Effect of Distance and Direction on Distress Keyword Recognition using Ensembled Bagged Trees with a Ceiling-Mounted Omnidirectional Microphone

Nadhirah Johari, Mazlina Mamat*, Yew Hoe Tung, Aroland Kiring
Faculty of Engineering, Universiti Malaysia Sabah, Kota Kinabalu, Malaysia

***Abstract*—Audio surveillance can provide an effective alternative to video surveillance in situations where the latter is impractical. Nevertheless, it is essential to note that audio recording raises privacy and legal concerns that require unambiguous consent from all parties involved. By utilizing keyword recognition, audio recordings can be filtered, allowing for the creation of a surveillance system that is activated by distress keywords. This paper investigates the performance of the Ensemble Bagged Trees (EBT) classifier in recognizing the distress keyword "Please" captured by a ceiling-mounted omnidirectional microphone in a room measuring 4.064m (length) x 2.54m (width) x 2.794m (height). The study analyzes the impact of different distances (0m, 1m, and 2m) and two directions (facing towards and away from the microphone) on recognition performance. Results indicate that the system is more sensitive and better able to identify targeted signals when they are farther away and facing toward the microphone. The validation process demonstrates excellent accuracy, precision, and recall values exceeding 98%. In testing, the EBT achieved a satisfactory recall rate of 86.7%, indicating moderate sensitivity, and a precision of 97.7%, implying less susceptibility to false alarms, a crucial feature of any reliable surveillance system. Overall, the findings suggest that a single omnidirectional microphone equipped with an EBT classifier is capable of detecting distress keywords in a low-noise enclosed room measuring up to 4.0 meters in length, 4.0 meters in width, and 2.794 meters in height. This study highlights the potential of employing an omnidirectional microphone and EBT classifier as an edge audio surveillance system for indoor environments.**

***Keywords*—Distress speech; ensemble bagged trees; audio surveillance; machine learning; distance; directions**

I. Introduction

Screaming, yelling, or shouting is a natural way to express agony. They are particularly useful for detecting distress events in audio-based surveillance and monitoring applications that use the pitch and intensity of voice. The audio key-event detection system using Gaussian Mixture Model (GMM) presented in [1] uses acoustic references to detect and examine abnormal events based on a binary classification technique of shot and normal classes. In [2], a monitoring service technology was developed to determine the stress level from the voice tone changes. An anomalous audio event detection using GMM in [3] discriminates screams and gunshot sounds from noises. Another study in [4] utilizes the personal computer equipped with a sound card and microphone to detect distress sounds like a cry for help or glass breaking. Real-time sound analysis was developed using eight mic channels to distinguish stress from normal situations [5]. Audio data are advantageous in assistive awareness systems for emergency recognition of falls and distressed speech expression in elder or patient care [6-7]. In [8], a two-stage learning-based method was proposed to detect screams and cry in urban environments. Indoor context based on Receiver Operating Characteristic (ROC) result gives the least false alarm rate compared to the other five contexts; Gathering, conversation, outdoors, machinery, and multimedia. Detection of speech distress was also conducted through remote monitoring [9]. According to the results obtained in [10], it can be concluded that in various interaction scenarios, voice pitch may serve as an accurate biosocial and individual identifier. The distress call is useful in emergencies, especially tracking a person by detecting cries for help in an enclosed environment [11].

However, people might also talk, laugh or scream loudly and high-pitched when they get excited. Surveillance systems triggered by the voice's pitch and intensity will create false alarms in those circumstances. Moreover, the existing surveillance system may also intrude on an individual's privacy and invasion of civil rights [12]. For example, always-listening devices like smartphones may accidentally wiretap the information of the surroundings [13]. Hence, the privacy-aware architecture was proposed to prevent privacy violations and potential recordings [14, 15]. The risks are apparent due to the availability of technology that can access sensitive data such as audio [16]. Overcoming these issues requires a simplified surveillance system that intelligently recognizes distress keywords apart from the voice's pitch and intensity. Such a system will enable a crime detection automation system that recognizes targeted distress speech and non-distress (high-toned) speech. This idea has been explored for outdoor surveillance [8] and should be extended to indoor environments when video surveillance is not viable [17-19]. Places like shared bathrooms and nursery rooms should be equipped with audio surveillance to reassure safety [20-21].

Having indoor audio surveillance that is always recording in the cloud is not well accepted by many due to privacy and security concerns. Edge Artificial Intelligence (AI) could be employed as a solution. Audio surveillance with edge AI will be able to detect specific keywords and trigger the system at the device level (locally). Among the AI algorithms, the Ensemble Bagged Trees (EBT) is one of the supervised machine learnings explored in audio data classification for safety-related applications. EBT has been applied to non-

speech data to investigate false speakers via spoofing sounds [22], classify non-speech audio data [23], monitor babies [24], and classify sounds [25-26]. EBT has also recently been employed to differentiate multiple emotions; anger, disgust, fear, joy, neutral, surprise, and sad from speech audio signals [27-29].

