Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Real time end-to-end glass break detection system using LSTM deep recurrent neural network

Wai Yan Nyein Naing*, Zaw Zaw Htike, Amir Akramin Shafie

Mechatronic Engineering Department, International Islamic University Malaysia (IIUM), Gombak, Malaysia

ARTICLE INFO

Article history:
Received 15 May 2018
Received in revised form
12 January 2019
Accepted 18 January 2019

Keywords:
Glass break detection system
Deep learning
Long-short term memory
Deep recurrent neural network

ABSTRACT

The aim of this paper is to propose a new design for a glass break detection system using LSTM deep recurrent neural networks at an end-to-end approach to reduce false positive alarm of state of the art glass break detectors. We utilized raw wave audio data to detect a glass break detection event in End-to-End learning approach. The key benefit of End-to-End learning is avoiding the need for hand-crafted audio features. To address the issue of a vanishing gradient and exploding gradient problem in conventional recurrent neural networks, this paper proposed deep long short term memory (LSTM) recurrent neural network to handle the sequence of the input audio data. As a real-time detection result, the proposed glass break detection approach has a clear advantage over the conventional glass break detection system, as it yields significantly higher precision accuracy (99.999988 %) and suffers less from environmental noise that might cause a false alarm.

1. Introduction

Glasses are increasingly used in the construction of offices and residential places because of the advantages of comfort, well-being, style and light sustainability. Despite its benefits, it is also prone to security risks at night or when a person is not present, as it would be easy for an intruder to smash the glass based door, then reach inside and open the latch lock. Therefore, glass break detection plays an important role in ensuring the security of offices and residential places, as most of the burglars or intruders enter the home through glass doors and windows. A glass break detector is an electronic sensor that detects breakage vibrations or shattering sounds of glass panes. A glass break detector can be used for the protection of the internal and external perimeter building. When the glass pane is shatters or breaks, it generates sound over a wide band of vibrations and frequencies. These shattering glass sounds have a kind of distant frequency. Generally, these can range from 3 to 5 kHz, depending on the type of glass and the presence of an interconnected plastic layer. Most conventional electronic glass break detectors process use pre-determined frequency, amplitude and vibration thresholds to determine whether the glass has broken. Generally, conventional glass break detectors can be grouped into three main categories, (such as Activate Detectors, Physical Vibration Detectors and Acoustic Detectors). Active detectors send a set of frequency energies towards the window glass panes and receive the reflected frequency energy. Any change observed in the reflected frequency energy triggers an alarm or activates another circuit (Clark and Lewis, 1996; Zidan, 2015). The Physical Vibration detector is composed of a piezoelectric element. Whenever the glass is broken, there is some vibration caused in the molecules. These vibrations are noted by the detectors and converted in to an electric signal, then the alarm system is triggered (Sharapov, 2011). Acoustic based glass break detectors contain one or more acoustic audio transducers that can detect an electrical signal in response to a high amplitude and frequency sound created due to breaking of the glass plate. If a burglar is trying to break through a window, the detector would pick up on the high-pitched shattering sound and a pre-determined frequency composition of breaking event to trip the alarm (Cecic and Fong, 1997; Matesa, 2015; Rickman, 1995). In fact, classification of glass breaking sounds and some loud anomalous audios (such as, gunshots, thunder, and people shouting, dropping and hitting objects) remains a challenging task despite decades. The

* Corresponding Author.
Email Address: waiyan.nn18@email.com (W. Y. N. Naing)

https://doi.org/10.21833/ijaas.2019.03.009

Corresponding author's ORCID profile:
https://orcid.org/0000-0001-5639-9899

chances of false alarms in glass break detectors are high, because shock and anomalous loud sounds have similar frequency and vibration thresholds of pre-defined glass breaking sounds (Clavel et al., 2005). The recent development in technology has improved towards overcoming this drawback. There is an ongoing success in the performance of Artificial Intelligence (AI) in dealing with video and audio surveillance applications (such as speech recognition, computer vision, voice translation, and much more in past few years), and smart security surveillance systems are shifting from conventional electronic sensor based classification techniques to modern machine learning and deep learning methods. Among them, Conte et al. (2012) proposed abnormal audio event detection in an urban area, Mahler et al. (2017) discussed a home interior security system, Dufaux et al. (2000) proposed an impulsive sound detection system in a public square, and Zidan (2015) studied protection of nuclear facilities using hardware sensors. Gestner et al. (2007) proposed Digital Signal Processing (DSP) based glass break detectors in homes and offices. Peng et al. (2014) focused on impulsive sound detection and surveillance system in public transport. Aurino et al. (2014) discussed anomaly detection in automatic surveillance application. Kiktova et al. (2015) proposed a gunshot and shout sounds detection system in a city environment which can be noted as particular applicable for surveillance responsibilities, wherein audio can continually make contributions. In this paper, we advocate for a new architecture of glass break detection system to reduce false detection alarm using long short term memory (LSTM) deep recurrent neural network in an end-to-end approach.

