The Sight for Hearing: An IoT-Based System to Assist Drivers with Hearing Disability

Osman Salem and Ahmed Mehaoua
Centre Borelli UMR 9010
Université Paris Cité
Paris, France
{osman.salem,ahmed.mehaoua}@u-paris.fr

Raouf Boutaba
David R. Cheriton School of Computer Science
University of Waterloo
Ontario, Canada
rboutaba@cs.uwaterloo.ca

2023 IEEE Symposium on Computers and Communications (ISCC) | 979-8-3503-0048-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISCC58397.2023.10218250

***Abstract*—The objective of this paper is to propose a new system to assist drivers with hearing disability, deaf or unfocused persons by recognizing and transforming audible signals, such as emergency vehicle sirens or honks into alerts displayed in the dashboard. Such conversion from audio to alert messages attracts the attention of unfocused drivers to hear the honking of other cars, and enhances the safety and the quality of life for deaf or hard-of-hearing drivers. We develop an IoT based system to identify, denoise and translate any significant voice signal around the driver's car into alert messages. The signal acquired by sensors is processed to identify the source and display the associated message through the use of several machine learning models with majority voting. Our experiments results show that the proposed solution is able to achieve a 95% accuracy when trained and validated against a real dataset of 600 files.**

***Index Terms*—IoT; Machine Learning; Deep Learning; EVD; ANN; CNN; KNN; SVM; RF; ROC.**

I. Introduction

The World Health Organisation (WHO) is expecting around 2.5 billion people with some degree of hearing loss, with most of young adults being unconscious of the risk associated with unsafe listening practices [1]. Even if the word "deaf" is often used in conjunction with "cophotic", there are mild and moderate deafness that are not complete deafness (i.e., total hearing loss).

A poorly accompanied hearing disability can lead deaf or hard of hearing adolescents to psychological distress, anxiety, depression and even to suicide. Unfortunately, the reported suicide rate among deaf or hard of hearing people between the ages of 10 and 20 is double the rate of normal teen. The world of work and entering active life are very difficult for hard of hearing adults, especially considering the need for independence and freedom as felt by young adults. Several research works and products have been developed to assist people with such disability in their daily life activities, including gloves and armband to translate sign languages into words on a SmartPhone [2].

Deaf people have highly developed visual acuity, but remain unable to hear honks and warning from other cars. This motivate us to develop a visual solution to help these people identify honks and warnings by displaying them on the dashboard. In this manner, they can feel safe and do not need human assistance daily when driving.

Life and death issues override traffic restrictions and fire engines need to rush to fight a fire and rescue the victims. Sirens are special signals emitted by emergency vehicles such as fire trucks, police cars, prisoner transfers and ambulances. They are used to alert drivers and request priority to pass. When emergency vehicles are transporting a patient, sirens warn other drivers or pedestrians on the road. However, car drivers are sometimes distracted and unable to hear nearby sirens when they listen to loud music or radio in the car. Furthermore, the sound-isolation capabilities of modern cars, on-road publicity and sms messages may distract car drivers. Poor communication and lack of cooperation may cause delay in transporting patients to hospital emergency and even traffic accidents.

In this paper, we propose an acoustic-based method to detect the presence of emergency vehicles on the road and to display the associated warning messages on the dashboard. We focus on the detection of siren sounds on emergency vehicles, including ambulances, fire trucks and police cars. Considering that each country may have its own regulations on the type and frequency of siren sounds, our goal is to develop a highly versatile Emergency Vehicle Detection (EVD) system that can achieve a good detection accuracy with different siren types and traffic conditions. For practical applications, we roughly divide the acquired audible signal into 4 categories, including emergency sirens, police sirens, honks (or horns), and silence or insignificant noise produced by common vehicles.

We formulate the EVD problem by detecting and distinguishing sirens of emergency vehicles, police cars and horns signals using 4 sensors installed at the front and rear of the car. We created a database containing data gathered from different car models, police/fire siren sounds from different countries, abnormal sound produced by the car, and other sounds higher than 0dB and lower than 130 dB considered as insignificant noise in our system.