These studies show that EBT performs considerably better than other classifiers, such as Boosted Trees, Bagged Trees, Subspace K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Quadratic SVM, and Quadratic Discriminant [24,29]. EBT can be a good option for real-time distress keyword recognition due to its ability to handle noise and variability in speech data. In speech recognition, there can be significant variation in speech patterns due to individual differences, accents, and other factors. Bagged trees can capture this variability by creating multiple decision trees and combining their outputs. Thus, it can improve accuracy and robustness in recognizing spoken keywords. EBT is computationally efficient and can be easily parallelized, making it a good option for real-time speech recognition applications. It can also be trained using small amounts of data, which is beneficial for scenarios where large amounts of labeled data may not be available.

Employing EBT in the edge audio surveillance system requires careful evaluation to get robust performance. To the best of our knowledge, there is still a shortage of work on distress speech detection using EBT because most studies focus on non-speech and non-distress signals. Moreover, the effectiveness of EBT on distances and directions of the sound sources from the microphone is still undiscovered. This paper investigates the effect of speech distance and direction on the EBT recognition performance, which was proven superior in a previous analysis [30]. For that, two experiments were conducted in a room that resembled a typical nursing room size. Experiment 1 studies the performance of EBT to detect a distress keyword from three different distances: 0m, 1m, and 2m. Experiment 2 analyzes the effect of different directions, facing toward and away from the microphone.

This paper is arranged as follows. Section II covers the materials and research methodology. Section III presents and discusses the results. Finally, Section IV concludes the findings, research's limitations, and future recommendations.

II. Materials and Methods

A. Experimental Setup

The experiment was conducted in a low-noise room with a dimension of 4.064m (Length) x 2.54m (Width) x 2.794m (Height). An omnidirectional microphone was installed at the center of the room's ceiling to capture audio signals from various angles at equal coverage and in a short range. However, omnidirectional microphones collect more noise than directional microphones [31]. Thus, it was suggested to place the omnidirectional away from the reflecting areas [32-33]. Speeches were uttered from horizontal distances of 0 m, 1 m, and 2 m by speakers in sitting condition, facing toward the microphone and away from it, as shown in Fig. 1 to 3. The microphone was connected to a Vivo V17 smartphone (Android 9.0 Pie & Octa-core processor) that performed as a speech recording device.

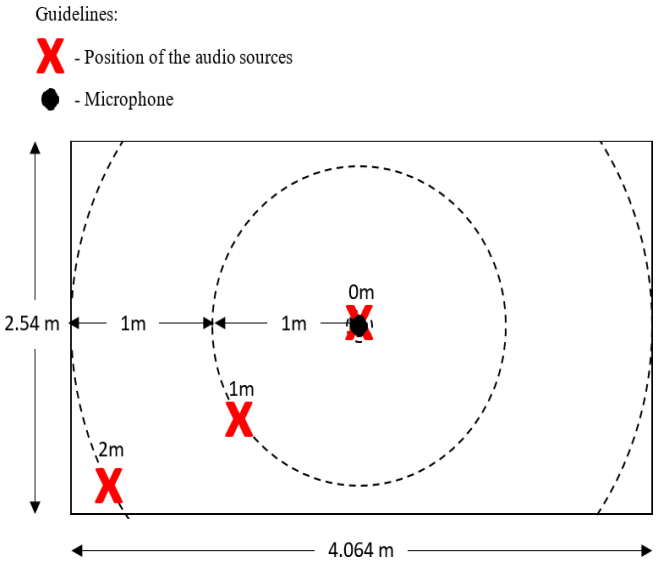

Fig. 1. Distance between speaker and microphone in a closed room.

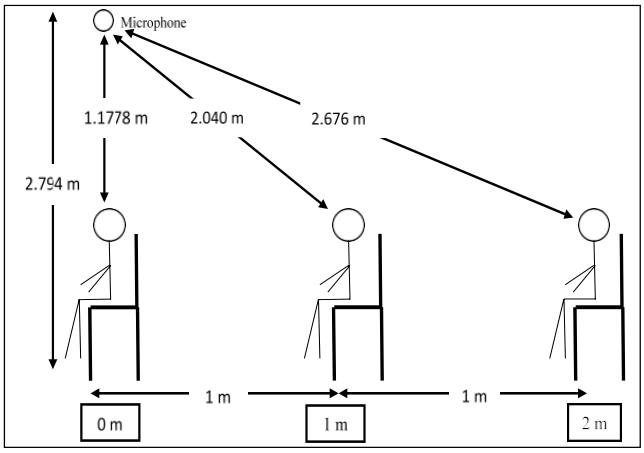

Fig. 2. Position of the speaker during data collection facing toward the microphone.