2. Outline

The rest of the paper is organized as follows. Section 3 will discuss data acquisition of acoustic audio signals. Section 4 will describe the LSTM deep recurrent neural network. Section 5 will explain end-to-end (LSTM) deep recurrent neural network approach on glass break detection events. Section 6 will summarize the experimental result of proposed deep learning model, and Section 7 will provide conclusions.

3. Data acquisition

For data acquisition, we manually collected annotated dataset of glass break and non-glass break acoustic audio for training, testing, and validation of the proposed system. Input audio signals are recorded with an acoustic sensing built-in microphone from a laptop. Collected audio signals for glass break detection is generally at a 44100 sampling rate per second at 2 sec time frames.

We collected 5000 audio (.wav) slices samples for audio dataset under various noise level environments, as shown in Fig. 1. This dataset is composed of two types of sound classes consisting of 2500 slices samples of breaking glass sounds data (breaking glass sounds with different noise level) and 2500 non-breaking glass sounds slices from environmental sound and noises (combining of shouted sounds, cars horn, household, alarm , animals, farm and child playing, people conversation sounds), as shown in Fig. 2.

Fig. 1: Data acquisition of glass breaking sounds

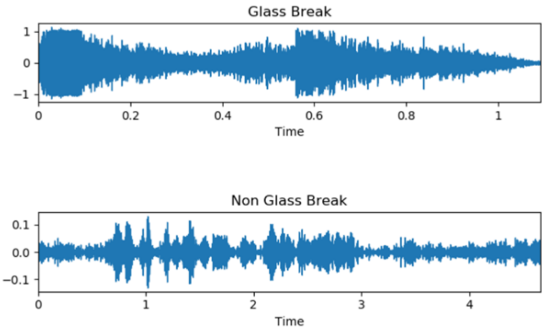

Fig. 2: Wave form of glass break and non-glass break sounds

4. Methodology

4.1. Deep recurrent neural network (long short term memory - LSTM)

Primary topics of research on deep learning were image and audio analysis. Although these are many sources of image data that can be collected in recent years, audio analysis has been restrained until lately. Few publicly collected sound data and complex sequence characteristics of audio data (such as frequency features, energy levels) are the causes for the limited research in this field (Graves et al., 2013). Recurrent networks are a kind of artificial neural network intended to understand patterns in continuous data (which includes text data, speech data or sequence of numerical sensors data, time series data of stock exchanges and social networks). Recurrent networks (RNNs) vary from conventional feed forward neural networks in that the feedback loop is associated with their previous decisions and takes its own outputs as an input for each timestamp. RNN discovers correlations between events separated by many timestamps, and these correlations can be denote as "long-term dependencies" (vanishing gradients) and "short-term dependencies (exploding gradients)" (Sak et al., 2014).

The first drawback with conventional recurrent neural network is finding the correlation between current events and long-term memory of past timestamp (vanishing gradients problems). Updated weights of RNN are too small (almost unchanged) and many iterations are needed to update the new weights. The second drawback in RNN is finding the correlation between current events and short-term memory of recent timestamp (exploding gradients problems). Updated weights of RNN are too large and the updated weights is too distant from current weights (Gers et al., 2000).

The architecture of an LSTM Network has been shown to be particularly effective when stacked into a deep configuration, towards handling the vanishing gradient and exploding gradient issues of traditional Recurrent Neural Network. In the LSTM structure, the recurrent hidden layer consists of a set of recurrently connected subnets called “memory blocks”. Each memory block includes one or more self-connected memory cells and three multiplicative gates to control the flow of information (Gers et al., 2000).

The processes of carrying memory forward of LSTM graphically is described in Fig. 3. An architecture of (LSTM) RNN is as follows. In the first gate, we decide what we need to forget from the data (forget gate); in the second gate, new information is stored into the cell state throughout the whole process (Input Gate); in the final gate, the new output is produced based what we decided (output gate). This is what basically how LSTM works to handle complex sequences of data at different timestamps.