To develop a prototype for real time experiments, and to evaluate the performance of our proposed system, we used a Raspberry Pi 4, microphones, and 3.5 inch screen to perform our tests. The acquired audible signal from the environment is preprocessed before the classification using majority voting between 5 machine learning algorithms, namely: Support Vector Machine (SVM [3]), K-Nearest Neighbors (KNN [3]),

Random Forest (RF [3]), Artificial Neural Network (ANN [3]) and Convolutional Neural Network (CNN [3]). These models are derived from a training database of 1.5GB containing more than 600 sounds. The classification result is translated into text displayed in the screen attached to the Raspberry Pi to help the deaf by identifying the source of the noise around them. These algorithms are computationally inexpensive with low classification delay and high accuracy [4].

The rest of this paper is organized as follows. In section II, we review recent related work, while Section III presents the building blocks of our proposed approach to convert detected sounds from driving environment to visual messages on the dash board. In Section IV we present experimental results from the application of our proposed system to real time acquired sound as well as performance analysis results. Finally, section V concludes the paper and presents some perspectives for future work.

II. Related Work

Deaf people represent 11.2% of the French population [5], yet there are only few technologies to assist them in their Activities of Daily Living (ADL) and are expensive and hence inaccessible for low-income people. Moreover, few people understand sign language, which contributes to the exclusion of these 11.2% of the population. There is a dive need for system to assist deaf, hearing impaired and age related hearing loss.

While several research works have been conducted on sound recognition of ambulance sirens [6], [7], [8], [9], only a few works on sound recognition of police cars and horns are available [10], [11].

Usaid *et al.* in [6] apply MultiLayer Perceptron (MLP) to detect the siren of ambulance on the road. Their model achieves 90% of detection accuracy with a dataset of only 300 files, but their model is limited to two classes: siren and noise. Islam *et al.* in [7] apply Extreme Learning Machines (ELM) for the detection of EVD. Their experimental results on dataset (of 2000 audio clips) show a detection accuracy of 97% during classification into 2 categories: EVD and urban sounds. Cantarini *et al.* in [8] propose an emergency siren detection using low computational complexity algorithm based on CNN, with Short-Time Fourier Transform (STFT) spectrograms as features, and harmonic percussive source separation technique to improve the accuracy of the classification. Jonnadula *et al.* in [9] compare different classification methods and features for EVD. They found that Artificial Neural Network (ANN) with 3 hidden layer presents high accuracy compared to one layer. They used Google Audio Dataset with noises like people talking and horns in their experiments.

Otoom *et al.* in [12] propose an assisting device to help deaf drivers to receive GPS directions using voice recognition and speech-to-vibration. The spirit of their work is similar to ours. They map each voice navigation to vibrotactile stimulus (vibrator motors) mounted on a bracelet, where the vibrations translated by deaf drivers into 6 instructions. They extract 13 features from audible signal and classify into one of six classes (turn left, turn right, slight right, slight left, straight and silence) using machine learning algorithms. They compare the accuracy of Naïve Bayes (NB), KNN, SVM, ZeroR, OneR and RF. They found that KNN with $K = 1$ outperforms the five others.

CNN classifiers have been privileged for the recognition of sounds as they achieve better accuracy [13] in noisy environments. Trand *et al.* [14] investigated emergency vehicles deletion using CNN with the background sound of the traffic. The sounds (horns and sirens) we want to perceive are often accompanied by external nuisances (car noise, trucks, etc.). They concluded that CNN has better accuracy than other models in noisy environments.

Therefore, to benefit from the advantage of existing models under different circumstances, we will use a majority voting between the top 5 classifiers: CNN, ANN, KNN, RF and SVM to identify noise, horn, police and emergency. Our approach must classify in real time horns, sirens, etc. and the response time must be as fast as possible (no more than 1 sec), otherwise it will be useless.

