Acoustic Sensing for Assistive Living: Investigating ML Models to Address Privacy and Data Collection Challenges

Pavlos Nicolaou

School of Engineering
University of Kent

2024

Abstract

Assistive living technologies have advanced significantly, utilising wearable sensors and other smart devices. While wearables are effective in monitoring human activity, they may become impractical when the elderly are resistant to them. In contrast, audio sensing can passively monitor daily routines without the burden of managing wearable devices. This research is divided into three parts: the first focuses on an unsupervised application of acoustic sensing, the second distinguishes data collection and application challenges in an acoustic environment, and the third presents a proposed pipeline that can solve privacy concerns regarding speech in an acoustic environment. Audio as a modality is rich with contextual information and can be very useful in sensing activities in a home environment. Because of the nature of an audio-sensing environment, there are some restrictions we have to keep in mind in terms of data collection and privacy. Firstly, having a large amount of labelled data is not possible. Therefore, this thesis's first part focuses on an unsupervised approach to detect changes in daily routine in an acoustic environment with a 14% improvement in the F1 score compared to the previous baseline. The second part of our research sets the challenges of any acoustic home environment regarding data collection and speech privacy, showcasing the solution deployed in the industry. Finalising this thesis with the creation of a three-stage methodology on developing a privacy pervasive system, which is a system that ensures privacy is maintained at all stages of data processing, regarding speech audio data and the possibility of having a passive audio sensing system that discards all speech data and then having the possibility to apply any audio sensing techniques without compromising privacy. The proposed methodology achieved this by eliminating all private conversations from the acoustic signal, resulting in a minor reduction of 2-13% in the F1 score for acoustic activity detection across various datasets.

Acknowledgements

I would like to take this opportunity to express my gratitude to everyone who has supported me throughout this thesis.

First and foremost, I am grateful to my supervisor, Dr Christos Efstratiou, for his unwavering support, guidance, and encouragement throughout this journey. His valuable insights and expertise have been instrumental in shaping my research and helping me grow as a researcher.

I'm also grateful for the opportunity and exposure to real deployments in smart care homes that were provided by MiiCare. I would like to thank you for sharing your knowledge regarding the deployments and the efforts that come with it.

I would also like to thank my second academic advisor Dr Chee Siang (Jim) Ang and my Progression Review Panel Dr Konstantinos Sirlantzis and Dr Xinggang Yan for their valuable advice throughout the years I was part of this process.

I am deeply grateful to my internal examiner, Dr Moinul Hossain, and my external examiner, Dr Paul McCullagh, for their time, effort, and thoughtful feedback during my viva. Their insightful comments and constructive criticism have been invaluable in refining my research.

I'm grateful to all those who stood by me during my time at Kent and beyond. This includes the researchers whose books and articles I have read. Although we may not have met in person, I believe that your work has undoubtedly added value to my life.

I am also thankful to my friends and colleagues, who have provided me with moral support, entertainment in down times and motivation when I needed it the most. I am particularly grateful to Alexandros Zenios, Giorgos Nicolaou, Nicolas Georgiades, Konstantinos Ashiotis, Aimilios Patsalides, Rafaellos Lympouras, Dr Marco Santopietro, Ethan Chung, Dr Rania Kolaghassi, Chantal Zorzi, Dr Paula Delgado-Santo, Dr Elhassan Mohamed, Dr Matthew Boakes, Dr Ayda

Majd Ardekani, Michael Liao, Dr Mayank Loonker, Dr Boyd Brunch and Dr Aymen Lakehal for their unwavering support and encouragement throughout this journey.

Lastly, I would like to express my deepest gratitude to my family, who have always supported me. I thank my parents, Petros and Anastasia, and my siblings, Neophytos and Evridiki, for giving me their ethos and pathos (passion) for whatever they do in their lives. I would also like to thank my significant other, Aine Maire Solomon, for always supporting me unconditionally. Their belief in me and constant support have been the pillars of my academic journey. Without them, I would not have been able to achieve this milestone.

Pavlos Nicolaou, April 2025.

Contents

Chapter 1

Introduction

1.1 Background

The current state-of-the-art in assistive living is based on a blend of smart devices, including wearable sensing technologies (smart watches and wristbands), smart home devices, for example, voice assistant virtual agents, IoT and mobile technologies [1]. However, real-world deployments with these technologies are limited. For instance, the elderly do not favour using wearable technologies because they can be cumbersome or forget to recharge them. Our research aims to explore a new paradigm in which people's daily activities are monitored passively through sensing technologies embedded within the environment without any form of wearable device or active user intervention. Instead, this research focuses on using the audio signal as a modality to detect human activity with data collected in a competent care home.

