute attention weights, from which final output is calculated using the weighted sum of hidden states of LSTM layers along the time dimension. The temporal attention for convolutional layers is introduced in [27], which proposed to calculate an attention vector based on the input spectrogram and produce the attention feature maps using dot-product between attention vector and output of convolutional layer along the time dimension. [28] applied temporal attention for both LSTM and convolutional layers to improve the classification performance. In addition, some works have shown that using multiple feature sets together can efficiently improve the performance of sound classification. The combined use of multiple features can come in different styles consisting of input-level feature aggregation [24], [29], high-level feature combination [27], and decision-level fusion [10], [29]. This work incorporates 2D-CNN with time-frequency audio features, attention mechanism, feature combination, and data augmentation techniques to build audio classifiers for A-TAD systems.

III. METHODOLOGY

A. CONVOLUTIONAL NEURAL NETWORKS FOR A-TAD

Convolutional neural networks (CNN), a variant of feedforward neural networks, are normally composed of convolutional (Conv) layers, pooling layers, and fully-connected (FC) layers. Roughly, typical CNN architectures for classification tasks have two parts: the first part contains different blocks of Conv layers and pooling layers for feature extraction and dimensional reduction; the second part is a simple neural network of FC layers that processes features extracted by

the first part and produce predictions on classification. This work proposes to build 2D-CNN classifiers for the A-TAD system, in which the networks' input can be MFCCs and/or Mel-spectrogram. Assuming that the input of the CNN is a tensor of the shape $(d_H^{(0)}, d_W^{(0)}, d_C^{(0)})$ where $d_H^{(0)}$, $d_W^{(0)}$, and $d_C^{(0)}$ indicate the input's height, width, and the number of channels, respectively. Note that $d_H^{(0)} > 1$ for 2D input. When the feature extraction part of CNN gets deeper, the dimension $(d_H^{(l)}, d_W^{(l)}, d_C^{(l)})$ of feature map at the l^{th} layer normally decreases in height and width but increases in the number of channels, i.e. $d_H^{(l)} < d_H^{(l-1)}$, $d_W^{(l)} < d_W^{(l-1)}$, and $d_C^{(l)} > d_C^{(l-1)}$. Given input-output pairs $(\boldsymbol{X}, y)$ in the training dataset, a CNN of L layers is trained to approximate the relationship between every input $\boldsymbol{X}$ and its corresponding output y. We describe this approximation by a composite nonlinear function $G(\cdot \mid \theta)$ of input $\boldsymbol{X}$ and parameter θ, as expressed in (1), in which the function or operation $g^{(l)}\left(\cdot \mid \theta^{(l)}\right)$ is referred to as the l^{th}, $l = \{1, 2, \ldots, L\}$ layer of the network, and $\theta^{(l)}$ represents parameters of this layer.

$$\begin{aligned} \mathrm{y} \approx \hat{y} &= G(\boldsymbol{X} \mid \theta) \\ &= g^{(L)}\left(g^{(L-1)}\left(\ldots\left(g^{(2)}\left(g^{(1)}\left(\boldsymbol{X} \mid \theta^{(1)}\right)\right)\right)\right)\right) \end{aligned} \tag{1}$$

$$\begin{aligned} \boldsymbol{a}^{(l)} &= g^{(l)}\left(\boldsymbol{a}^{(l-1)} \mid \theta^{(l)}\right) \\ &= \left(\boldsymbol{W}^{(l)} \circledast \boldsymbol{a}^{(l-1)} + \boldsymbol{b}^{(l)}\right), \theta^{(l)} = \left[\boldsymbol{W}^{(l)}, \boldsymbol{b}^{(l)}\right] \end{aligned} \tag{2}$$

$$\boldsymbol{a}^{(l)} = \psi^{(l)}\left(\left(\boldsymbol{W}^{(l)}\right)^T \boldsymbol{a}^{(l-1)} + \boldsymbol{b}^{(l)}\right), \theta^{(l)} = \left[\boldsymbol{W}^{(l)}, \boldsymbol{b}^{(l)}\right] \tag{3}$$

$$\begin{aligned} &p(c = ATW \mid \boldsymbol{X}) \\ &= \boldsymbol{a}^{(L)} = \sigma\left(\left(\boldsymbol{W}^{(L)}\right)^T \boldsymbol{a}^{(L-1)} + \boldsymbol{b}^{(L)}\right) \end{aligned} \tag{4}$$

$$c^* = \begin{cases} ATW \text{ if } p(c = ATW \mid \boldsymbol{X}) \geq \mathit{threshold} \\ NonATW \text{ if } p(c = ATW \mid \boldsymbol{X}) < \mathit{threshold} \end{cases} \tag{5}$$

If the l^{th} layer is a Conv layer, its operation is expressed by (2) where the layer's input $\mathbf{a}^{(l-1)}$ is a 3D tensor of $d_C^{(l-1)}$ channels, $\mathbf{W}^{(l)}$ is a set of $d_C^{(l)}$ filters, $\mathbf{b}^{(l)} \in R^{d_C^{(l)} \times 1}$ is a bias vector, $\circledast$ indicates the convolutional operation, and $\psi^{(l)}$ is the element-wise activation function. In case the l^{th} layer is an FC layer of $d^{(l)}$ nodes, its operation is expressed by (3), where the input $\mathbf{a}^{(l-1)}$ is a column vector of length $d^{(l-1)}$, $\mathbf{W}^{(l)}$ is a 2D weight matrix of shape $(d^{(l-1)}, d^{(l)})$, and $\mathbf{W}^{(l)}\mathbf{a}^{(l-1)}$ is a matrix product. Similar to Conv layers, $\mathbf{b}^{(l)} \in R^{d^{(l)} \times 1}$ and $\psi^{(l)}$ in FC layers are the bias vector and the element-wise activation function, respectively. Note that at the input layer ($l = 0$) we have $\mathbf{a}^{(0)} = \boldsymbol{X}$, and the number of neurons at the output layer is normally equal to the number of sound classes, except for the binary classifier which may have a single output node. Since we treat A-TAD as a binary classification problem the objective model is designed with an output layer of one neuron and a sigmoid activation function. For each audio segment $\boldsymbol{X}$, the model generates the probability $p(c = ATW \mid \boldsymbol{X})$ representing how likely the input $\boldsymbol{X}$ is an audible train warning (ATW), from which the system can determine the status of train arrival detection based on the decision rule (5), where $c \in \{ATW, NonATW\}$ and $\mathit{threshold} \in [0, 1]$.