The contributions of this paper are: i) a database for EVD, where a total of 600 sounds were recorded and used to derive the classification model, ii) compared to previous work, our model distinguishes 4 categories of voice instead of focusing only in EVD, namely: police, emergency, horns and noise, iii) our prototype using Raspberry Pi proves that the proposed model has low complexity and can be implemented in any micro-controller for real life deployment, iv) our prototype has an accuracy of 97% and an average response time of 0.25 seconds.

III. Proposed Approach

We assume a real deployment scenario where 4 sensors are installed on a car, 2 in the front and 2 in the back (as shown in Figure 1 in order to detect audible signals and their directions to draw the attention of hard of hearing and deaf persons.

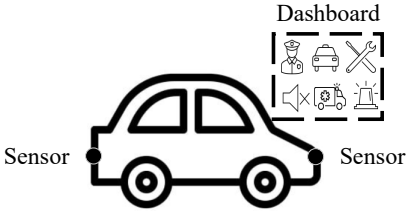


Fig. 1: model of sound assistance

We create a database containing more than 600 audible files that have been recorded from different sources: horn, emergency siren, police siren sound, city sound and neutral recording (silent). These files are converted into a spectrogram, and used as training data to derive 5 classification models: CNN, ANN, k-Nearest Neighbor (KNN), Random Forest (RF) and Support Vector Machine (SVM). The CNNs are a set of layers containing highly interconnected neurons, and they are similar to traditional Deep Neural Networks (DNN) where they receive input images, detect features in each image, and

use the extracted features to train a classifier and derive the classification model (as shown in Figure 2). However, instead of simply using a series of fully connected layers, it uses additional layers, namely convolution and pooling layers. The convolution layers (denoted by kernel) extract visual features and pooling layers reduce the size of output by retaining the most important features in each area. Pooling layers reduce the height and width of each feature map, while keeping the depth intact.

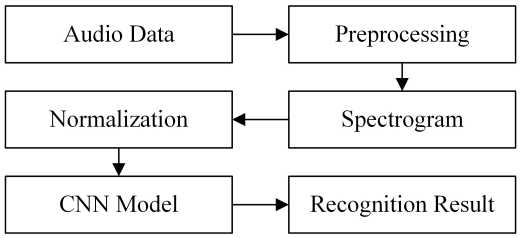


Fig. 2: Block diagram of our approach

The CNN classification can be divided in two main parts: feature learning and classification as shown in Figure 3. In the first part, a series of convolutional layers learn appropriate representations by extracting useful features from the input. In the second part, the fully connected layers act as a classifier, which processes the extracted features and assigns probabilities to each class for prediction.

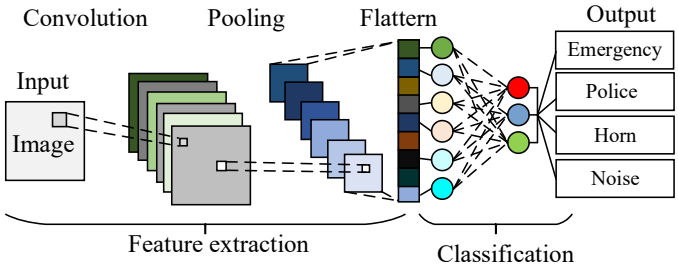
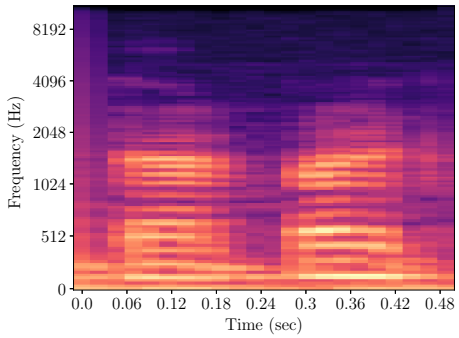