NHS started Virtual wards, also known as a hospital at home [2], allowing patients to get hospital-level care while at home. The motivation of virtual wards is to support patients safely and conveniently while they are home instead of visiting the hospital. In addition, there has been significant interest in smart healthcare homes in recent years [3]. For example, Smartphone-based devices and wearables have been successfully used to monitor and manage blood pressure, weight, body analysis, pulse rate, electrocardiograph, blood glucose, blood glucose saturation, sleeping, and physical activity. Additionally, with the use of cloud computing and Bluetooth technology, smartphone-based devices and applications can provide real-time analysis of user health [4].

Recently, there has been an increasing interest in voice assistants, such as Amazon Echo and Google Home, which millions of people have adopted [5]. These devices offer helpful services like conversational assistants, media players, and home automation with a caveat of the user's audio data. The smart home device can not only support the general population but also can be specified to support the elderly. MiiCare [6] is an example of a deployment of voice assistant devices in single-occupancy elderly homes. Meanwhile, real-world deployments of wearable technologies for the elderly have had limited success. A noninvasive approach would be more appropriate, using microphones and multi-modal sensors for passive audio sensing. MiiCare is in the process of understanding and setting up the mechanisms for receiving anonymous data from them.

Our research aims to build systems for tracking acoustic activity changes in elderly people in a privacy-preserving way. According to previous work, acoustic sensing can be achieved by passive always-on audio recording, which can pose privacy concerns, or by using short audio clips when the voice assistant is recording [7]. In a real world setting, such as a home environment, it is not always possible to collect curated data without the presence of cameras or human intervention. For this reason, a combination of public datasets and experimental data from a real deployment is being used. Additionally, two data collections will be conducted, utilising a data collection app that has been created to support the evaluation of our methodologies.

1.2 Motivation

As the number of elderly people continues to rise yearly [8], the need for practical solutions to assistive living becomes increasingly important. While wearable technologies have been implemented in the past, these solutions have yet to see widespread success due to the burden of daily management and charging. However, a passive approach to this problem offers a promising alternative for improving elder care and enhancing the quality of life for our ageing population.

How is it possible to apply a passive sensing approach in assisted living? A wearable has various embedded sensors, such as an accelerometer, which can detect movement. A smart device with a microphone sensor integrated into the room can also detect some activities using audio, which is an example of passive sensing. The home occupant can go through their day without the need to remove, charge or interact with the sensing device. Also, sound from the environment can be conceptually rich of information, which can be used in various sensing aspects. By expanding our understanding of this modality and its potential benefits, a more supportive environment for the elderly can be created.

1.2.1 Application Domain

The research in passive activity sensing focuses on the application domain of smart homes. Smart homes are equipped with various sensors and devices that enable residents to be assisted with automation and monitoring. By integrating passive sensing technologies within the smart home environment, this research aims to leverage existing infrastructure to monitor and track the activities of the habitats. Smart homes provide an ideal setting for this research as they offer a controlled environment where sensor data can be collected and analysed to better understand daily activities and health status.

The ideal home environment would be one where smart devices equipped with microphones are installed in every room. These devices could continuously record audio when a person is in the room and upload it to the server for further analysis. For people with dementia, one example of further analysis is to understand changes in their routine, which can lead to identifying a change in the severity of their health issue.

1.2.1.1 Challenges

Passive activity sensing in smart homes presents several challenges. Firstly, the need for labelled data poses a significant challenge for supervised acoustic sensing. Labelling audio files with ground truth information is often impractical and sometimes impossible, especially in real-world deployments requiring continuous monitoring. This challenge makes the development of unsupervised techniques and algorithms that can analyse and interpret audio data without the need for explicit labels.

Furthermore, privacy concerns are a significant challenge in deploying passive sensing technologies. Always-on audio recordings can capture private conversations and sensitive information, raising concerns among individuals, caregivers, and family members. Striking a balance between monitoring activities for health and well-being while preserving privacy becomes crucial in designing and implementing passive activity sensing systems.