B. THE MSNET AND CONVOLUTIONAL ATTENTION BLOCK

Figure 2 illustrates the general structure of the proposed 2D-CNN model, namely the MSNet which can be trained with MFCCs or spectrogram inputs. For ease of explanation, we illustrate the MSNet with the spectrogram input. The extraction of spectrogram input for the MSNet is described as follows. Generally, an input segment $\boldsymbol{X}$ of t seconds is split into smaller overlapping frames using a sliding window of P data samples and the hop length of Q data samples, resulting in $M = \lceil (t \times SR)/Q \rceil$ successive frames, where $\lceil \cdot \rceil$ is the ceiling function and SR is the sampling rate. For each frame, we extract log-mel features of N components, so the spectrogram for the whole input segment $\boldsymbol{X}$ has the shape of $(N, M, 1)$ corresponding to the (feature, time, channel) representation which is the channel-last input format for the 2D-CNN-based MSNet. In our experiments, the input length is 1 second ($t = 1$), and the sampling rate SR is 22,05 kHz which is commonly used in the audio domain. We set the window length to 552 samples or 0.25ms, and the hop length to 276 samples, so a 1-second input segment is split into $\lceil (1 \times 22,050)/276 \rceil = 80$ frames of 50% overlapping. Consequently, the extracted spectrogram has the shape of $(128, 80, 1)$ in the (feature, time, channel) format. The extraction of MFCC features is conducted similarly to the spectrogram features, in which each frame is extracted with 40 MFCCs, resulting in the MFCC features of shape $(40, 80, 1)$ for a segment of 1 second.

From Figure 2 we can see that the sp ectrogram is processed by a series of 2D-Conv layers to extract useful discriminative features which are then flattened and fed into the fully connected layers for classification. The MSNet contains eight 2D-Conv layers organized in four pairs separated by pooling layers, in which layers in the next pair double the number of filters in the previous pair. 8, 16, 32, and 64 are the number of filters in each layer of the first, second, third, and fourth pairs of 2D-Conv layers, respectively. In each 2D-Conv layer, we set the receptive field to $(3, 3)$, and stride to $(1, 1)$. Among three FC layers of the MSNet, the last layer has a single neuron with the sigmoid activation function to output classification probability $p(c = ATW \mid \boldsymbol{X})$ for audible train warning class.

We also investigate the use and efficiency of the attention mechanism in the MSNet, in which attention blocks can be placed at the input layer or after convolutional layers. The attention used in this work is inspired by the idea of temporal attention and frame-level attention proposed in [27] and [28]. Figure 3 illustrates the structure of an attention block at the output of the l^{th} convolutional layer. It is assumed that

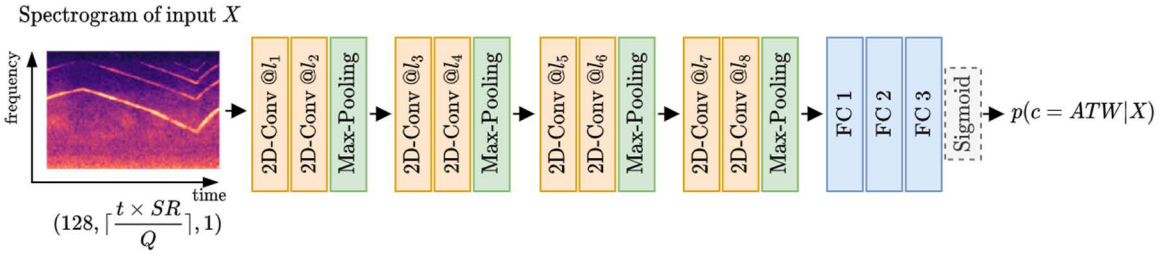


FIGURE 2. Illustration of the MSNet working with spectrogram input. *SR* is the sampling rate, *Q* is the hop length, and *t* is the input length (in second).

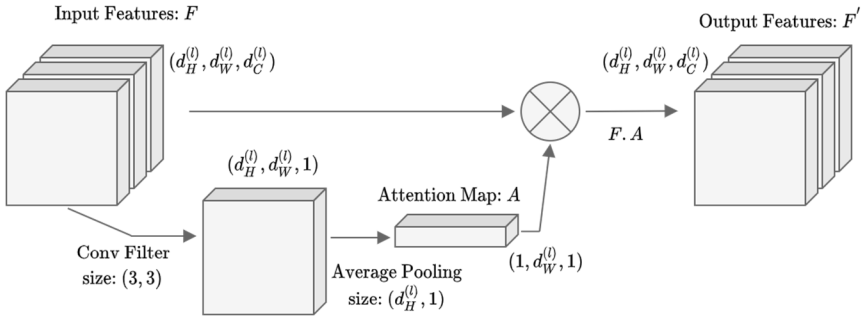


FIGURE 3. The illustration of attention at convolutional layers.

the input features F of attention block at the l^{th} convolutional layer has the shape $(d_H^{(l)}, d_W^{(l)}, d_C^{(l)})$ corresponding to frequency, time, and channel dimensions, respectively. First, a convolutional filter of size $(3, 3)$ is applied to input F to extract feature map M of shape $(d_H^{(l)}, d_W^{(l)}, 1)$. Next, the feature map M is processed using average pooling along the frequency dimension to create a lower-dimensional feature map which is further processed by the sigmoid function to form the attention map or attention vector A of shape $(1, d_W^{(l)}, 1)$. Note that each element of attention map A represents the attention weight for the corresponding frame of input features F along the time dimension. Finally, the input features F are multiplied with the attention vector A to create the attention-weighted feature F'. The operation of the attention block is presented by (6) and (7), where σ denotes the sigmoid function. This work examines the effect of attention blocks when they are applied in different positions of the MSNet.

$$\begin{aligned} A &= \sigma\left(AvragePool\left(M\right)\right) \\ &= \sigma\left(AvragePool\left(Conv\left(Filter^{3\times3}, F\right)\right)\right) \end{aligned} \tag{6}$$

$$F\prime = F.A \tag{7}$$

C. THE COMBINED USE OF MFCC AND SPECTROGRAM FEATURES

It is assumed that the widely-used MFCC and spectrogram features could possess different patterns and information which are useful for A-TAD tasks, so we further explore the combined use of those two feature sets by building and evaluating various CNN models trained with both MFCC and spectrogram inputs. Considering different positions in the models where two sets of features or their high-level representations are combined, we examine three types of combinations as follows.