Fig. 3: CNN explanation

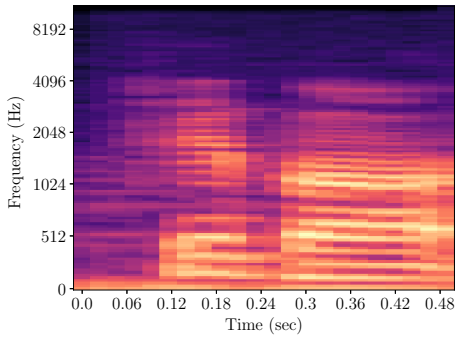
Unlike supervised learning techniques, CNN extracts automatically the features of each image in the first phase. The CNN does all the tedious work of features extraction (as shown in Figure 3) for data description through convolution and pooling. In the learning phase, the classification error is minimized to optimize the parameters and features of the classifier. Moreover, the specific architecture of the learning phase in CNN allows the extraction of features with varying complexity, from the simplest to the most complex.

The first step in our proposed approach is to transform audio file into a spectrogram, which is the visual representation of the spectrum of frequencies (or loudness of signal) over time at various frequencies. The variations are shown in Figure 4.

We are looking for data extracted from our audio tracks containing enough information to be able to separate the audible sound into 4 categories. We extracted and used the recommended 23 features to achieve the best performance.



(a) Emergency



(b) Police

Fig. 4: Image representation of audio files

The used features are: Zero Crossing Rate (ZCR), Spectral Centroid (SC), Spectral Roll-Off Point (SROP) and Mel-frequency Cepstral (20 coefficients).

The ZCR is the number of times the sign of a measurement in a time series change within a window. It measures the number of times the amplitude of the speech crosses the value zero in a given window. It is widely used as feature for speech recognition and classification. The value of ZCR over a window of n measurements is given in Equation 1:

$$ZCR = \frac{1}{2n}\sum_{i=1}^{n}|sign(x_i) - sign(x_{i-1})| \quad (1)$$

The SC is a measure used in digital signal processing to characterize a spectrum. It indicates the location of the center of mass of the spectrum and considered as the most discriminating factor between voices as it reflects the amount of high frequencies that make up the sound. The formula of the coefficient is given in Equation 2:

$$SC = \frac{\sum_{i=0}^{n-1} f_i x_i}{\sum_{i=0}^{n-1} x_i} \quad (2)$$

The Spectral Roll-Off point measures the right-side asymmetry of a sound spectrum. It is used for example to distinguish sound from normal speech. The Mel-Frequency Cepstral (MFCC) are cepstral coefficients that correspond to a sinu-

soidal transformation of the power of a signal. The MFCC are the most important features where the frequency bands are distributed according to the Mel-scale. We used 20 MFCC to derive the shape of the signal. For clarification, the variations of audio sound from horn file is shown in Figure 5, and the spectrogram with the extracted features are presented in Figures 6, 7, 8, 9 and 10.

The KNN model is used to classify test data into one of the 4 classes by looking at the values of the k nearest neighbors. The RF classifier derives a multitude of decision trees from the training data, and the class of the majority voting of these trees is the output class of the associated input test record. SVM identifies the hyperplane that achieves the best separations among classes.

To improve the accuracy of the CNN in classifying the captured sounds, we applied a majority voting between 5 classifiers. The choice of 5 classifiers has been achieved through experiments and performance analysis, where the accuracy of voting between 5 classifiers outperforms that of 3 classifiers, and achieves a tradeoff between complexity and classification delay. The slight or insignificant increase in the accuracy when increasing the number of classifiers justify our choice.

If the majority vote for the horns or sirens (3/5, 4/5 or 5/5), the system display the associated message on the dashboard within 0.5 sec. To improve the speed of driver reflection, we split the files into pieces of sound that do not exceed 0.5 sec with an overlap of 0.1 sec to be sure that there is a complete sound to improve the classification efficiency. This will allow us to optimize our model in order to improve its efficiency and to process the data in a safer and faster way.