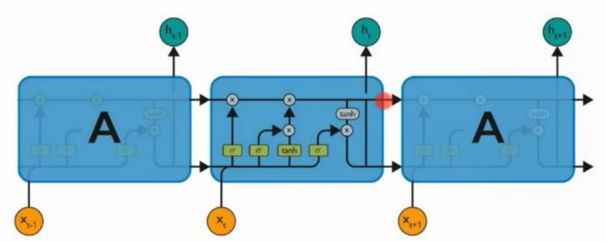

Fig. 3: LSTM (RNN) at different timestamps (Edureka, 2017)

4.2. Forget gate

The first stamp in LSTM is to identify information that is not required and will be discarded from the cell state. This decision is made by a sigmoid layer called a Forget Gate Layer.

Graphically representation of the Forget Gate is shown in Fig. 4.

Based on Gers et al. (2000) Eq.

$$f_t = \partial(W_f\,[h_{t-1}, x_t] + b_f) \qquad (1)$$

the Forget gate is denoted as f_t and cell state as c_t. The hidden state at a previous timestamp is h_{t-1}, and the current input is denoted as x_t. The previous hidden state h_{t-1} are cascade together at a same timestamp, modified by a Weight matrix W_f and summed with a bias value of the forget gate. The result of the function is squashed by the sigmoid activation function ∂ which is a standard tool for considering very large or very small values of the Forget gate, as well as rendering gradients workable for back propagation through time. The final value of forget gate (f_t) can be between 0 and 1 according to the output of sigmoid activation function. If the value of f_t is 0, then the value of the event is necessary to forget; 1 means complete info of previous timestamp is needed to remember for current state (Pascanu et al., 2013).

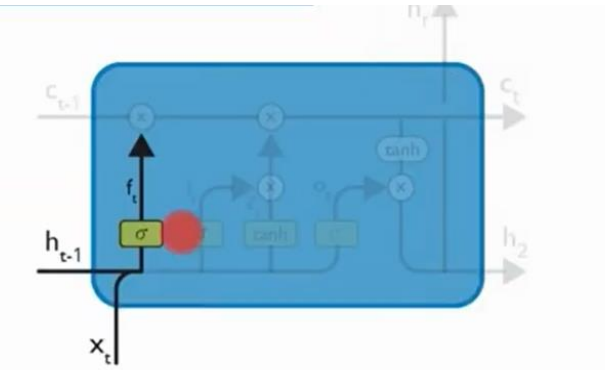

Fig. 4: Forget gate of LSTM (RNN) block (Edureka, 2017)

4.3. Input gate

This step is to decide what new information that we are going to store in the cell state.

Graphical representation of the Input Gate is shown in Fig. 5.

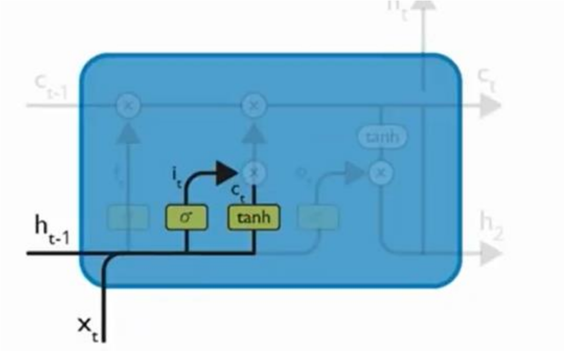

Fig. 5: Input gate of LSTM (RNN)

Using Eq. 2 from Gers et al. (2000)

$$i_t = \partial(W_i[h_{t-1}, x_t] + b_t) \qquad (2)$$

i_t denotes a sigmoid layer called “input gate layer” that decides which value will be updated. And from Eq. 3 by Gers et al. (2000)

$$c_t^{\sim} = tanh(W_c[h_{t-1}, x_t] + b_c) \qquad (3)$$

$c_t^{\sim}$ denotes a tanh layer that creates a vector of “new candidate values” that could be added to the state. Then, we’ll combine these two gates to update the state.

From Eq. 4 (Gers et al., 2000), We then update the old cell state (c_{t-1}) into the new cell state(c_t).

$$c_t = (f_t * (c_{t-1})) + (i_t * \tilde{c_t}) \qquad (4)$$

First, we multiply the old state (c_{t-1}) by f_t, forgetting the things we decide to forget earlier. Then, we add $i_t * \tilde{c_t}$. This is the new candidate value c_t, scaled by how much we decided to update each state value.