The first case of combination is illustrated in Figure 4, in which the spectrogram and MFCC features are aggregated at the input level, then the combined features are fed into 2D-Conv layers of the network. We refer to this network as the InCom-TADNet. As shown in Figure 4, from the position of the combined features, the InCom-TADNet works similarly to the MSNet. In the second combination approach, we utilize a two-stream structure where each stream processes a feature set and outputs a high-level feature vector. We then concatenate output vectors of two network streams and feed the result into a fully connected layer for computing the final output, as described in Figure 5. Since we combine features from fully connected (FC) layers, the model in the second approach is referred to as FCCom-TADNet. Lastly, we examine an ensemble model, namely DLCom-TADNet, which is based on the decision-level combination between a network stream of the MFCCs input and another network stream that processes spectrogram input. The structure of DLCom-TADNet is presented in Figure 6, showing that the predictions produced by two streams are combined to generate the final classification probability. Specifically, the DLCom-TADNet's classification decision on an audio segment is based on the average value of outputs from the two network streams. We denote $p(c)$ as the final classification probability for class c, $c \in \{ATW, NonATW\}$. Similarly,

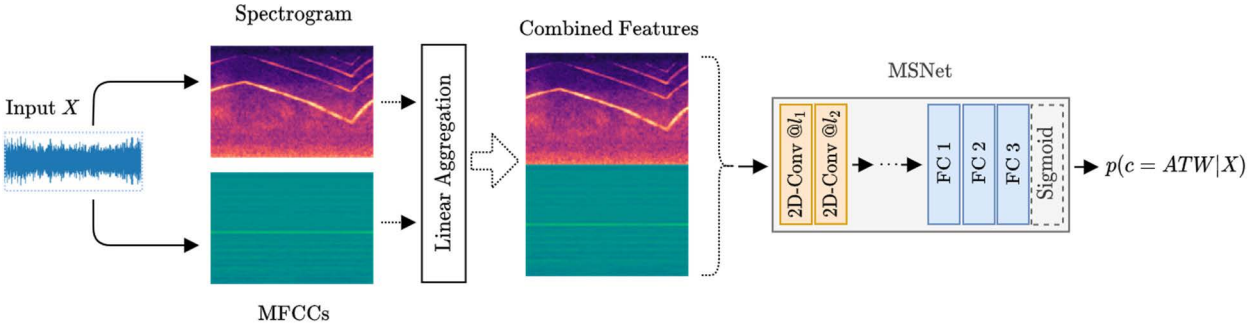


FIGURE 4. Illustration of the InCom-TADNet.

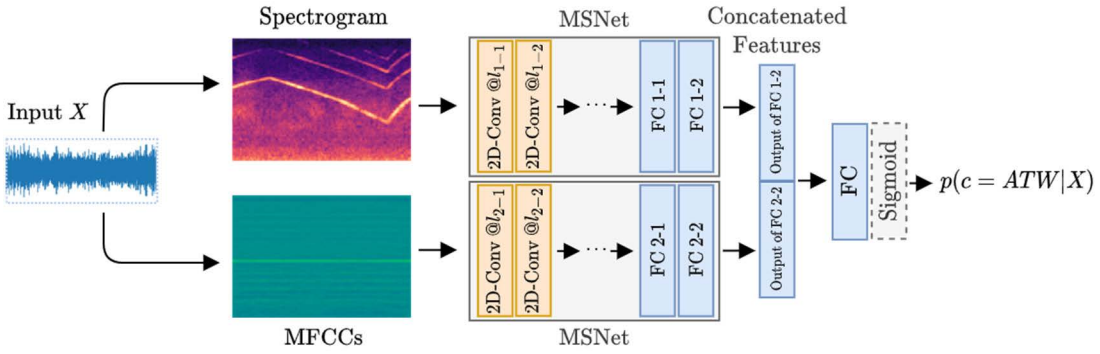


FIGURE 5. Illustration of the FCCom-TADNet.

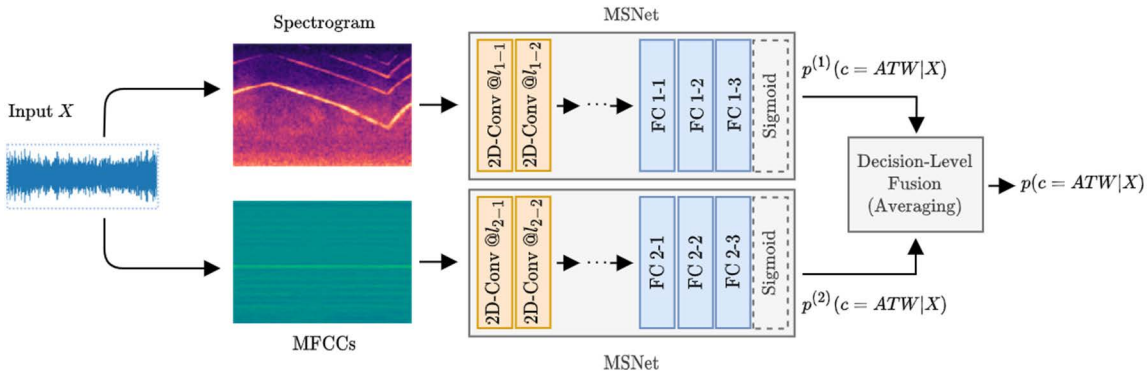


FIGURE 6. Decision-level fusion model for A-TAD (DLCom-TADNet).