Algorithm 1: EVD approach

```
Input  : Audible signal
Output: Emergency, Police, Horn, Noise
for i in range(4) do
    Record an audio chunk of 100ms
end
while True do
    Record a chunk of 100ms
     Concatenate it with the 4 previous chunks
     Extract features
     Res = [0,0,0,0]
     C=[]
     for i in range(5) do
        C.append(Classification[i])
     end
     if majority_voting(C) is i then
        Turn on led[i]
     end
end
```

IV. Experimental results

To conduct experiment and performance analysis of our proposed model for the detection and classification of voice to assist deaf and hard-of-hearing drivers, we start by building a database containing audio files collected from different sources and labeling them using 4 categories (or separating them into 4 directories): Emergency, Police, Horn and Noise. The training dataset must have large number of diverse audible files for building deep neural networks, and thus must be sufficiently large and variable in terms of sound specifications and recording conditions. The dataset is collected from real scenarios and from online sources that professionally provide audio clips of emergency vehicles. In other words, the dataset was collected from different sources and it does not contain simulated measurements. The classes are equally distributed in the training dataset. To display the shape of some raw audio data in the training set, Figures 11, 12 and 13 show the variations of the audio content for siren from emergency car, police car, and horn respectively. The waveforms are the amplitude variations with time. The preprocessing phase takes place after the collection of data, where un-exploitable and distorted files are discarded after manual check. Afterward, the files have been cropped to remove noise and to split them into 500 ms for achieving better accuracy.

As we extracted 23 features from sounds and used them as input for the classification model, the variation of the ZCR, SC and SROP are presented in Figure 14, the variations of the first 5 coefficients of Mel-frequency Cepstral are presented in Figure 15, while the variations of the remaining 14 coefficients (from 6 to 20) are presented in Figure 15.

We first conduct experiments to evaluate and analyse the accuracy of each classification model on the dataset. The ROC for each class (emergency, noise, police and Horns), as well as the micro and macro-average are shown in Figure 17 for SVM, Figure 18 for KNN, Figure 19 for RF. These algorithms achieve a good accuracy for noise detection (higher than 96%), with lower accuracy in the identification of other classes (77%, 84% and 93%). The SVM has been proven to be optimal and able to achieve high accuracy and low false alarms. However, when comparing the ROCs in Figures 17, 18 and 19, we notice that RF achieves the highest performance (AUC=0.96) followed by KNN (AUC=0.89) and SVM (AUC=0.85). The RF achieves better than linear SVM over the used dataset, and the derivation of the RF classification model was faster than SVM. In fact, the computational complexity for deriving the classification model in SVM is $\mathcal{O}(n^3)$, and n is the number of records in the training dataset.

The ROCs curves for the deep learning neural network ANN and CNN and our approach are presented in Figures 20 and 21 respectively. Their detection accuracy outperforms the previous 3 algorithms (SVM, KNN & RF), where the AUC is 0.96 for ANN and 0.99 for CNN. Finally, we analyse the classification accuracy of our majority voting in the identification of the four classes. The ROC curve is presented in Figure 22, where the AUC reaches 99% in the recognition of emergency cars and 98% for police cars. Figure 22 proves that the accuracy of MV approach outperforms the accuracy of CNN and ANN.