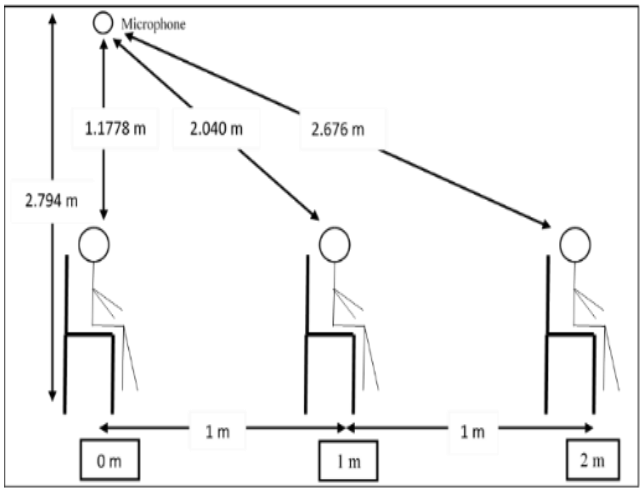

Fig. 3. Position of the speaker during data collection facing away from the microphone.

B. Signal Preprocessing – Audio Filtering

The recorded audio signal was initially stereo with MP4a format, 44100 Hz sampling rate, and 32-bit rate. The audio signal was then exported in WAV format with 44100 Hz for further processing. WAV format audio files are stored in a large space because the signals are uncompressed, maintaining strong sound quality [34]. Due to the differences in time-frequency representation, both left and right channels were calculated to find their average to produce only a mono data channel instead of separating the channels into two parts [35-36].

C. Feature Extraction and Feature Reduction

Each preprocessed audio signal was divided into smaller frames using a Hamming window of 1024 ms and an overlap length of 512 ms [37-39]. For every frame, thirteen Mel-Frequency Cepstral Coefficients (MFCCs) were extracted and organized into an f * c matrix, where f is the number of frames and c is the thirteen MFCCs. The total number of frames was set to 120 by adding or deleting frames at regular intervals, resulting in a matrix of 120 * 13 MFCCs representing the audio signal features [30]. Principal Component Analysis (PCA) was used to reduce the features to 13*13 [40]. The pseudocode for feature extraction is given in Fig. 4.

```
Obtain a speech signal of 1.5s length, and do the following:
    Detect Endpoints, obtain the voiced part
    Divides into f frames (Hamming window: 1024ms, overlap 512ms)
    For n = 1 to f
        Extract 13-MFCCs
        Construct an f x 13 matrix
    End For loop
    If f > 120, do the following,
        Obtain number of rows to be removed, r
        Determine the interval, d
        For i = 1 to f
            Delete the 13-MFCCs at every d
        End for loop
    Else
        Obtain number of rows to be added, r
        Determine the interval, d
        For i = 1 to 120
            For every d, compute new MFCCs using the average
            of MFCCs(d-1) and MFCCs(d+1)
            Add the new MFCCs at row d
        End For loop
    End if-else
```

Fig. 4. Pseudocode for feature extraction.

D. Data Collection

The distress keyword "Please" was chosen for analysis, as it was found to be the most distinct compared to other distress keywords such as "Oi", "Help", "Tolong", and "No [30]. A total of 3600 speeches containing the targeted distress keyword "Please" and non-targeted speeches were collected and divided into four datasets: Dataset 1, Dataset 2, Dataset 3, and Dataset 4. Dataset 1 comprises 100 samples for each distance and 300 for each direction of distressed "Please" speeches, recorded by five female speakers, with each speaker producing 20 samples for each distance and direction.

Dataset 2 includes 600 samples of recorded distressed "Please" speeches played at three different distances and directions. Dataset 3 has the same design as Dataset 1, except the speakers uttered "Please" in a non-distressed tone. Dataset 4 contains 20 samples of each of the five words "One", "Two", "Three", "Okay", and "Yes", spoken in a distressed or high tone captured from three female speakers at each position. Each dataset was labeled '1' for the targeted "Please" and '0' for the non-targeted speeches. The datasets were then divided into training and testing data in an 80 to 20 ratio, as shown in Table I. The EBT was trained using a combined training data of 2880 speeches and tested on various groups of unseen speeches.

TABLE I. DATA COLLECTION OF DISTRESS KEYWORD "PLEASE"

Dataset	Characteristic	Label	Sample	Data Partition	
				Training (80%)	*Testing (20%)*
1	Distress "Please"	1	600	480	120
2	Distress "Please"	1	600	480	120
3	Non-distress "Please"	0	600	480	120
4	Distress/High-tone Words	0	1800	1440	360

E. EBT Classifier

Z 5Breiman presented the ensemble technique in 1996, intending to improve the Decision Trees (DT) classification performance [41]. The word 'Bagging' itself is a Bootstrap Aggregation that reduces the decision tree variance.

The bootstrap method randomly creates minor groups of data with replacements from the overall dataset from the training dataset. Each set created with equal probability will undergo a parallel training of DT classifiers. It produces a robust performance compared to an individual DT model [42]. Each DT model will independently produce different features. Then, the aggregation process was applied by accumulating the predictions of all different DT groups and taking the mean of the outcomes to get the final bagging result. Likewise, this machine learning classifier is a highly precise model combining various decision trees [43]. For this study, 30 DT classifiers were used for the parallel ensemble technique, as illustrated in Fig. 5.