$p^{(i)}(c)$ presents the prediction for class c, which is yielded by the $i^{th}, i \in \{1, 2\}$ network stream. The combination of two streams' outputs is expressed by (8), and the final prediction of the DLCom-TADNet is determined by decision rule (5).

$$p\left(c = ATW \mid \boldsymbol{X}\right) = \frac{1}{2}\sum_{i=1}^{2} p^{(i)}\left(c = ATW \mid \boldsymbol{X}\right) \qquad (8)$$

IV. EXPERIMENTS AND RESULTS

A. EXPERIMENTAL DATA COLLECTION

The experimental dataset contains two sound classes, audible train warning (ATW) and noises (NonATW). We set several criteria for data collection as follows. First, all data must be real-field recordings captured at the traffic nearby railroad systems and related urban environments. Second, in order to build deep learning models with a high level of generalization, the dataset must be adequately large and covers various recording conditions. To achieve those two primary goals, we collect data using three approaches: (1) extracting data from online resources specialized in audio/video clips of train arrivals; (2) directly recording data in real traffic; (3) combining the self-collected data with published datasets that contain various sources of urban noise.

TABLE 1. Data for A-TAD experiments.

Data Class	Data Sources			
	Our ATAD dataset	UrbanSound8K [30]	ESC-50 [31]	Total (#samples)
ATW	15,183	0	0	15,183
NonATW (Noise)	5,252	8,732	2,000	15,984
Total (#samples)	20,435	8,732	2,000	31,167
Total duration	11.35 hours	9.7 hours	2.8 hours	23.85 hours
Clip length	2 seconds	1-4 seconds	5 seconds	-
Sampling rate	44.1 kHz	8-192 kHz	44.1 kHz	-

TABLE 2. Data separation for A-TAD experiments.

Subset	ATW	NonATW	Total
Train	11,245	10,951	22,196
Validation	1,937	2,488	4,425
Test	2,001	2,545	4,546
Total	15,183	15,984	31,167

In the first approach, thanks to the availability of YouTube channels providing a large number of videos about train arrival recorded all over the world, we have access to a diverse database of train horns, railway alarms, and traffic noises. In the second data preparation approach, we captured noise recordings of Taiwan's traffic using mobile phones or a microphone plugged into a laptop. 60 recordings were conducted nearby the railway stations, road intersections, or when we drove on the streets, in which each recording lasted for three minutes under normal weather conditions. Thus, approximately 3 hours of noise recordings were recorded. Finally, we combined the data collected by ourselves with that of Urbansond8K [30] and ESC-50 [31] datasets. UrbanSound8K and ESC-50 contain useful subsets of urban noises and environmental sounds, such as sounds of car horns, engine idling, siren, street music, air conditioner, drilling, jackhammer, and thunderstorm. Therefore, data in those two datasets can complement the traffic noises recorded by ourselves to form a diverse set of background noises.

The data extracted from online sources and our recordings were split into non-overlapping clips of 2 seconds, resulting in 15,183 and 5,252 samples for ATW and NonATW classes, respectively. Then, we combined the self-collected data with that of Urbansond8K [30] and ESC-50 [31] to form the final dataset of 31,167 samples, as shown in Table 1. We organize the experimental data into three subsets for training, validation, and testing, respectively, each of which has a similar amount of audio length for ATW and NonATW classes. Also, the original recordings in a subset are entirely different from those of the remaining subsets, to assure the isolation between data for development and inference. The detail of the training set, validation set, and testing set are shown in Table 2.

B. EXPERIMENT SETUP

Note that the experimental data is collected from different sources and has various properties like sampling rates, the number of channels, and bit-depth. Thus, we standardize the whole dataset into monophonic signals at a sampling rate of 22,05 kHz, which is performed using Librosa [32], a library for audio signal processing. Librosa is also utilized in the process of acoustic feature extraction, in which we set the frame length to 25ms and the frame overlapping rate to 50%. Although most of the recordings in the experimental dataset are between 2s and 5s, we only examine the input length of 1s for two reasons. It is more computationally efficient to utilize short inputs as they help to reduce the computational complexity of the models. Short inputs are also suitable for practical A-TAD applications with requirements of a quick response and real-time processing. As described in section III. B, for a sample of 1 second, we extract spectrogram and MFCCs of shapes (128, 80, 1) and (40, 80, 1), respectively.

We performed experiments using a desktop PC built with 16 GB RAM, an Intel Core i7-9700K CPU (8 cores @3.60 GHz), and NVIDIA GeForce GTX 2080 Ti. The basic setup to train deep networks in our experiments is as follows: we use the cross-entropy loss function; Adam optimizer [33] is employed to update model parameters, in which the initial learning rate is 0.00001. Data augmentation is applied to increase the amount and the diversity of training data, which is achieved by randomly adding noise recordings to the original training samples at random signal-to-noise ratios (SNRs) to generate noisy samples. We additionally utilize batch normalization to speed up the training process, and dropout regularization [34] is applied to alleviate overfitting. To evaluate the robustness of the proposed models, we report their performances on noisy testing sets of different SNRs consisting of +15dB, +10dB, +5dB, 0dB, -5dB, -10dB, and -15dB. The original testing samples reflect the real traffic soundscape, so they already contain certain levels of noise. Therefore, noisy testing sets generated by the noise addition approach create more challenging evaluation conditions for the proposed models. Equally important, the SNRs used in noisy test sets do not duplicate with SNR values used for training data augmentation, which means that the SNRs of testing sets are unseen by the pre-trained models.

Since the experimental dataset is relatively balanced, we employ classification accuracy as the primary metric for model evaluation, the accuracy (in %) of the A-TAD system is characterized by (9).

$$\text{Accuracy} = \frac{\#\text{Correctly\_classified samples}}{\#\text{Testing samples}} \times 100 \quad (9)$$

C. PERFORMANCE OF THE PROPOSED MSNET

At first, we performed an initial experiment to find the optimal configuration for the MSNet which is generally illustrated in Figure 2 of section III.2. We evaluated the performance of MSNet variants with 4, 6, 8, and 10 2D-Conv layers, respectively. We stopped examining the number of 2D-Conv layers at 10 layers because with this configuration

TABLE 3. Results of MSNet for different configurations.