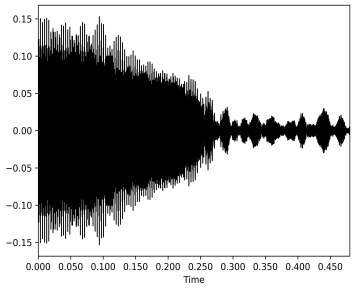


Fig. 5: Horn

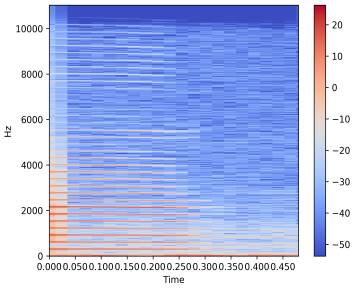


Fig. 6: Spectrogram

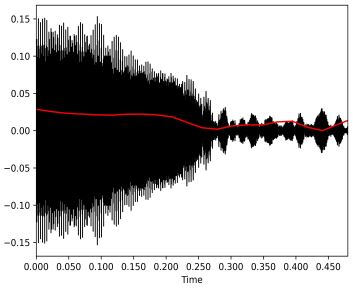


Fig. 7: Spectral Centroid

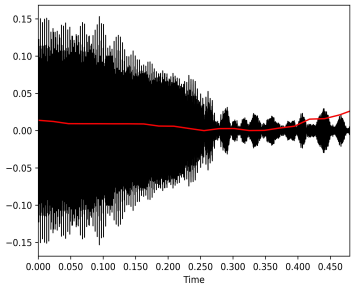


Fig. 8: Spectral Rolloff

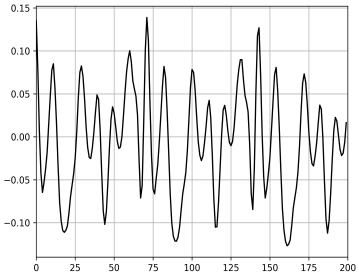


Fig. 9: Zero Cross Rate

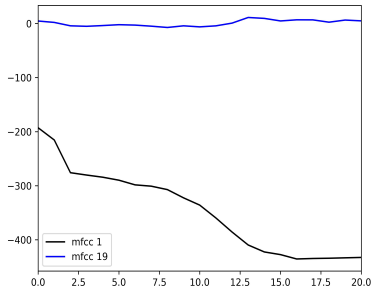


Fig. 10: MFCCs 1 & 20

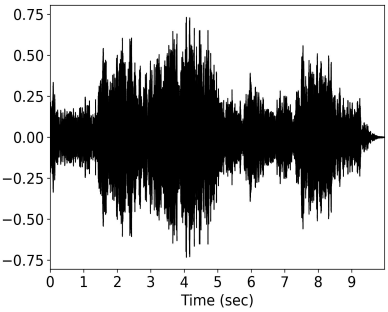


Fig. 11: Emergency car

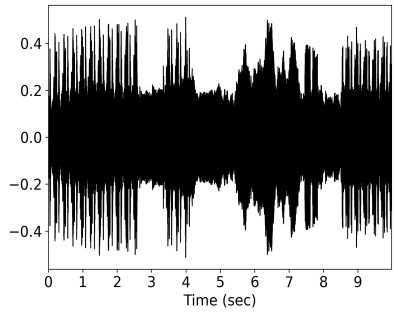


Fig. 12: Police car

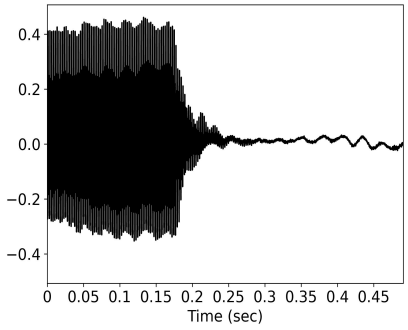


Fig. 13: Horns

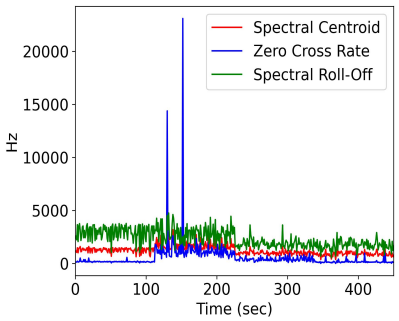


Fig. 14: SC, ZCR & SROP

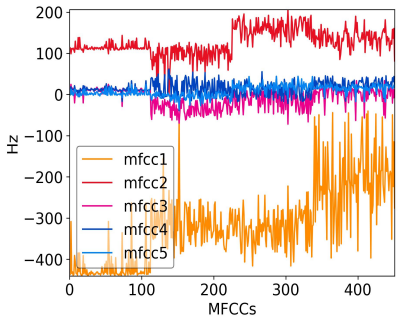


Fig. 15: MFCCs {1-5}

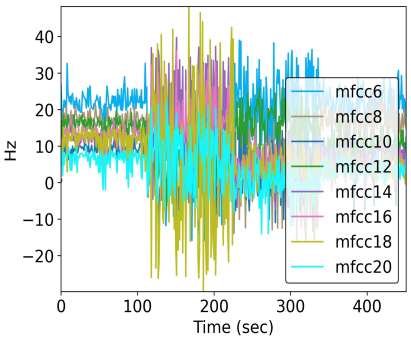


Fig. 16: MFCCs {6-20}

V. Conclusion

In this paper, we proposed a deep learning based model to identify horns, police cars and emergency car sirens to assist hearing impaired or deaf people. However, the hearing impaired or deaf are not the only people who can benefit from this work, where distracted people with all their hearing

