PRIVACY AND VOICE SCREENING

Contents

November 2022

Authors: Natasha Pegler, Steven Liu, Chen Si, Xuan He (list the name of group members)

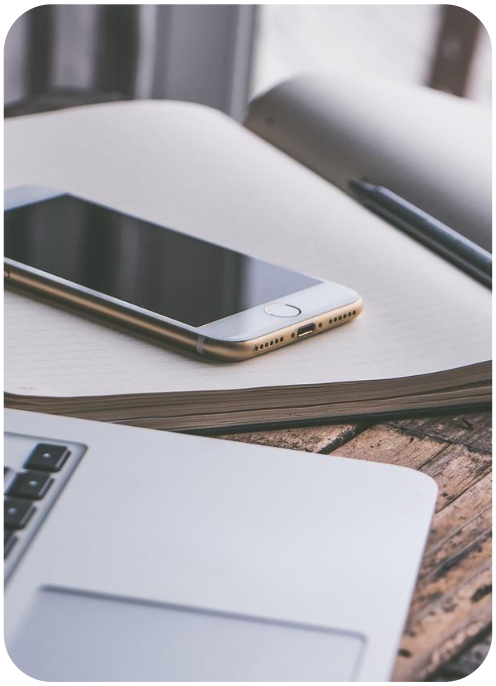
This document is produced as one of the deliverables of the 'responsible innovation project' in the course 'Professional Practice 2' convened by Dr Ehsan Nabavi at Australian National University.

Address Information

The Australian National University, Peter Baume Building #42A, Linnaeus Way
Acton ACT 2601 Australia
Phone: +61 2 6125 5394 | **Email:** ehsan.nabavi@anu.edu.au, cpas@anu.edu.au

ABSTRACT

Audio recognition will be a key part of the future of communication. It can listen to what people say, and interpret it to a digitized version that reads and analyses their words. Environmental sound identification is a field based on using machine learning models to identify sources of noise. This paper firstly presents the problem statement of a privacy issue that should be considered in audio recognition. Secondly, this paper briefly describes the background to the application of sound recognition in the environment. Thirdly, we will outline the proposed solution based on deep learning and machine learning algorithms. In addition, the paper identifies key stakeholders, and discusses their influence on the development and application of environmental sound recognition technology. Finally, this report will connect the knowledge of responsible innovation with machine learning in order ensure our solution can help in responsible environmental sound recognition and achieve greater results and progress in the industry.



PROBLEM STATEMENT.

With the rapid development of technology in the 21st century, voice recognition technology has been iterated and updated to reach a state of usability in noisy environmental scenarios. Since the rise of deep learning technologies, the technical shortcomings of voice recognition in terms of accuracy and speed have been reduced and industry acceptance of voice recognition has increased. However, with the rapid development of the technology, the hidden privacy concerns have been exposed to society and have attracted significant ethical and legal attention.

In particular, environmental sound recognition technology is generally focused on non-human sources of noise (such as industrial or environmental noise). As a result, the privacy impacts of environmental sound recognition technology have not been well explored. People tend to ignore the importance of their privacy when they are in an external environment, or may not be aware that they are under surveillance. So, in a sense, the detrimental consequences of private information being known to others will inevitably arise in the context of the ambient sound recognition technology currently being developed.

BACKGROUND & CONTEXT

- **Environmental sound recognition (ESR):** Environmental sound identification is a cutting-edge technology that integrates multidisciplinary knowledge, covering basic and cutting-edge disciplines such as mathematics and statistics, acoustics and linguistics, computers and artificial intelligence, and is a key link to classify audio in the technology of human-computer and natural interaction. A monitor unit is installed in a particular environment, where it records and analyses detailed sound levels, and typically then analyses the source of the sound.

- **Industrial applications:** ESR monitors can measure noise pollution, track vehicle movement or motor vehicle sounds. They can also monitor different scenarios, such as street scenes, indoor scenes and car scenes. The main purpose of industrial sound event detection is to detect the presence of a target sound event within a continuous audio stream, e.g., to detect anomalous sounds from faulty equipment or sounds from an accident scene.

- **Environmental applications:** ESR can be used for surveying bird or frog populations and listening for illegal logging. For example, monitoring the impact of logging on biodiversity in a particular area is necessary to better protect forest biodiversity[1]. Acoustic data is collected to understand the impact of forestry reforms on forest biodiversity. Sound recognition can also be used to save endangered animals. Developments in bioacoustics are now already changing the way conservation works, and scientists predict that using this method has great potential to change the way we monitor species, assess the health of ecosystems and evaluate the impact of humans on nature.

PROPOSED SOLUTION

Introduction of solution

In the process of our research, we found that at this stage there is no reasonable technology that has been developed to address the privacy concerns surrounding ESI. So, we come up with the solution that helps us to avoid the privacy issue.