Model configuration		Accuracy (%) associated with features	
#2D-Conv Layers	#FC Layers	Spectrogram	MFCC
4	3	91.82	92.76
6	3	92.63	92.81
8	3	**93.58**	**93.71**
10	3	92.74	92.21

the output of the last convolutional layer reaches a very small size already. The results of this experiment are summarized in Table 3, from which we can see that the image classification-based approach using 2D time-frequency input and 2D-CNN is useful for the A-TAD problem as all models yield accuracies above 90%. Among all models, the MSNet with 8 2D-Conv layers achieves the highest accuracy of 93.58% and 93.71% for spectrogram input and MFCC input, respectively. For the first three models configured with 4, 6, and 8 2D-Conv layers, the larger the number of Conv layers is, the higher the accuracies models achieve. However, when it comes to 10 Conv layers configuration, the network gets a significant decrease in accuracy, by around 1%. This can be explained by the overfitting problem due to a large-capacity network trained on a moderate amount of data. For ease of explanation in later experiments and analysis, MSNet is referred to as the 2D-CNN network of 8 2D-Conv layers and three FC layers as presented in Figure 2.

We evaluated the performances of some baseline models on the A-TAD dataset to make the comparison with the proposed MSNet. We considered several models based on image-like inputs (i.e. spectrogram and MFCCs), including 2D-CNN models [12], [35], [36] and can RNN model [20], in which models in [12], [20] were directly proposed for environmental sound classification tasks, while AlexNet [35] and VGG [36] are well-known models in the field of image classification. For the models which work with raw audio input, two 1D-CNN models [37], the SoundNet variants, and a model with the combined use of 1D-CNN and 2D-CNN [23] were also taken into consideration.

From Table 4 we can see that, except for VGG, the proposed MSNet yields better accuracies compared to those of baseline models for all kinds of input features, showing the efficiency of the MSNet in A-TAD. VGG and MSNet achieve comparable accuracies of above 93%, but the MSNet has a much smaller number of parameters. For example, with the spectrogram input of shape $(128, 80, 1)$, the VGG network has 15.79 million parameters while that for the MSNet is 1.47 million. Similarly, another high-capacity model, the AlexNet attains smaller accuracies compared to those of the MSNet. Thus, it is likely not necessary to utilize very high-capacity models, such as VGG and AlexNet, for a moderate dataset of two audio classes in A-TAD. In addition, for all four 2D-CNN models, employing MFCCs input results in slightly higher accuracies than using the spectrogram input.

TABLE 4. Results of baseline models on the A-TAD dataset. The number of model parameters attribute is shown in millions (M).

Models	#Parameters	Features	Input shape	Accuracy (%)
2D-CNN (AlexNet [35])	22.15M	Spectrogram	$(128, 80, 1)$	92.15
2D-CNN (AlexNet [35])	12.97M	MFCCs	$(40, 80, 1)$	92.30
2D-CNN (VGG16 [36])	15.79M	Spectrogram	$(128, 80, 1)$	93.84
2D-CNN (VGG16 [36])	15.01M	MFCCs	$(40, 80, 1)$	93.88
2D-CNN ([12])	0.58M	Spectrogram	$(128, 80, 1)$	91.83
2D-CNN ([12])	0.21M	MFCCs	$(40, 80, 1)$	92.21
RNN ([20])	0.16M	Spectrogram	$(128, 80, 1)$	85.90
RNN ([20])	0.16M	MFCCs	$(40, 80, 1)$	82.36
1D-CNN (SoundNet 5 layers [37])	6.38M	Raw data	$(1, 22050, 1)$	90.21
1D-CNN (SoundNet 8 layers [37])	13.48M	Raw data	$(1, 22050, 1)$	90.58
1D-CNN and 2D-CNN (EnvNet [23])	46.14M	Raw data	$(1, 22050, 1)$	90.84
2D-CNN (MSNet of this work)	1.47M	Spectrogram	$(128, 80, 1)$	93.58
2D-CNN (MSNet of this work)	0.49M	MFCCs	$(40, 80, 1)$	93.71

TABLE 5. Effects of attention blocks on performances of the MSNet.

Model Settings	Accuracy (%) associated with different features	
	Spectrogram	MFCC
No attention	93.58	93.71
Attention at l_0 layer (input)	93.75	94.19
Attention at l_2 layer	93.84	94.13
Attention at l_4 layer	94.15	93.97
Attention at l_6 layer	94.61	94.24
Attention at l_8 layer	94.28	94.48
Attention at l_0, l_2, l_4, l_6, l_8 layers	94.74	94.88

By contrast, the opposite trend is observed for the RNN model. Also, among all models, 2D-CNN models are more performant than the RNN model and 1D-CNN model.

D. EFFECTS OF ATTENTION MODULES

Next, we analyzed the effect of convolutional attention blocks in the MSNet, in which we compared the results of the proposed MSNet without and with attention blocks applied at different positions of the network. Specifically, we performed separate experiments about applying attention blocks to l_0 layer (i.e. input layer), l_2, l_4, l_6, and l_8 layers of MSNet, respectively.

As shown in Table 5, the use of attention blocks results in considerable improvement in the classification accuracies. For both types of input features, when using an attention block at only one position, l_0, l_2, l_4, l_6, or l_8 layer, we always obtain better accuracy compared to that of the network without attention. For spectrogram input, except for l_8 layer,

TABLE 6. Effects of feature combination for A-TAD.

Models	Features	Position of combination	Accuracy (%)
InCom-TADNet	Spectrogram +MFCC	at input layer	96.39
FCCom-TADNet (Concat)	Spectrogram +MFCC	at FC layer	96.55
FCCom-TADNet (Add)	Spectrogram +MFCC	at FC layer	95.40
DLCom-TADNet	Spectrogram +MFCC	Decision-level fusion	95.32
MSNet without attention	Spectrogram	Not applicable (NA)	93.58
MSNet with attention	Spectrogram	NA	94.74
MSNet without attention	MFCC	NA	93.71
MSNet with attention	MFCC	NA	94.88

applying attention to the deeper layers tends to achieve higher accuracies. Whereas, no specific increasing trend is observed for experiments with MFCC input. We further investigated the case that we utilized attention blocks for all l_0, l_2, l_4, l_6, and l_8 layers. In this case, MSNet yields accuracies of 94.74% for spectrogram input and 94.88% for MFCC input, which are higher than the results of all cases in that we apply a single attention block. Using attention at all five positions brings about improvements of 1.16% (spectrogram input) and 1.17% (MFCC input) compared to the results of MSNet without attention. This experiment has shown the efficiency of convolutional attention blocks to boost the accuracy of the standard 2D-CNN network in A-TAD application.