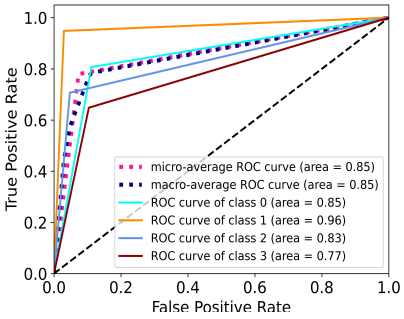


Fig. 17: SVM

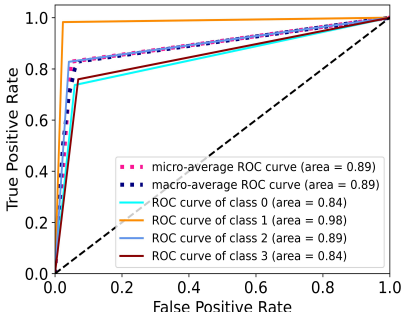


Fig. 18: KNN

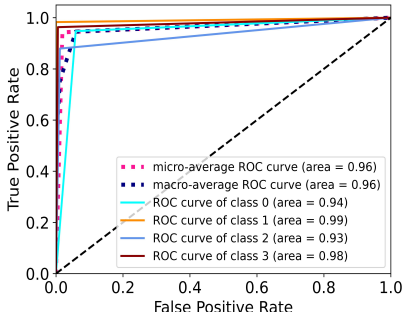


Fig. 19: Random Forest

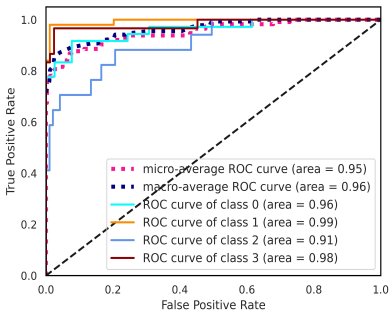


Fig. 20: ANN

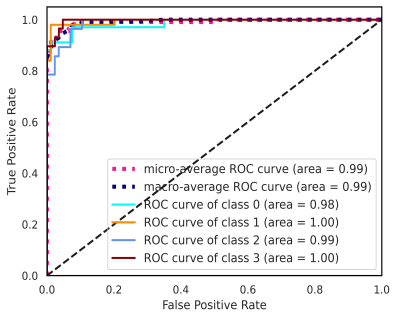


Fig. 21: CNN

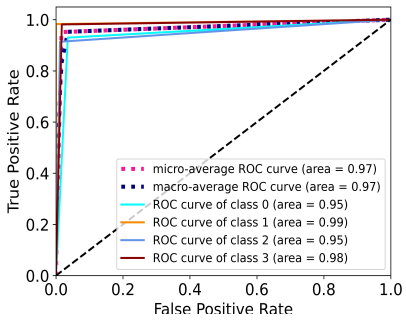


Fig. 22: MV

faculties can miss a horn or a siren noise. These include people listening too loudly to radio, people distracted by noise from within the car. We build a database containing 600 different sounds to derive the classification. The files are recorded in real environments with background noise, and used to extract features as input to the classification model. We derive five classification models using the top 5 optimal algorithms and we apply a majority voting. We conduct performance analysis using the ROC curve and achieved detection accuracy of 97% for sounds with noisy environment. We built a prototype using a Raspberry PI 4 demontrating the feasibility of our approach for large-scale deployment to the public.