Unlabelled Data

The availability of unlabelled data is a significant challenge in passive activity sensing for smart care homes. Different from controlled experimental setups or datasets generated from videos, collecting labelled data in real-world smart care home environments is often impractical or unfeasible. However, unlabelled data can still provide valuable insights into the acoustic activity patterns of residents. Therefore, this research strongly focuses on leveraging unsupervised learning techniques that can detect changes and patterns in activity using unlabelled audio data. By exploring approaches such as clustering, Dynamic Time Warping and Support Vector Machines with Transfer Learning, this research aims to extract meaningful information from unlabelled data and develop robust pattern recognition models for smart care homes.

Lack of labelling

More accurate and comprehensive labelling in smart home scenarios presents a significant challenge for supervised learning techniques. The nature of daily activities and the complex interactions within a smart home environment make manual annotation of audio files with ground truth labels difficult, if not impossible. This challenge poses a barrier to the application domain of traditional supervised machine learning algorithms that heavily rely on labelled data. However, even with a small amount of labelled data, transfer learning can be leveraged to address this limitation. Transfer learning allows us to retrain a pre-existing model, previously trained on a large dataset, with the limited labelled data available in the smart home scenario. The parameters of a larger, more general model can be fine-tuned, allowing knowledge learned from the larger dataset to be effectively transferred to the specific activity recognition task in the smart home environment. This approach enables the model to learn and generalise from the limited labelled data, capturing important patterns and features specific to the smart home context. By leveraging transfer learning and training a shallow model, this research aims to improve the performance and adaptability of activity recognition models in smart home scenarios with limited labelling resources.

1.2.1.2 Concerns

Audio recording will be essential in a smart home environment where audio is the main modality. Continuous audio recording could be used for passive acoustic sensing and audio analysis without the need for human intervention or the use of wearable devices. Having such a system could bring some privacy concerns. Based on previous experience in deploying acoustic sensing in care homes, recording private conversations was the main concern. Similar to carers visiting the smart home environment, they prefer not to be recorded, specifically when they have a private phone call or a conversation. The idea of constant conversation monitoring in their homes may make individuals feel uncomfortable or intruded upon. Addressing these concerns and implementing privacy-preserving measures, such as speaker anonymisation, is crucial to ensuring user acceptance and ethical considerations in assistive living systems.

In the context of smart care homes, where individuals with conditions like dementia reside, there may be concerns from family members and caregivers. It's important to be mindful of privacy concerns when family members or caregivers are present in the home, which may be recorded by passive sensing systems without their knowledge. To build trust and address concerns from caregivers and family members, it's crucial to have open communication about data collection and privacy policies.

1.2.1.3 Technology

Collecting data in smart home environments presents its own set of challenges. Due to the complexity and diversity of daily activities, obtaining labelled ground truth data for training and evaluation purposes is not easy or feasible. It is time-consuming and expensive to manually annotate and label audio files accurately and consistently. As a result, researchers often rely on alternative approaches, such as synthetic datasets or controlled data collection setups, to augment real-world data for training and testing acoustic sensing models.

1.2.1.4 Health Related Motivation

The motivation for passive activity sensing in smart care homes is particularly relevant for individuals with dementia. People with dementia often struggle with memory loss, confusion, and disorientation, making it difficult to carry out daily activities independently [9]. By monitoring their everyday activity patterns, changes to their daily routine or missed tasks can be distinguished. It has been observed that the elderly are resistant to using wearables, and when they are used, they may be forgotten to be charged. With the use of a passive acoustic sensing system in a home environment, where the management of any wearable device is not required by the elderly, activities can be detected, and changes to their sequence can be identified. Ultimately, this information could be used to develop personalised care plans and interventions that enhance quality of life and ensure timely assistance.

1.3 Research Questions

This PhD research aims to investigate novel approaches using privacy-aware acoustic sensing in smart care home environments, considering the challenges posed by the lack of labelled data and the privacy concerns associated with speech capture. The research questions of this study are as follows:

- **Q1: Is it feasible to identify changes in daily routine through unsupervised acoustic sensing with performance levels sufficient for detecting significant deviations, such as change of order of activities or missing activities?** It is highly beneficial to have a comprehensive understanding of the changes in the daily routine of an individual living with dementia. This knowledge can be instrumental in identifying and addressing the difficulties they may face in their daily lives. It can also facilitate the customisation of the support they receive, significantly improving their quality of life. Therefore, caregivers, family members, and healthcare professionals could be proactive in observing and documenting changes in the individual's routine to ensure that they receive the necessary support and care.