E. EFFECTS OF THE COMBINED USE OF SPECTROGRAM AND MFCC FEATURES

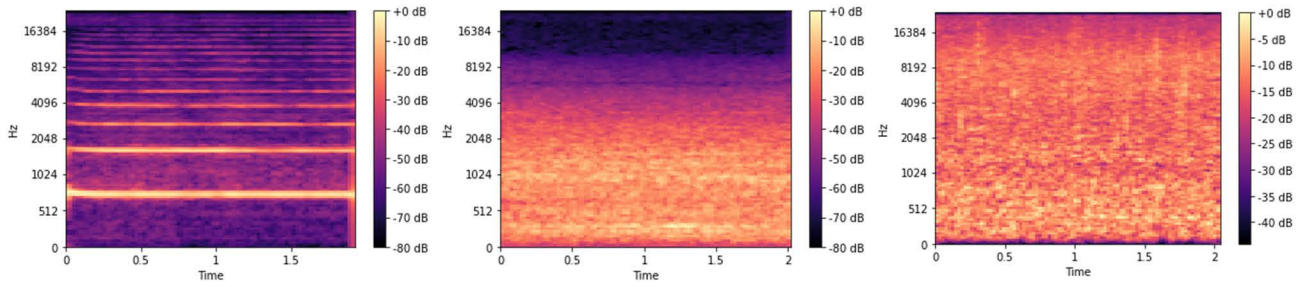
Table 6 shows the performances of proposed InCom-TADNet, FCCom-TADNet, and DLCom-TADNet, three models based on the combined use of spectrogram and MFCC features. All three models achieve promising classification accuracies, in which models based on feature-level combinations are more performant than the model based on the decision-level combination. Specifically, the InCom-TADNet and FCCom-TADNet (Concat) obtain the accuracies of 96.39% and 96.55% which are 1.07% and 1.23% higher than the result of the DLCom-TADNet, respectively. Among three cases of using spectrogram and MFCC together, FCCom-TADNet (Concat) achieves the highest accuracy. For the method of combining high-level features at fully-connected layers (FCCon-TADNet), we further compare the efficiency of two combination approaches, including the concatenation (FCCom-TADNet (Concat)) and addition (FCCom-TADNet (Add)) of high-level features from two network streams. Our experiments show that the concatenation approach is much more effective as the accuracy of FCCom-TADNet (Concat) is 1.15% higher than that of FCCom-TADNet (Add). It is assumed that concatenating high-level features is better at preserving the useful discriminative features for classification. It is worth mentioning that although FCCom-TADNet (Add) is not as effective as FCCom-TADNet (Concat), it is still more performant than DLCom-TADNet. In comparison with performances of models trained on a single feature set (i.e. spectrogram or MFCC), the accuracies of InCom-TADNet and FCCom-TADNet (Concat) are much higher, by around 2% to 3%, showing that utilizing spectrogram and MFCC features together brings about better performance for A-TAD's classifiers.

F. THE ROBUSTNESS EVALUATION

This experiment evaluates the robustness of the proposed networks by testing the pre-trained models with testing sets of different noise levels. Recall that the noisy testing sets were created by mixing the recordings in the original testing set with noise recordings at different signal-to-noise ratios (SNRs). There were two sources of noise recordings consisting of traffic noises and weather noises (i.e. heavy rain sounds and strong wind sounds). It is worth mentioning that the frequency range of currently manufactured train horns is between 200 Hz to 4000 Hz, in which popular train horns are in the 200-700 Hz range, and the most high-pitch horns are in the range of above 700 Hz to 2000 Hz. Figure 7 shows the spectrograms of a train horn sound and two samples of the used noises. Since the frequency content of the used noises is in the range of some thousand Hz, which covers the most common train horns' frequency range, mixing those noises with data in the original testing set definitely influence the spectral characteristic of the original recordings.

From the statistic in Table 7, we can see that models based on the combined use of spectrogram and MFCC inputs have better robustness. Across all noise levels ranging from -15dB to +15dB, models trained with both spectrogram and MFCC features yield much higher accuracies compared to those of models trained with only spectrogram or MFCC features. Taking the average accuracy of all testing sets into consideration, it is shown that InCom-TADNet, FCCom-TADNet variants, and DLCom-TADNet are more performant than MSNet variants, by 1% to 4%. Among all models, the FCCom-TADNet (Concat) shows the best robustness since this model obtains the highest average accuracy of 95.11%, which is 2.5% to 4% higher than the results of MSNet variants. At the moderate noisy conditions, i.e. the SNRs of +15dB, +10dB, and +5dB, the performances of all models drop slightly. For the noise level of 0dB, the performances of MSNet variants without attention decrease significantly by approximately 2%, while the accuracies of the remaining models decline slightly.

In more noisy conditions, i.e. SNRs of −5dB, −10dB, and −15dB, all models experience more considerable performance degradation, but models based on feature combination still attain high accuracies, in which the FCCom-TADNet (Concat) achieves the highest accuracies of 95.03%, 93.84, and 91.31% for the test sets of −5dB, −10dB, and −15dB, respectively. At the highest noise level (-15dB), the accuracies of InCom-TADNet, FCCom-TADNet variants,



(a). Train horn (b). Traffic noise (c). Heavy rain noise

FIGURE 7. Spectrograms of a train horn sound and two samples of the used noises.

TABLE 7. Performances of proposed models across various levels of noise.

Models	Features	Accuracy (%) for each SNR								
		-15dB	-10dB	-5dB	0dB	+5dB	+10dB	+15dB	Original data	Average
InCom-TADNet	Spectrogram+MFCC	90.98	93.42	94.30	94.37	94.46	95.47	95.91	96.39	94.41
FCCom- TADNet (Concat)	Spectrogram+MFCC	91.31	93.84	95.03	95.82	95.93	96.17	96.22	96.55	95.11
FCCom-TADNet (Add)	Spectrogram+MFCC	90.85	92.83	93.84	94.26	94.44	94.70	94.83	95.40	93.89
DLCom-TADNet	Spectrogram+MFCC	88.87	92.74	93.86	94.36	94.52	94.59	94.90	95.32	93.65
MSNet without attention	Spectrogram	86.01	88.95	89.22	91.57	93.07	93.15	93.45	93.58	91.13
MSNet with attention	Spectrogram	87.83	90.23	92.49	93.75	94.23	94.36	94.58	94.74	92.78
MSNet without attention	MFCC	85.46	89.30	90.49	91.79	92.78	93.04	93.48	93.71	91.26
MSNet with attention	MFCC	86.20	90.23	93.02	93.92	94.12	94.28	94.61	94.88	92.66

TABLE 8. Performances of several baseline models for various levels of noise.