4.4. Output gate

The Output Gate decides what part of the cell state to output.

Graphical representation of the Output Gate is shown in Fig. 6.

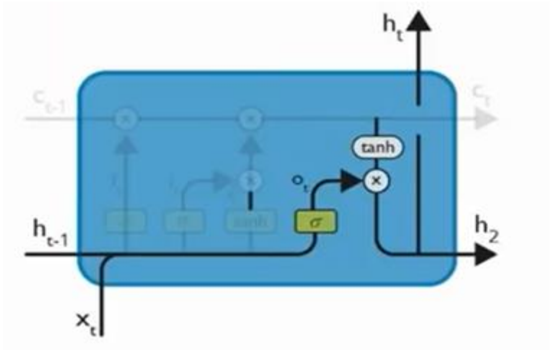

Fig. 6: Output gate of RNN (LSTM)

From Eq. 5 (Gers et al., 2000), O_t denotes an output Gate that will run as a sigmoid layer that decides what part of the cell state going to output.

From Eq. 6 (Gers et al., 2000) , the updated cell state (c_t) will pass through tanh activation function to get push values (between -1 and 1), so that LSTM (RNN) only produces the new output information (h_t) related to the goal of coming next (Li and Wu, 2015).

$$O_t = \partial(w_0 * [h_{t-1}, x_t] + b_0) \qquad (5)$$
$$h_t = O_t * \tanh(c_t) \qquad (6)$$

5. Experiments

In the experiment, 2 second variable-length size audio sequences with 50% overlap time frame are recorded (glass breaking and non-glass breaking sound samples) and the dataset transformed into (299 × 299 × 3) shape fixed-length raw temporal image form and further reshaped from a fixed-length input image into a bottleneck tensor size (2048 dimensional) byte vector array form. If sufficient training data are available, treating raw temporal acoustic audio wave directly to the entire neural network works well for target classification, as opposed to hand crafted heuristics spectral audio features. Before training with LSTM RNN, 5000 samples of audio in the dataset (2500 glass break, 2500 non glass break) are randomly split into 10-fold cross validation form with training (70%), validation (10%), and test (10%) sets. Training set and validation set data are used in training with three time-delay hidden layers LSTM recurrent neural net, which computes the sigmoid and tanh activation functions of a weighted sum for each timestamp. For network training, we tried to set the specific initial and final learning rates in a range from 0.0005 to 0.001 for stable convergence. To prevent over fitting during training, we used the early termination method during training and L2 Regularization dropout. Then, the truncated back-propagation through time (BPTT) learning algorithm is adopted to reduce the cost function and optimization process. The gradients are computed for each subsequence and back-propagated to its start. After the model finally updates the parameters in the LSTM networks, the output gate of the LSTM decides the N sequence of audio as a glass break or not. All experiments were conducted with tensor flow library in the Python environment.

The system architecture of proposed end-to-end glass break detection system using deep LSTM (Recurrent Neural Network) and conventional hand-crafted feature based glass break detection system are shown in Fig. 7 and Fig. 8.

6. Result and discussions

To measure detection accuracy of the proposed LSTM model in offline during training, we split the data into 10-fold cross validation form with training/validation/test sets. The experiment on 10-fold cross-validation without replacement can prevent the use of sub-segments from the same recordings in training and validation. Cross entropy and the mean square error rate are used as an accuracy measure of the proposed classification criteria. Experimental results of proposed glass break detection system show that we obtained a trained set accuracy of 100%, validation set accuracy of 100% and invisible test accuracy of 99.999% correct detection result for 5000 samples of audio dataset. In the online experiments, a microphone is used to record at every 2 sec time frame of audio (.wav) with sampling rate (44100 kHz). Recorded audio is analyze with proposed LSTM (Deep Recurrent Neural Network) end-to-end learning approach to detect glass breaking sound using laptop built-in microphone. To measure the accuracy of the online system, we ran our proposed model on the raspberry-pi device and test with non-stop 48 hours detection with different noise (such as, human speaking, clap sounds, Door opening/Closed, Bell sounds, horns).

During the two days (48 hours) of testing, only two false glass break alarm detection alarm is occurred (e.g., sensitive to cough sounds). That means that online proposed glass break detection model correctly detects the glass breaking sound at a 99.999988% detection accuracy. To solve the false positive alarm of new environmental noise (such as cough sounds), we recorded and added this false alarm sound to the proposed training system and retrained to perform the detection model better.