Equation (1) defines its principle where DT learners, $f_d(x)$ are trained based on the architecture in Fig. 5 with the bootstrapped dataset. The mean of the total predictions from every DT learner is taken as the result.

$$f(x) = \frac{1}{D}\sum_{d=1}^{D} f_d(x) \qquad (1)$$

Where; D = sets of bootstrapped data, d = DT learners.

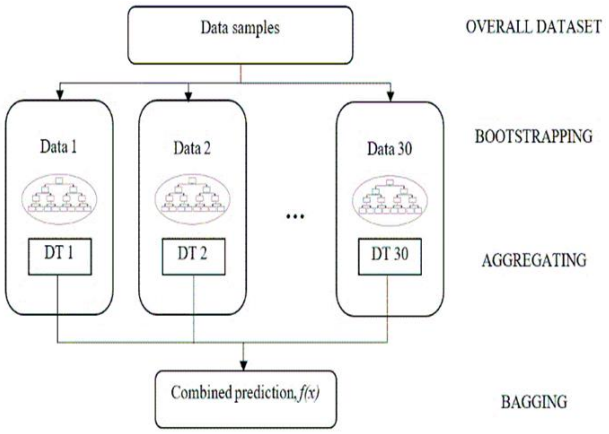

Fig. 5. The ensemble bagged trees.

F. Evaluation Metrics

The efficiency of the EBT as a distress keyword recognizer is measured based on the following metrics: accuracy, precision, recall, and F1 score, as stated by equations (2) until (5). In an audio-based surveillance application, all metrics are important, but precision and recall are two indicators defining performance. Precision and recall values will determine whether such an audio-based surveillance system would have minimal false alarms or unidentified incidents.

1) Accuracy: It measures the overall correctness of the EBT's output. The calculation of accuracy is based on the following formula:

$$Accuracy = \frac{No.of\ samples\ predicted\ correctly}{Total\ number\ of\ samples} x\ 100\% \quad (2)$$

2) Precision and recall: These metrics provide information on how effectively the EBT performs when categorizing specific classes. Precision measures the proportion of true positives among all the positive results. In other words, precision measures the percentage of correctly identified events out of all the events detected by the EBT. The recall measures the percentage of true positives that were accurately detected. In an audio-based surveillance system, precision is important to avoid false alarms, which can be a nuisance and result in unnecessary responses by security personnel. On the other hand, recall is important to identify all critical events, even if they are rare or infrequent. The precision and recall are given by equations 3 and 4, respectively:

$$Precision = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP)+False\ Positive\ (FP)} x\ 100\% \quad (3)$$

$$Recall = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP)+False\ Negative\ (FN)} x\ 100\% \quad (4)$$

3) F1 Score: It is considered a weighted average based on precision and recall and is calculated with the following:

$$F1\ Score = \frac{2(Precision*Recall)}{(Precision+Recall)} \quad (5)$$

In an audio-based surveillance system, the F1 score can be a useful metric to optimize a balance between precision and recall, especially when there is a trade-off between the two.

III. Results and Discussion

All four datasets described in Table I were merged to form a training data set of 2880 samples. This combined data set was used for EBT training and validated with a 20-fold technique. To evaluate the performance of the trained EBT, 720 unseen samples were tested based on the distances and directions as defined in the subsequent subsections. Table II presents the results that were deduced from the confusion matrices in Fig. 6. Excellent accuracy, precision, and recall values exceeding 98% were observed during validation.

Of the testing data, the EBT has a satisfactory recall rate of 86.7%. It means the system has moderate sensitivity to every possible incident, adequate but not achieving the desired characteristic of a surveillance system. Nevertheless, a precision of 97.7% was observed, indicating the system is less susceptible to generating a false alarm, a feature of a trusted surveillance system.

TABLE II. Performance on Different Distances and Directions

Validation Performance (%)					
Metric		***Accuracy***	***Precision***	***Recall***	***F1-Score***
		99.4	99.2	99.9	99.5
Testing Performance (%)					
Metric		***Accuracy***	***Precision***	***Recall***	***F1-Score***
Distance: 0m	Facing Toward	97.5	95.1	97.5	96.3
	Facing Away	87.5	96.3	65.0	77.6
Distance: 1m	Facing Toward	95.0	97.2	87.5	92.1
	Facing Away	93.3	100.0	80.0	88.9
Distance: 2m	Facing Toward	99.2	97.6	100.0	98.8
	Facing Away	96.7	100.0	90.0	94.7
Average					
Testing data		94.9	97.7	86.7	91.9