Models	Features	Accuracy (%) for each SNR							
		-15dB	-10dB	-5dB	0dB	+5dB	+10dB	+15dB	Original data
2D-CNN (AlexNet [35])	Spectrogram	73.60	78.61	84.62	87.79	90.03	91.11	91.58	92.15
2D-CNN (AlexNet [35])	MFCCs	77.47	81.39	85.90	89.15	91.18	91.60	92.01	92.30
2D-CNN (VGG16 [36])	Spectrogram	82.73	89.26	90.32	92.48	93.07	93.11	93.27	93.84
2D-CNN (VGG16 [36])	MFCCs	85.70	89.79	91.26	92.21	92.80	92.98	93.64	93.88
2D-CNN ([12])	Spectrogram	84.47	85.83	88.40	90.76	91.50	91.61	91.75	91.83
2D-CNN ([12])	MFCCs	85.65	86.38	89.17	90.13	91.61	91.64	91.77	92.21
RNN ([20])	Spectrogram	61.81	64.98	71.23	76.95	80.33	81.46	82.24	82.36
RNN ([20])	MFCCs	68.08	71.95	76.22	80.07	83.08	84.87	85.74	85.90
1D-CNN (SoundNet 5 layers [37])	Raw data	68.24	69.20	75.01	83.28	88.56	89.31	90.21	90.21
1D-CNN (SoundNet 8 layers [37])	Raw data	69.38	70.70	75.80	82.73	87.84	89.75	90.45	90.58
1D-CNN and 2d-CNN (EnvNet [23])	Raw data	73.53	77.84	81.54	85.08	88.60	88.93	90.06	90.84

and DLCom-TADNet remain above 90%, while the figures for models trained on a single feature set are smaller than 88%. Note that although MSNet variants with attention are less performant than InCom-TADNet, FCCom-TADNet variants, and DLCom-TADNet, their accuracies across all noise levels are higher than those of MSNet without attention for either spectrogram or MFCC features. This observation further proves the efficiency of attention blocks in the MSNet.

In addition, to compare the robustness of the proposed models with the baseline networks, we evaluated the performances of baseline models on the same testing sets and summarized the results in Table 8. We can see that the baseline models have lower levels of robustness compared to our proposed models. Especially, at the negative SNRs, the performances of baseline models degrade dramatically, resulting in much lower accuracies than those of proposed models.

V. CONCLUSION & FEATURE WORK

This work studied approaches for acoustic-based train arrival detection (A-TAD) which was formulated as a binary audio classification problem, in which two audio classes were audible train warning sounds and noises. A self-collected dataset of train horn sounds, railway alarm sounds, and traffic noises was prepared and combined with published

datasets to create the A-TAD dataset for system development and evaluation. We first investigated the performance of the proposed 2D-CNN model (i.e. the MSNet) whose input can be either the spectrogram or MFCC features, and we found that the MSNet outperformed almost all baseline models. We then examined the effects of frame-level attention on the performance of the MSNet, in which a comparative analysis of applying attention blocks to different network layers was conducted. The experimental results showed that, for both spectrogram and MFCC inputs, the MSNet with attention at every examined 2D-Conv layer was more performant than the MSNet without attention. Especially, higher accuracies were achieved when multiple attention blocks are used together in the network.

Next, we investigated three different approaches to utilize spectrogram and MFCC features together as the input of the A-TAD system's classifier, including input-level feature combination, high-level feature combination, and decision-level fusion with the proposed InCom-TADNet, FCCom-TADNet, and DLCom-TADNet, respectively. Those three models achieved significant improvements in classification accuracies compared to the results of models with a single feature set. In comparison with models only trained with spectrogram or MFCCs, InCom-TADNet, FCCom-TADNet, and DLCom-TADNet also have higher resistance to noise as they produced better performances across various levels of environmental noise, especially at considerable noisy conditions like −15dB and -10dB. Among three models, the FCCom-TADNet obtained the highest accuracies and showed the best robustness. Some existing models were also evaluated with the same experimental conditions, showing that our proposed models achieved much better robustness. All in all, incorporating 2D-CNN, attention mechanism, and the combined use of spectrogram and MFCCs brought about promising results in A-TAD, in which the proposed method with low computational complexity outperformed baseline methods.

Although promising results have been achieved, further examination and development are still needed to maximize the accuracy and robustness of the A-TAD system, so its applicability can be extended. In future work, we would collect a more comprehensive dataset that covers more complex experimental conditions, such as different traffic scenarios and weather conditions, so the A-TAD system built on such an extensive dataset can reach a high level of generalization. We would also examine the problem of determining the direction and distance from the audible train warning sounds.

REFERENCES

[1] J. Santos, M. Hempel, and H. Sharif, "Sensing techniques and detection methods for train approach detection," in *Proc. IEEE 78th Veh. Technol. Conf. (VTC Fall)*, Sep. 2013, pp. 1–5.

[2] K. Chetty, Q. Chen, and K. Woodbridge, "Train monitoring using GSM-R based passive radar," in *Proc. IEEE Radar Conf. (RadarConf)*, May 2016, pp. 1–4.

[3] K. Sato, S. Ishida, J. Kajimura, M. Uchino, S. Tagashira, and A. Fukuda, "Initial evaluation of acoustic train detection system," in *Proc. ITS Asia–Pacific Forum Fukuoka*, Dec. 2018, pp. 1–12.

[4] F. Peng, N. Duan, Y.-J. Rao, and J. Li, "Real-time position and speed monitoring of trains using phase-sensitive OTDR," *IEEE Photon. Technol. Lett.*, vol. 26, no. 20, pp. 2055–2057, Oct. 15, 2014.

[5] B. Yang and Y. Lei, "Vehicle detection and classification for low-speed congested traffic with anisotropic magnetoresistive sensor," *IEEE Sensors J.*, vol. 15, no. 2, pp. 1132–1138, Feb. 2015.