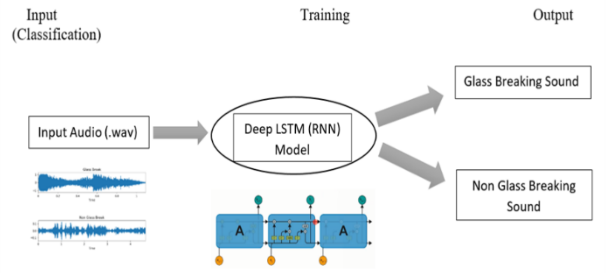

Fig. 7: End to end glass break detection system using deep LSTM RNN

Table 1 describes the experimental results of the state of the art and the methodological comparison of hand crafted features and sensor based glass break detection system. According to the experimental results from Table 1, the proposed End-to-End glass break detection system can perform good detection with the least false alarm errors as compared to other conventional electronic glass break detectors and hand crafted feature based Machine Learning methods.

7. Conclusion

The major drawback of conventional glass break detectors is false alarms. Sounds such as thunder, shouting, gunshot, hitting objects are similar in frequency and threshold value to glass breaking sounds events that may cause false positives in the alarm system. Therefore, this research proposed a new architecture for glass break detection approach based on LSTM deep recurrent neural network, to improve the correct detection accuracy with less false alarms. In this approach, we utilized raw wave audio data to detect a glass break detection event in End-to-End learning approach.

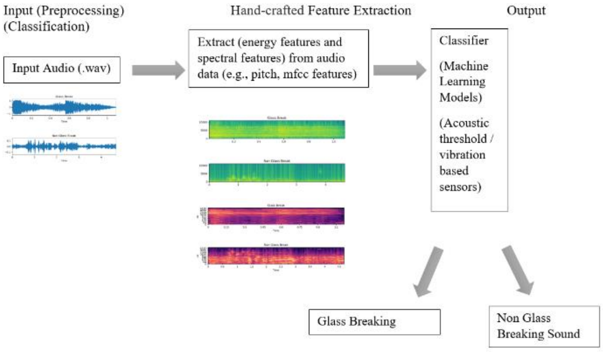
Input (Preprocessing)
(Classification)
Hand-crafted Feature Extraction
Output
Input Audio (.wav)
Extract (energy features and spectral features) from audio data (e.g., pitch, mfcc features)
Classifier
(Machine Learning Models)
(Acoustic threshold / vibration based sensors)
Glass Breaking
Non Glass Breaking Sound

Fig. 8: Conventional glass break detection system using hand crafted audio features

The key benefit of End-to-End learning is avoiding the need of hand-crafted audio features. To address the issue of a vanishing gradient and exploding gradient problem in conventional recurrent neural networks, this paper proposed deep long short term memory (LSTM) recurrent neural network to handle the sequence of the input audio data. As a real time detection result, the proposed glass break detection approach has a clear advantage over the conventional glass break detection system, as it yields significantly higher precision accuracy (99.999988 %) and suffers less from environmental noise that might cause a false alarm. With the availability of sufficient computational power of embedded applications and data, deep learning has become practical and ever more present in powerful and intelligence applications to security surveillance.

Table 1: Experimental results of different strategies of constructing glass break detection system

Authors	Sensors / Hand crafted Features	Classification	Accuracy %
Conte et al. (2012)	centroid, total, energy ERSB features, Zero crossing rate; spectral centroid, spectral pitch	LVQ Neural Network	54.93
Peng et al. (2014)	Spectral Features (Audio Spectrum Flatness(ASF))	HMM (Hidden Markov Model)	80
Aurino et al. (2014)	Mel-Frequency Cepstral Coefficient (MFCC)	Support Vector Machine (SVM)	91.7
Zidan (2015)	Vibration Threshold data	Vibration based Glass Break Detector(GB)	65
Kiktova et al. (2015)	Mel-Frequency Cepstral Coefficient (MFCC)	SVM-1 with Weighted Majority Voting (WMV) strategy	69.93
Mahler et al. (2017)	Threshold Based Features from Sensors (Vibration threshold, Accelerometer data, Magnetometer data, Air pressure data)	Feed forward Neural Network (K-Nearest Neighbors + Dynamic Time Warping)	92
Proposed Glass Break Detection System	End-to-End System based on raw temporal audio data	Long Short Term Memory (LSTM) (Deep Recurrent Neural Network)	99.999988

Acknowledgement

This work was supported by the Ministry of Higher Education Malaysia under PRGS17-002-0042 and International Islamic University Malaysia under RIGS16-350-0514.