As future work, we intend to work on improving the detection performance using larger datasets containing more sounds and additional categories of sound. We also intend to work on noise filtering to enhance the detection accuracy of in-exploitable voice with audible noises, and to enhance message transmission for deaf people safety.

REFERENCES

[1] World Health Organization, "Deafness and hearing loss," https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss, April 2021.

[2] Park, HyeonJung and Lee, Youngki and Ko, JeongGil, "Enabling real-time sign language translation on mobile platforms with on-board depth cameras," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 2, pp. 1–30, 2021.

[3] A. Geron, *Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems*. Sebastopol, CA: O'Reilly Media, 2017.

[4] Z. K. Maseer, R. Yusof, N. Bahaman, S. A. Mostafa, and C. F. M. Foozy, "Benchmarking of machine learning for anomaly based intrusion detection systems in the cicids2017 dataset," *IEEE access*, vol. 9, pp. 22 351–22 370, 2021.

[5] Christel Colin, Roselyne Kerjosse, "Handicaps-incapacités-dépendance," DREES, Report 16, 2020, Last visited November 2022. [Online]. Available: https://drees.solidarites-sante.gouv.fr/sources-outils-et-enquetes/02-les-enquetes-handicap-sante

[6] M. Usaid, M. Asif, T. Rajab, M. Rashid, and S. I. Hassan, "Ambulance siren detection using artificial intelligence in urban scenarios," *Univ. Research Jour. of Eng. & Technology*, vol. 12, no. 1, pp. 92–97, 2022.

[7] Z. Islam and M. Abdel-Aty, "Real-time emergency vehicle event detection using audio data," *arXiv preprint arXiv:2202.01367*, 2022.

[8] M. Cantarini, A. Brocanelli, L. Gabrielli, and S. Squartini, "Acoustic features for deep learning-based models for emergency siren detection: An evaluation study," in *12th Inter. Symposium on Image and Signal Processing and Analysis (ISPA)*. IEEE, 2021, pp. 47–53.

[9] E. P. Jonnadula and P. M. Khilar, "Comparison of various techniques for emergency vehicle detection using audio processing," in *Cloud Security*. CRC Press, 2021, pp. 64–75.

[10] C. A. Dim, R. M. Feitosa, M. P. Mota, and J. M. d. Morais, "A smartphone application for car horn detection to assist hearing-impaired people in driving," in *International Conference on Computational Science and Its Applications*. Springer, 2020, pp. 104–116.

[11] C. A. Dim, R. M. Feitosa, M. P. Mota, and J. M. de Morais, "Alert systems to hearing-impaired people: a systematic review," *Multimedia Tools and Applications*, pp. 1–20, 2022.

[12] M. Otoom, M. A. Alzubaidi, and R. Aloufee, "Novel navigation assistive device for deaf drivers," *Assistive Technology*, vol. 34, no. 2, pp. 129–139, 2022.

[13] T. Kim, M. Yoo, D. K. Shin, G. Park, and S. Kim, "A study on the sound recognition method of autonomous vehicle using cnn," *Int. J. of Intelligent Systems and Applications in Engineering*, vol. 10, no. 1s, pp. 158–162, 2022.

[14] V.-T. Tran and W.-H. Tsai, "Acoustic-based emergency vehicle detection using convolutional neural networks," *IEEE Access*, vol. 8, pp. 75 702–75 713, 2020.