Environment Sound Recognition (ESR) has been widely used in audio retrieval, audio forensics and other situational awareness and wearable based applications as an effective method to perceive the surrounding environment. Currently, simpler classifiers have been intended for use in ESR problems, but do not reflect and recognize human and environmental sounds well, let alone extract human voices and avoid privacy issues. In our exploration of machine learning algorithms as a high performance, multi-layer technology development industry, a more effective way to better characterize raw data and solve model recognition problems is provided. To this end, this paper will apply a simple machine learning model to the ambient sound recognition problem with feature separation of audio features, and train the model to deepen its ability to recognize ambient scenes.

Application of solution

Our solution consists of a simple machine learning model that can distinguish between audio with voice and audio without voice. The model would be installed on devices in the field, and would check data immediately after recording. If voice is detected in a sample, then that data is excluded from further analysis; if not, the data goes through the process as normal. This means the monitor mostly functions as normal. But voice data is excluded at the earliest possible point in the process, and is never stored – so privacy impacts are minimized.

A more detailed view of the process is as follows. The monitor records data in small time steps (less than one second). However, successfully identifying voice data will require longer time steps. So, the input data is put in a cache until there is enough to run our screening model. Once there is enough (approximately 5-10 seconds), run the screening – if it's voice data, then set the values of the cache data to 0 and flag them as voice screened. Then release the cache into the monitor's normal process.

Here, we also propose an approach algorithm to achieve the human speech filter. Firstly, to minimize the energy consumption, TensorFlowLite is one of the best solutions since it is designed to suit small-size IoT(Internet of Things) devices[2]. Moreover, TensorFlowLite allows low-latency inference of on-device machine learning models.

To train the mode, we employ the Mel-frequency cepstrum algorithm. This algorithm uses a short-time Fourier transform to convert the audio data in time-domain signal into frequency-domain signal. Then the frequency signal is mapped onto the Mel frequency scale, using a logarithmic representation of the sound power at each frequency. Through Mel cepstrum analysis, discrete cosine transformation is used to separate the DC (linear) signal component and the sinusoidal (oscillating) signal component[3]. At the end, we would extract the sound spectrum feature vector and convert the vector into an image. The feature statistics then form the inputs to our speech audio filter model.

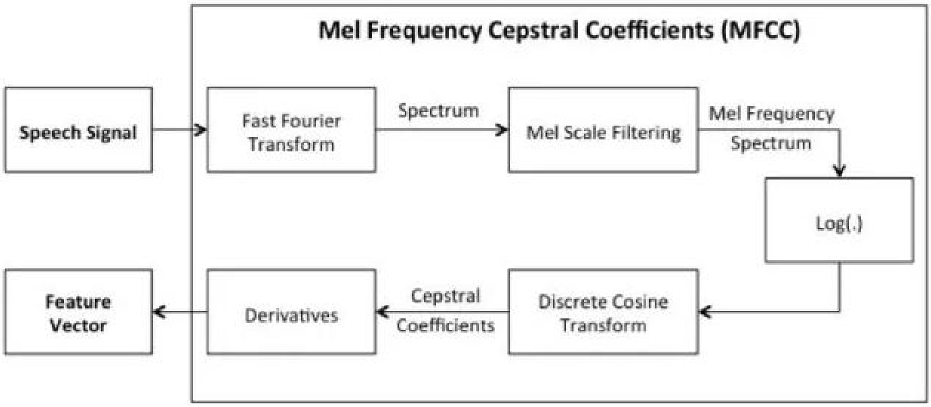


Diagram. Mel Frequency Cepstral Coefficients.

Overall, our goal is to use the speech audio filter model to detect whether the audio contains any human voice. The filter would directly discard the audio data if it contains human speech, otherwise the data would progress through to the rest of the analysis. The diagram below gives a straightforward expression illustrating the logical data filter process.

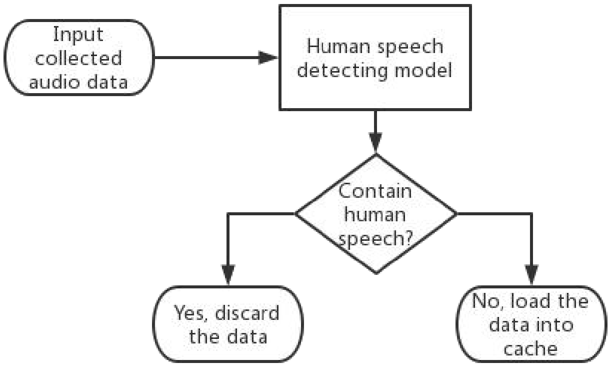

Input collected audio data
Human speech detecting model
Contain human speech?
Yes, discard the data
No, load the data into cache

Diagram. Filter process.