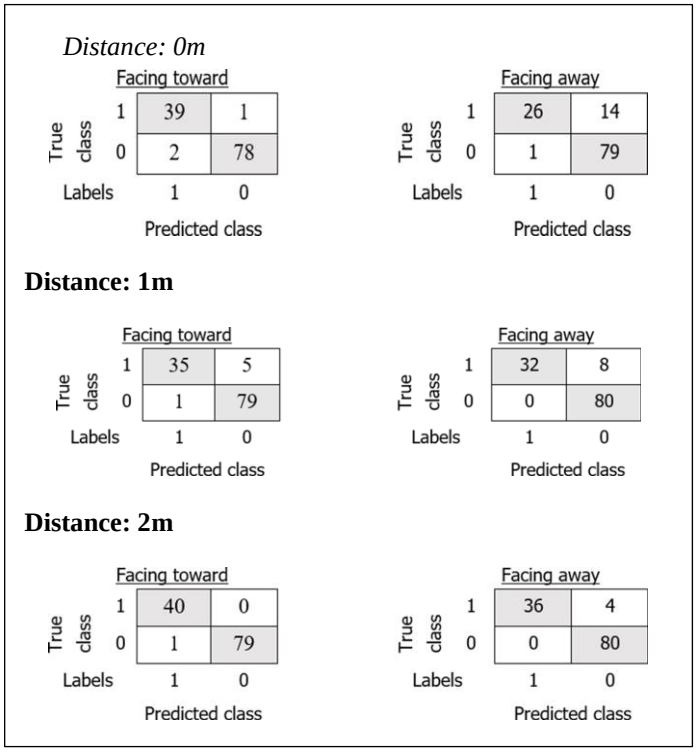

Fig. 6. The confusion matrices for different distances and directions.

A. Effect of Sound Distance

Fig. 7 shows the evaluation metrics versus distances of the testing results in Table II. Overall, EBT demonstrated high recognition accuracy with excellent precision for all distances examined. The lowest accuracy was observed at a 0 m distance, right under the microphone, possibly due to sound energy spreading from the speaker. When moving a little farther, the recognition performance improved. It is interesting to note that recall values are lowest at 0 m, slightly increased at 1 m, and highest at 2 m from the microphone.

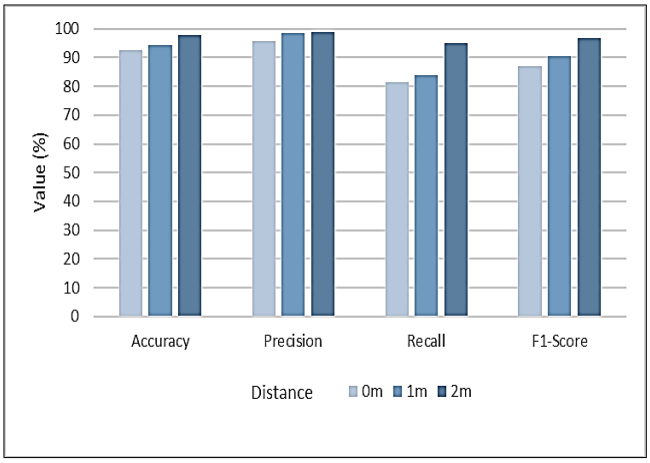

Fig. 7. Evaluation metrics versus distance.

This observation indicates that the system is more sensitive and could better recognize the targeted signals when the signals are a little farther, at one or 2-meter from the microphone. The findings corroborate that the omnidirectional microphone can capture signals from a wider distance, contributing to EBT's recognition performance in this study.

B. Effect of Sound Direction

Fig. 8 displays the evaluation metrics values for the facing toward and facing away directions. Based on the chart, it can be observed that facing toward the microphone will produce a better recognition rate. This is expected as the coverage area of the signals emitted toward the microphone is much wider. Sounds emitted from sources facing toward the micro-phone contain higher intensity than the sounds emitted from sources facing away.

Facing away from the microphone apparently affects the sensitivity of the system to recognize the targeted keyword. The overall recall value falls under 80%, in which a distance of 0 m scores the lowest. However, the recall value is gradually improving as the distance increases. At closer distances, facing away from the microphone can result in a weaker signal-to-noise ratio, which can make it more difficult to distinguish speech from background noise or interference. This can result in a lower recall of speech and a lower overall quality of the recording. However, at farther distances, the reduction in speech intensity due to facing away may be less pronounced, and the ambient noise level may also be lower, resulting in a clearer and more intelligible recording.

C. Proposed Edge Audio Surveillance

From the results, it can be inferred that a single omnidirectional microphone equipped with an EBT classifier is adequate for capturing audio in a low-noise enclosed room measuring up to 4.0 meters in length, 4.0 meters in width, and 2.794 meters in height. An audio surveillance system with an omnidirectional microphone and a processing unit that contains algorithms for signal preprocessing, feature extraction, feature reduction, and a pre-trained EBT can be developed. Fig. 9 presents the proposed edge audio surveillance.

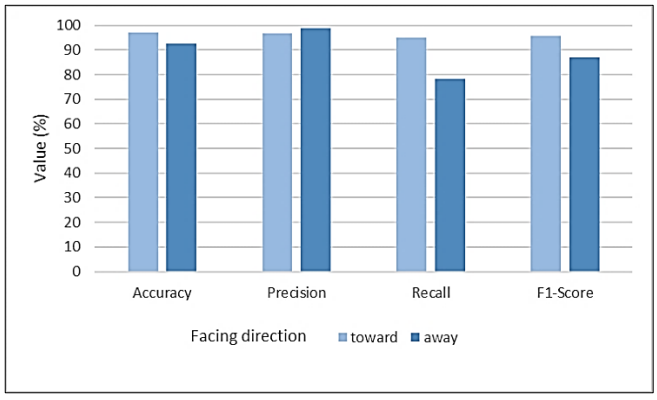

Fig. 8. Evaluation metrics versus direction.

