Privacy and Voice Screening

For Environmental Sound Identification

**Contents**

[1. Abstract 3](#_Toc523734708)

[2. Problem statement 3](#_Toc523734709)

[3. Background 4](#_Toc523734710)

[4. Proposed Solution(s) 4](#_Toc523734711)

[a. Introduction of Solution…… 4](#_Toc523734712)

[b. Application of Solution……..5](#_Toc523734713)

[5. Stakeholders & Impacts 5](#_Toc523734714)

[6. Responsibility & Conclusion 5](#_Toc523734714)

[7. Reference 6](#_Toc523734715)

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](mailto:ehsan.nabavi@anu.edu.au), [cpas@anu.edu.au](mailto:cpas@anu.edu.au)

## Abstract

|  |  |  |
| --- | --- | --- |
| Voice 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. For environmental voice identification, it is a field based on using machine models to identify sources of noise. This paper firstly presents the problem statement of privacy issue that should be considered in voice recognition. Secondly, this paper briefly describes the background to the application of sound recognition in the environment. Thirdly, the article will outline the proposed solutions based on the deep learning and machine learning algorithm. It helps to show the ideas from a professional perspective on how solutions can be clearly expressed and understood. In addition, discussion about stakeholders and their impacts can influence the development and application of environmental sound recognition technology. Finally, this report will connect the knowledge of responsible innovation with machine learning in order to provide the technology that can help in 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' protections have been exposed to society and have attracted significant ethical and legal attention.

In particular, in environmental sound recognition technology, which was originally used to study various sounds in ecological environments over time, for the public good. People tend to ignore the importance of their privacy when they are in an external environment. The definition of privacy is that it represents a type of personal information that is not wanted to be known by others, including not being known in a particular form. So, in a sense, the detrimental consequences of privacy being known to others will inevitably arise in the context of the ambient sound recognition technology currently being developed.

## Background & context

* **Environmental sound identification:** Environmental voice recognition 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. Recognition targeting the sounds of the human surroundings. A monitor unit is installed in a particular environment, where it records and analyses detailed sound levels.
* **Industrial applications:** They 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:** They include 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 environmental sound recognition technology that has been developed to meet the desired effect. 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, leave 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

The simple machine learning model is developed 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 the situation of identifying the voice sample happens, then that data is excluded from further analysis; if it doesn’t happen, 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.

Here’s showing more detailed view of the process. 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, 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 audio sensor energy consumption, invoking TensorFlowLite is one of the best solutions since it is designed for suiting for small-size IoT(Internet of Things) device[2]. Moreover, TensorFlowLite allows low-latency inference of on-device machine learning models.

To train the speech audio data, we employ the Mel-frequency cepstrum algorithm. The nature of this algorithm is using short-time Fourier transform to convert the audio data in time-domain signal into frequency-domain signal. Then the frequency can be transformed to the linear relation through Mel-frequency cepstrum algorithm, which can be perceived by human ears. Through Mel cepstrum analysis, DCT transformation is used to separate the DC signal component and the sinusoidal signal component[3]. At the end, we would extract the sound spectrum feature vector and convert the vector into an image. With the feature statistics, we could input the features to generate our speech audio filter model.

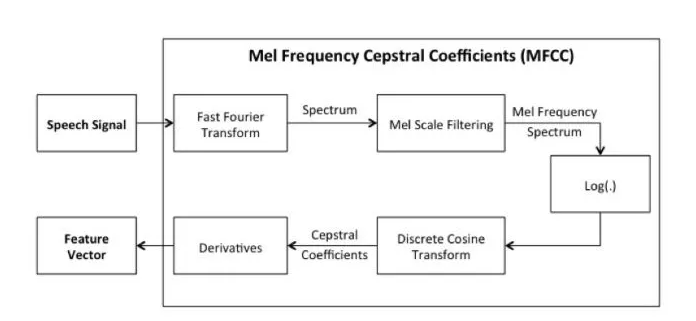


Diagram. Mel Frequency Cepstral Coefficients.

Overall, we’d like 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 be loaded to cache. Below diagram gives a straightforward expression illustrating the logical data filter process.

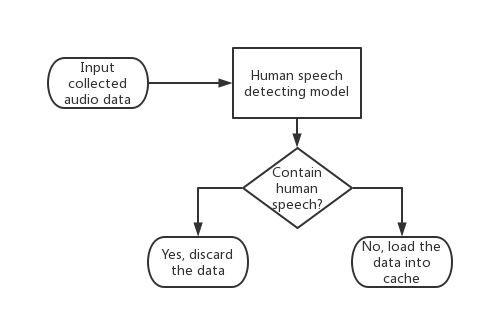


Diagram. Filter process.

## STAkeholders & impacts

Representing the importance of the stakeholders and their impacts is necessary for developing the environmental voice recognition technology. The mapping for stakeholders of the project shows the different level of power corresponding with different level of interest.

#### 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:

·     **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.

## Responsibility & conclusion

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

**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.

**What happens if capturing voice data is desirable?** For example, the monitor might record 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.

In summary, voice recognition will tend to develop in the direction of far-field and incorporation in the future, which cannot only be the advancement of algorithms, but requires the common technology upgrade of the whole industry chain. In terms of technological progress and industrial development, there is a good positive iterative effect between technology and industry, and linking technological innovation with social responsibility can better help speech recognition technology to progress rapidly and solve practical problems. Environmental sound identification is a new field with many applications. However, it has the potential 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: 2022.10.18)
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.

#### Level 4 heading goes here

On the Insert tab, the galleries include items that are designed to coordinate with the overall look of your document. You can use these galleries to insert tables, headers, footers, lists, cover pages, and other document building blocks. When you create pictures, charts, or diagrams, they also coordinate with your current document look.

#### Level 4 heading again

You can easily change the formatting of selected text in the document text by choosing a look for the selected text from the Quick Styles gallery on the Home tab. You can also format text directly by using the other controls on the Home tab. Most controls offer a choice of using the look from the current theme or using a format that you specify directly.

## Level 2 Heading AGain

### Level 3 heading goes here

On the Insert tab, the galleries include items that are designed to coordinate with the overall look of your document. You can use these galleries to insert tables, headers, footers, lists, cover pages, and other document building blocks. When you create pictures, charts, or diagrams, they also coordinate with your current document look.

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Data Set 1** | **Data Set 2** | **Data Set 3** |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

#### Level 4 heading goes here

On the Insert tab, the galleries include items that are designed to coordinate with the overall look of your document. You can use these galleries to insert tables, headers, footers, lists, cover pages, and other document building blocks. When you create pictures, charts, or diagrams, they also coordinate with your current document look.