STAKEHOLDERS & IMPACTS

Representing the importance of the stakeholders and their impacts is necessary for developing the environmental voice recognition technology. We have considered the various stakeholders of the project, including their level of impact as well as the level of influence they have over the project; they are shown in the table below.

	High influence	Low influence
High power	Audio monitoring manufacturers/developers Audio ML developers	Government
Low power	Public users	Data customers Audio monitoring customers

Low power/ Low interest

Customers of audio monitoring and data services are placed in the low power/low interest quadrant. There is a weak relationship between them and us. These customers concern about how the screening process impacts on their business (minimal interest).

Low power/ High interest

In the low power/high interest quadrant are the general public – people who may be recorded by the monitors. Our primary goal is to protect their privacy.

High power/ Low interest

Government plays the role in the high power/low interest quadrant. Legal requirements could make our solution particularly desirable, or undesirable.

High power/ High interest

In the high power/high interest quadrant are our immediate customers – operators and manufacturers of audio monitoring systems. Machine learning developers are also in this category. Both of these groups stand to benefit from our system, so long as their requirements are met. The benefits include:

- **Assisting with responsible practice** – they can use the privacy screening to promote themselves to potential customers or to the general public

- **Reduced data collection** – Less data means less storage space required and lower network bandwidth use

Legal protections – they don't have to worry about legal requirements for collecting and storing personal data. For example, some areas of the USA require 'two-party consent' for recording conversations – that is, everyone being recorded must give consent to the recording. Our solution would let noise monitors operate in these areas without worrying about potential legal issues.

RESPONSIBLE PRACTICE

The paper shows the understanding of responsible innovation for the design of the environmental voice recognition. The concept of responsible innovation is a deepening and extension of 'sustainability' in this day and age, trying to bridge the gap between technological systems and effective deployment by analyzing the whole range of stakeholders with whom users interact most closely with technology, taking full account of the consequences of new technological applications in the natural and social environment and the range of impacts they face. The results and choices involving social needs and ethical values are also fully considered, based on an integrated assessment of both, in order to find functional needs as a basis for the design and development of new research, products and services.

The point of view for customers' needs

In order for our system to be viable for use, it needs to meet some key customer requirements. These will be considered in the design and development of the system.

Minimal computation and power usage: Our system will run on devices in the field, which are likely to have only limited battery charge and computational power. Our model should be able to run on these devices without impacting their normal operation.

Traceability: It should be clear when data was excluded due to the screening process (as opposed to a device fault, for example).

The point of view for social responsibility

Avoid marginalizing anyone: Ideally our system would work equally well detecting male and female voices across all languages and accents. We should use a diverse set of training data, and if the screening process doesn't work as well for certain groups, we should make sure our stakeholders are informed. For example, if performance is poor for certain languages, then our system may not be suitable for regions with significant populations that speak those languages.

What happens if capturing voice data is desirable? For example, the monitor might record voice data from a crime. Screening out this data means erasing evidence. Whether the social benefits outweigh the privacy concerns depends on the culture – there isn't one right answer.

CONCLUSION

In summary, sound recognition will continue to develop into the future and will see increasing application in many different contexts. These advances will be the result not only of improvements in the identification algorithms, but also technology upgrades across the entire industry chain. These rapid advances mean that responsible innovation practices will be important in ensuring the new technologies benefit society. Our proposed system seeks to address one specific problem: the potential for sound recognition technology to impact privacy of people in the vicinity. Our screening system would minimize privacy impacts and help sound identification operate in a more ethical and responsible way.

REFERENCE

[1] Adam Welz, "Listening to Nature: The Emerging Field of Bioacoustics", (website: https://e360.yale.edu/features/listening-to-nature-the-emerging-field-of-bioacoustics Accessed Time: 2022.09.02)

[2] https://www.tensorflow.org/mobile/tflite.

[3] Ming Y. ,Yuqian Y., Hao D., Zhe W., "Text-dependent speaker recognition method using MFCC and LPCC features"[J]. Computer Applications, 2006(04) 883-885.

[4] Ketan Doshi, "Audio Deep Learning Made Simple: Sound Classification, Step-by-step",(website: https://towardsdatascience.com/audio-deep-learning-made-simple-sound-classification-step-by-step-cebc936bbe5. Accessed Time: 2022.09.02)

[5] Mporas, I., Perikos, I., Kelefouras, V., & Paraskevas, M. (2020). Illegal logging detection based on acoustic surveillance of forest. *Applied Sciences, 10*(20), 7379. doi:https://doi.org/10.3390/app10207379.

[6] Zgank A. Bee Swarm Activity Acoustic Classification for an IoT-Based Farm Service. Sensors (Basel). 2019 Dec 19;20(1):21. doi: 10.3390/s20010021. PMID: 31861505; PMCID: PMC6982799.