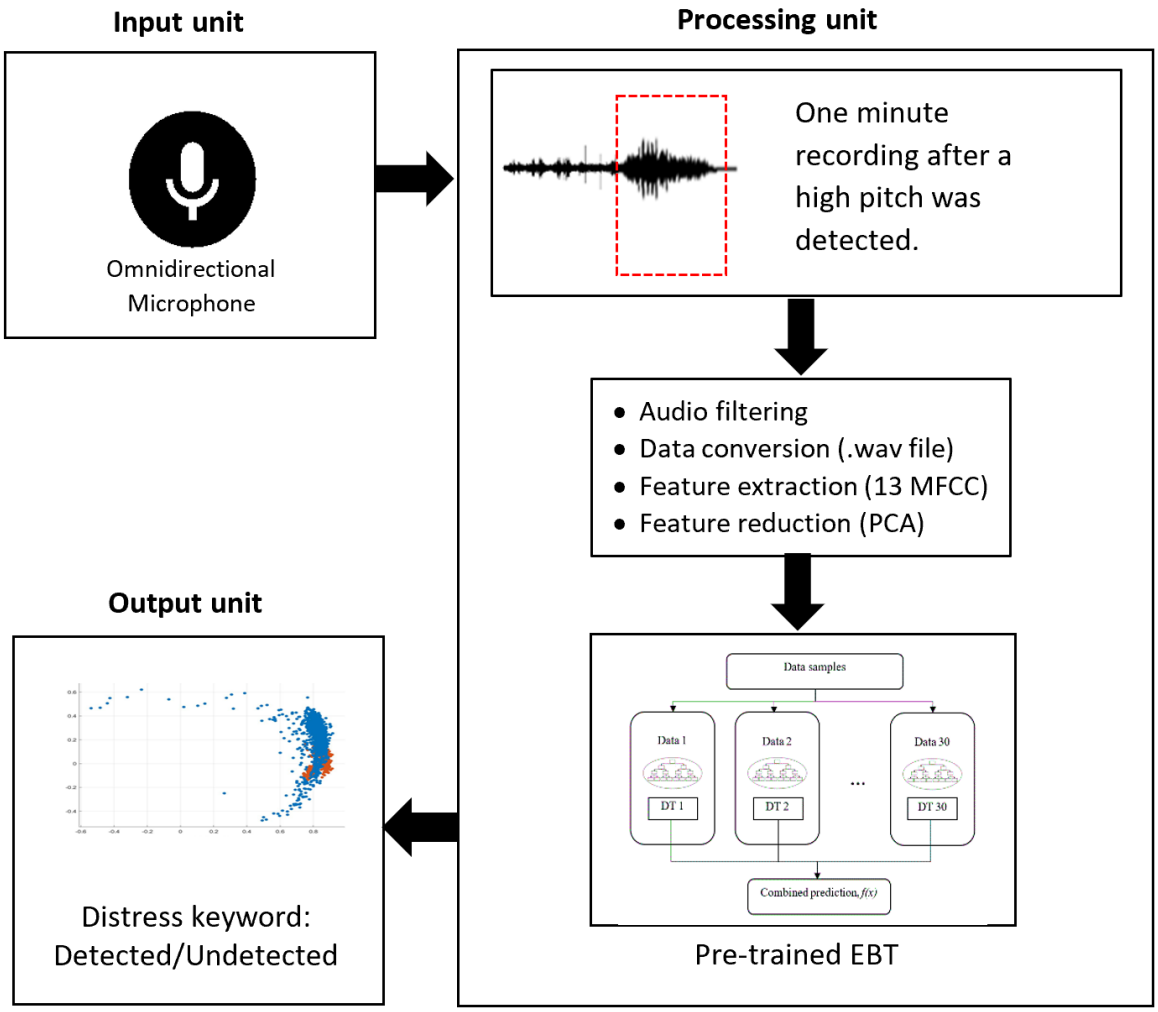

Fig. 9. Block diagram of the proposed audio surveillance system.

The processing unit can be assembled into several devices to form an edge surveillance system. A potential device would be the Raspberry Pi 4. Raspberry Pi 4 is a low-cost, small-sized computer that can run various operating systems and programming languages. The Raspberry Pi 4 has up to 8GB SDRAM and is clocked at 1.5GHz, enough processing power and memory to run simple speech preprocessing and pre-trained EBT, and can be easily connected to microphones and other output devices. The efficiency of the pre-trained EBT on the Raspberry Pi 4 will depend on factors such as the size of the EBT, the complexity of the individual trees, and the available resources on the Raspberry Pi. However, with proper optimization and tuning, it is possible to achieve efficient and accurate inference on the Raspberry Pi 4 with a pre-trained EBT model.