[6] L. Angrisani, D. Grillo, R. S. Lo Moriello, and G. Filo, "Automatic detection of train arrival through an accelerometer," in *Proc. IEEE Instrum. Meas. Technol. Conf.*, May 2010, pp. 898–902.

[7] S. U. Hassan, M. Zeeshan Khan, M. U. Ghani Khan, and S. Saleem, "Robust sound classification for surveillance using time frequency audio features," in *Proc. Int. Conf. Commun. Technol. (ComTech)*, Mar. 2019, pp. 13–18.

[8] D. Henze, K. Gorishti, B. Bruegge, and J.-P. Simen, "AudioForesight: A process model for audio predictive maintenance in industrial environments," in *Proc. 18th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2019, pp. 352–357.

[9] M. A. Sehili *et al.*, "Sound environment analysis in smart home," in *Proc. Ambient Intell.* Berlin, Germany: Springer, 2012, pp. 208–223.

[10] V.-T. Tran and W.-H. Tsai, "Acoustic-based emergency vehicle detection using convolutional neural networks," *IEEE Access*, vol. 8, pp. 75702–75713, 2020.

[11] V.-T. Tran and W.-H. Tsai, "Audio-vision emergency vehicle detection," *IEEE Sensors J.*, vol. 21, no. 24, pp. 27905–27917, Dec. 2021.

[12] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Process. Lett.*, vol. 24, no. 3, pp. 279–283, Mar. 2017.

[13] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *Proc. IEEE 25th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2015, pp. 1–6.

[14] V. Boddapati, A. Petef, J. Rasmusson, and L. Lundberg, "Classifying environmental sounds using image recognition networks," *Proc. Comput. Sci.*, vol. 112, pp. 2048–2056, Jan. 2017.

[15] O. Abdel-Hamid, A. R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," in *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 10, pp. 1533–1545, Oct. 2014.

[16] P. Dhanalakshmi, S. Palanivel, and V. Ramalingam, "Classification of audio signals using AANN and GMM," *Appl. Soft Comput.*, vol. 11, no. 1, pp. 716–723, Jan. 2011.

[17] S. P. Mohanapriya, E. P. Sumesh, and R. Karthika, "Environmental sound recognition using Gaussian mixture model and neural network classifier," in *Proc. Int. Conf. Green Comput. Commun. Electr. Eng. (ICGCCEE)*, Mar. 2014, pp. 1–5.

[18] A. Temko and C. Nadeu, "Classification of acoustic events using SVM-based clustering schemes," *Pattern Recognit.*, vol. 39, no. 4, pp. 682–694, Apr. 2006.

[19] N. Jakovljević and T. Lončar-Turukalo, "Hidden Markov model based respiratory sound classification," in *Proc. Precis. Med. Powered pHealth Connected Health*. Singapore: Springer, 2018, pp. 39–43.

[20] I. Lezhenin, N. Bogach, and E. Pyshkin, "Urban sound classification using long short-term memory neural network," in *Proc. Federated Conf. Comput. Sci. Inf. Syst. (FedCSIS)*, Sep. 2019, pp. 57–60.

[21] W. Dai, C. Dai, S. Qu, J. Li, and S. Das, "Very deep convolutional neural networks for raw waveforms," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 421–425.

[22] S. Abdoli, P. Cardinal, and A. L. Koerich, "End-to-end environmental sound classification using a 1D convolutional neural network," *Expert Syst. Appl.*, vol. 136, pp. 252–263, Dec. 2019.

[23] Y. Tokozume and T. Harada, "Learning environmental sounds with end-to-end convolutional neural network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2721–2725.

[24] V.-T. Tran and W.-H. Tsai, "Stethoscope-sensed speech and breath-sounds for person identification with sparse training data," *IEEE Sensors J.*, vol. 20, no. 2, pp. 848–859, Jan. 2020.

[25] Z. Mushtaq and S.-F. Su, "Environmental sound classification using a regularized deep convolutional neural network with data augmentation," *Appl. Acoust.*, vol. 167, pp. 1–13, Oct. 2020.
[26] J. Guo, N. Xu, L.-J. Li, and A. Alwan, "Attention based CLDNNs for short-duration acoustic scene classification," in *Proc. Interspeech*, Aug. 2017, pp. 469–473.
[27] X. Li, V. Chebiyyam, and K. Kirchhoff, "Multi-stream network with temporal attention for environmental sound classification," in *Proc. Interspeech*, Sep. 2019, pp. 3604–3608.
[28] Z. Zhang, S. Xu, S. Zhang, T. Qiao, and S. Cao, "Attention based convolutional recurrent neural network for environmental sound classification," *Neurocomputing*, vol. 453, pp. 896–903, Sep. 2021.
[29] Y. Su, K. Zhang, J. Wang, and K. Madani, "Environment sound classification using a two-stream CNN based on decision-level fusion," *Sensors*, vol. 19, no. 7, p. 1733, 2019.
[30] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 1041–1044.
[31] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 1015–1018.
[32] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "Librosa: Audio and music signal analysis in Python," in *Proc. 14th Python Sci. Conf.*, 2015, pp. 18–25.
[33] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.
[34] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2004.
[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, vol. 25, pp. 1097–1105.
[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.
[37] Y. Aytar, C. Vondrick, and A. Torralba, "SoundNet: Learning sound representations from unlabeled video," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 892–900.

VAN-THUAN TRAN received the M.S. degree in electrical engineering and computer science and the Ph.D. degree in electronic engineering from the National Taipei University of Technology (NTUT), Taipei, Taiwan, in 2018 and 2022, respectively. From 2015 to 2016, he worked as an Engineer with JGCS Consortium, NSRP Project, Vietnam. He is currently a Postdoctoral Researcher with the Department of Electronic Engineering, NTUT. His research interests include multimedia signal processing and applied artificial intelligence.

WEI-HO TSAI (Member, IEEE) received the Ph.D. degree in communication engineering from the National Chiao Tung University, Hsinchu, Taiwan, in 2001. From 2001 to 2003, he was with Philips Research East Asia, Taipei, Taiwan, where he worked on speech processing problems in embedded systems. From 2003 to 2005, he served as a Postdoctoral Fellow for the Institute of Information Science, Academia Sinica, Taipei. He is currently a Professor with the Department of Electronic Engineering, National Taipei University of Technology, Taiwan. His research interests include spoken language processing and music information retrieval.