Compliance with ethical standards

Conflict of interest

The authors declare that they have no conflict of interest.

References

Aurino F, Folla M, Gargiulo F, Moscato V, Picariello A, and Sansone C (2014). One-class SVM based approach for detecting anomalous audio events. In the International Conference on Intelligent Networking and Collaborative Systems, IEEE, Salerno, Italy: 145-151. https://doi.org/10.1109/INCoS.2014.59

Cecic D and Fong HUS (1997). Glass break detector (U.S. Patent No. 5,675,320A). Patent and Trademark Office, Washington, DC, USA.

Clark FB and Lewis KT (1996). Glass break detector and a method therefor (U.S. Patent No. 5,543,783A). Patent and Trademark Office, Washington, DC, USA.

Clavel C, Ehrette T, and Richard G (2005). Events detection for an audio-based surveillance system. In the IEEE International Conference on Multimedia and Expo, IEEE, Amsterdam, Netherlands: 1306-1309. https://doi.org/10.1109/ICME.2005.1521669

Conte D, Foggia P, Percannella G, Saggese A, and Vento M (2012). An ensemble of rejecting classifiers for anomaly detection of audio events. In the IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance, IEEE, Beijing, China: 76-81. https://doi.org/10.1109/AVSS.2012.9

Dufaux A, Besacier L, Ansorge M, and Pellandini F (2000). Automatic sound detection and recognition for noisy environment. In the 10th European Signal Processing Conference, IEEE, Tampere, Finland: 1-4.

Gers F, Schmidhuber JA, and Cummins F (2000). Learning to forget: Continual prediction with lstm. Neural Computation, 12(10): 2451–2471. https://doi.org/10.1162/089976600300015015 **PMid:11032042**

Gestner B, Tanner J, and Anderson D (2007). Glass break detector analog front-end using novel classifier circuit. In the IEEE International Symposium on Circuits and Systems, IEEE, New Orleans, USA: 3586-3589. https://doi.org/10.1109/ISCAS.2007.378528

Graves A, Mohamed AR, and Hinton G (2013). Speech recognition with deep recurrent neural networks. In the IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE: 6645-6649. https://doi.org/10.1109/ICASSP.2013.6638947

Kiktova E, Lojka M, Pleva M, Juhar J, and Cizmar A (2015). Gun type recognition from gunshot audio recordings. In the 3rd International Workshop on Biometrics and Forensics (IWBF 2015), IEEE, Gjovik, Norway: 1-6. https://doi.org/10.1109/IWBF.2015.7110240

Li X and Wu X (2015). Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition. In the IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, Brisbane, Australia: 4520-4524. https://doi.org/10.1109/ICASSP.2015.7178826

Mahler MA, Li Q, and Li A (2017). Secure house: A home security system based on smartphone sensors. In the 2017 IEEE International Conference on Pervasive Computing and Communications (PerCom), IEEE, Kona, HI, USA: 11-20. https://doi.org/10.1109/PERCOM.2017.7917846

Matesa JM (2015). Alarm detection device and method (U.S. Patent No. 9,191,762B1). Patent and Trademark Office, Washington, DC, USA.

Pascanu R, Mikolov T, and Bengio Y (2013). On the difficulty of training recurrent neural networks. In the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 28: 1310-1318. **PMCid:PMC4517175**

Peng L, Yang D, and Chen X (2014). Multi frame size feature extraction for acoustic event detection. In the Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific, IEEE, Siem Reap, Cambodia: 1-4. https://doi.org/10.1109/APSIPA.2014.7041574

Rickman SA (1995). Direction-sensing acoustic glass break detecting system (U.S. Patent No. 5,471,195A). Patent and Trademark Office, Washington, DC, USA.

Sak H, Senior A, and Beaufays F (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In the Fifteenth Annual Conference of the International Speech Communication Association, Singapore: 338-342.

Sharapov V (2011). General information about piezoelectric sensors. In: Sharapov V (Ed.), Piezoceramic Sensors: 1-24. Springer, Berlin, Heidelberg, Germany. https://doi.org/10.1007/978-3-642-15311-2_1

Zidan WI (2015). Estimation of cluster sensors' probability of detection for physical protection systems evaluation. Journal of Physical Security 8(1): 40-54.