IV. Conclusion

This paper presents audio-based surveillance based on distress keyword recognition using an EBT classifier and an omnidirectional microphone. The experiments were conducted in a setting similar to a typical nursing room, enclosed and low-noise. The recognition performance of EBT was evaluated under different conditions, explicitly varying distances and directions of the sound sources from the microphone. Results show that the system is more sensitive and could better recognize the targeted signals when the signals were a little farther, at a one or 2-meter distance, and facing toward the microphone. It can be inferred that a single omnidirectional microphone equipped with an EBT classifier is adequate for capturing the distress keyword "Please" in a small, low-noise enclosed room.

To further enhance the sensitivity of the developed audio surveillance, expanding the dataset by incorporating various sources of both targeted and non-targeted signals is recommended. This approach will help to mitigate the occurrence of false alarms. Additionally, it is advised to increase the number of samples with background noise to address common issues related to high-pitched vocalizations, such as screaming in joy or surprise, that may trigger the distress event detector. Furthermore, implementing an adaptive noise filtering mechanism can facilitate the system's learning about the surrounding noises associated with the threshold of the distress signal speeches, thereby enhancing recognition accuracy.

REFERENCES

[1] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," IEEE International Conference on Multimedia and Expo, pp. 1306-1309, July 2005.

[2] N. Matsuo, S. Hayakawa, and S. Harada, "Technology to detect levels of stress based on voice information," Fujitsu Sci. Tech. J, vol. 51(4), pp. 48-54, 2015.

[3] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," IEEE Conference on Advanced Video and Signal Based Surveillance, pp. 21-26, 2007.

[4] D. Istrate, M. Vacher, and J. F. Serignat, "Embedded implementation of distress situation identification through sound analysis," The Journal on Information Technology in Healthcare, vol. 6(3), pp. 204-211, 2008.

[5] M. Vacher, A. Fleury, F. Portet, J. F. Serignat, and N. Noury, Complete sound and speech recognition system for health smart homes: application to the recognition of activities of daily living, 2010.

[6] C. Doukas and I. Maglogiannis, "An assistive environment for improving human safety utilizing advanced sound and motion data classification," Universal Access in the Information Society, vol. 10(2), pp. 217-228, 2011.

[7] A. Shaukat, M. Ahsan, A. Hassan, and F. Riaz, "Daily sound recognition for elderly people using ensemble methods," IEEE 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), pp. 418-423, August 2014.

[8] A. Sharma and S. Kaul, "Two-stage supervised learning-based method to detect screams and cries in urban environments," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24(2), pp. 290-299, 2015.

[9] Y. Alkaher, O. Dahan, and Y. Moshe, "Detection of distress in speech," IEEE International Conference on the Science of Electrical Engineering (ICSEE), pp. 1-5, November 2016.

[10] K. Pisanski, J. Raine, and D. Reby, "Individual differences in human voice pitch are preserved from speech to screams, roars and pain cries," Royal Society open science, vol. 7(2), p. 191642, February 2020.

[11] A. Izquierdo, L. Del Val, J. J. Villacorta, W. Zhen, S. Scherer, and Z. Fang, "Feasibility of discriminating UAV propellers noise from distress signals to locate people in enclosed environments using MEMS microphone arrays," Sensors, vol. 20(3), p. 597, January 2020.

[12] S. A. Heidari, "Video surveillance in the Iranian law; crime prevention or abuse of civil rights," Ejovoc, vol. 5(6), pp. 80-85, 2016.

[13] L. Barrett and I. Liccardi, "Accidental wiretaps: the implications of false positives by always-listening devices for privacy law and policy," Okla. L. Rev., vol. 74, p. 79, 2021.

[14] N. Nower, "Supporting audio privacy-aware services in emerging IOT environment," IJ Wireless and Microwave Technologies, vol. 3, pp. 22-29, 2021.

[15] S. Prange, A. Shams, R. Piening, Y. Abdelrahman, and F. Alt, "Priview–exploring visualisations to support users' privacy awareness," In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, , pp. 1-18, 2021.

[16] T. Kompara, J. Perš, D. Susič, and M. Gams, "A one-dimensional non-intrusive and privacy-preserving identification system for households," Electronics, vol. 10(5), p. 559, 2021.

[17] Y. Iravantchi, K. Ahuja, M. Goel, C. Harrison, and A. Sample, "Privacymic: utilizing inaudible frequencies for privacy preserving daily activity recognition," In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1-13, 2021.

[18] S. J. Neville, "Eavesmining: a critical audit of the amazon echo and alexa conditions of use. surveillance and society," vol. 18(3), pp. 343-56, 2020.

[19] B. P. Knijnenburg, X. Page, P. Wisniewski, H. R. Lipford, N. Proferes, and J. Romano, "Modern socio-technical perspectives on privacy," 2022.

[20] A. Rahman and C. T. M. Ismail, "Combating domestic violence in malaysia: issues and challenges," Man in India, vol. 99(2), 2019.

[21] A. J. Marganski and L. A. Melander, "Technology-facilitated violence against women and girls in public and private spheres: moving from enemy to ally," In The Emerald International Handbook of Technology Facilitated Violence and Abuse. Emerald Publishing Limited, 2021.

[22] A. Javed, K. M. Malik, A. Irtaza, and H. Malik, "Towards protecting cyber-physical and IoT systems from single-and multi-order voice spoofing attacks," Applied Acoustics, vol. 183, p. 108283, December 2021.

[23] A. Alsalemi, Y. Himeur, F. Bensaali, and A. Amira, "Smart sensing and end-users' behavioral change in residential buildings: an edge-based internet of energy perspective," IEEE Sensors Journal, vol. 21(24), pp. 27623-27631, September 2021.

[24] A. Osmani, M. Hamidi, and A. Chibani, "Machine learning approach for infant cry interpretation," In 2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 182-186, November 2017.

[25] C. S. Chin, X. Y. Kek, and T. K. Chan, "Scattering transform of averaged data augmentation for ensemble random subspace discriminant classifiers in audio recognition," In 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), Vol. 1, pp. 454-458, March 2021.

[26] S. L. Ullo, S. K. Khare, V. Bajaj, and G. R. Sinha, "Hybrid computerized method for environmental sound classification," IEEE Access, vol. 8, pp. 124055-124065, June 2020.

[27] A. Bhavan, P. Chauhan, and R. R. Shah, "Bagged support vector machines for emotion recognition from speech," Knowledge-Based Systems, vol. 184, p. 104886, November 2019.

[28] M. M. Chalapathi, M. R. Kumar, N. Sharma, and S. Shitharth, "Ensemble learning by high-dimensional acoustic features for emotion recognition from speech audio signal. Security and Communication Networks, 2022.

[29] Pathak, B. V., Patil, D. R., More, S. D., and Mhetre, N. R. Comparison between five classification techniques for classifying emotions in human speech. In 2019 International Conference on Intelligent Computing and Control Systems (ICCS), pp. 201-207, February 2019.

[30] N. Johari, M. Mamat, and A. Chekima, "Performance of machine learning classifiers in distress keywords recognition for audio surveillance applications," In 2021 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), pp. 1-5, September 2021.

[31] T. A. Ricketts, E. M. Picou, and J. Galster, "Directional microphone hearing aids in school environments: working toward optimization," Journal of Speech, Language, and Hearing Research, vol. 60(1), pp. 263-275, January 2017.

[32] S. C. Loeb, B. A. Hines, M. P. Armstrong, and S. J. Zarnoch, "Effects of omnidirectional microphone placement and survey period on bat echolocation call quality and detection probabilities." Acta Chiropterologica, vol. 21(2), pp. 453-464, December 2019.

[33] Smith, T. Guide To Using Omnidirectional Microphones Available Online from https://www.movophoto.com/blogs/movo-photo-blog/guide-to-using-omnidirectional-microphones (accessed on January 8, 2023).

[34] A. D'mello, A. Jadhav, J. Kale, and R. Sonkusare, "Marathi and Konkani speech recognition using cross-correlation analysis," In 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1-5, July 2021.

[35] A. Wieczorkowska, E. Kubera, and T. Słowik, "Spectral features for audio based vehicle and engine classification," vol. 50, pp. 265–290, April 2018.

[36] G. Sharma, K. Umapathy, and S. Krishnan, "Trends in audio signal feature extraction methods," Applied Acoustics, vol. 158, p. 107020, January 2020.

[37] R. Rahman, M. A. Rahman, and J. Uddin, "Automated cockpit voice recorder sound classification using mfcc features and deep convolutional neural network," In Proceedings of International Conference on Computational Intelligence, Data Science and Cloud Computing: IEM-ICDC 2020, vol. 62, p. 125, 2021.

[38] R. Sharma, K. Hara, and H. Hirayama, "A machine learning and cross-validation approach for the discrimination of vegetation physiognomic types using satellite based multispectral and multi-temporal data," Scientifica, p. 9806479, June 2017.

[39] M. Maseri and M. Mamat, "Performance analysis of implemented mfcc and hmm-based speech recognition system," 2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), pp. 1-5, September 2020.

[40] L. Van Der Maaten, E. Postma, and J. Van den Herik, "Dimensionality reduction: A comparative," J. Mach. Learn. Res., vol. 10, nos. 66–71, p. 13, October 2009.

[41] L. Breiman, "Bagging predictors machine learning", vol. 24(2), pp. 123-140, 1996.

[42] N. Saeed, "Automated gravel road condition assessment: a case study of assessing loose gravel using audio data," Doctoral dissertation, Dalarna University, 2021.

[43] A. Nagpal, "Decision tree ensembles-bagging and boosting: random forest and gradient boosting," Towards Data Science Available online https://towardsdatascience.com/decision-tree-ensembles-bagging-and-boosting-266a8ba60fd9 (accessed on December 12, 2022).