- **Q2: What are the implications of an acoustic system in a home environment that is designed to protect against misuse of speech data?** Deploying an acoustic system in a home environment with the aim to support the occupant through passive sensing comes with some cost. Privacy preservation is a big part of any new system that can record sensitive data in a home environment, and these need to be addressed.

- **Q3: What is the impact on the performance of acoustic activity recognition while providing speech-related privacy in continuous acoustic sensing?** Passive sensing in an acoustic environment benefits from understanding the context of the environment, but when it comes to privacy, speech is something that needs to be preserved.

- **Q4: Is it feasible to apply speech-related privacy in continuous acoustic sensing while maintaining a drop in acoustic activity recognition accuracy of no more than 10%?** Understanding the challenges of preserving privacy in an environment where audio is the main modality. Valuable information may be lost while speech is being removed, and methods can be employed to minimise this. The preservation of speech does not always involve the destruction and scraping of audio data; rather, the parts of the audio that are prohibited can be removed.

1.4 Contributions

1.4.1 Major Contributions

- **Tracking activity changes through an acoustic environment.** As per the first contribution to our research, the challenge of understanding the variations of a normal sequence of activities was tackled through acoustic sensing. In people with dementia, an understanding of a consistent daily routine is considered crucial. With a health condition involving cognitive disability or deterioration, the simple tasks that are done on a daily basis can be easily forgotten. It is recommended that the routine of the person living with dementia be monitored to ensure that healthy daily habits are not forgotten.

- **Novel dimensionality reduction algorithm that maintains high levels of accuracy for tracking activities through acoustic sensing.** A novel dimensionality reduction algorithm was invented to track changes in daily routine, utilising features extracted from the audio data and a commonly used pre-trained model. The number of dimensions was reduced from

128 to 3, and high levels of accuracy in detecting changes in audio-recorded activities occurring in various sequences were successfully achieved.

- **Analysing unique challenges that occur in a smart home environment where acoustic sensing is used as the sole modality for human sensing.** Some acoustic sensing challenges in smart home environments change based on the material of the building, e.g. carpet or wooden floors, how many people live together or if it's in a quiet or loud neighbourhood, but some specific challenges are the same everywhere. The main challenge that was looked into is privacy regarding speech in a single occupant household, where care is provided to the elderly by visiting carers, introducing some unique challenges.
- **Developed a system that removes speech audio content from an acoustic environment based on limited data while keeping the background sounds intact.** The last part of our research is developing a system that can detect and remove speech from home environment sounds, where collecting data is not always possible while keeping the acoustic activity recogniser accuracy high.

1.4.2 Minor Contributions

- **Designed and developed an App for iOS and Android that is used for remote data collection of labelled acoustic data.** Developed a proposed app for our data collection on both iOS and Android platforms.
- **Finding Public Datasets relevant to indoor acoustic sensing.** The acoustic sensing field is popular, but when it comes to a specific task in single-occupant households, the data is available or limited. Various public datasets, along with data from a real deployment, could be accessed by us.
- **Achieved two data collections remotely during the COVID lockdown** During the COVID lockdown, our plans were changed because everyone was forced to stay home, and data collections were conducted remotely.

- **Created a methodology of masking chunks of audio containing speech for privacy preservation of the person recording in an acoustic environment.** A methodology was created to mask audio chunks containing speech content as the first action on preserving privacy, to understand its effect on acoustic activity classification.

- **Developed an acoustic activity recognition methodology focused on acoustic environments where collecting labelled data is not possible.** Acoustic activity recognition state of the art has advanced massively in recent years. The latest methodologies using deep learning need massive amounts of data, but in our real environment scenario, collecting labelled data is not always possible. A variation of acoustic activity classification was created for this reason to be used on a small amount of data from the home environment.

- **Found that masking chunks of audio containing speech is not the best solution for an acoustic environment that needs to use audio for further analysis.** The baseline acoustic activity recogniser was not performing well after the speech sections were removed from the audio signal.

- **Synthesised our purposed datasets with various speech volumes using public datasets and data from a real deployment.** The synthesis of the dataset was done to develop our speech removal model and to fulfil the needs of our methodology.

- **Used a voice activity recognition model with the synthesised datasets to understand the impact of speech before and after it is added to the background audio signal.** Analysed the impact of speech using a public voice activity detection model on our various datasets before and after introducing speech data.

1.5 Publications

- **Related Publications**

  - Chapter 4: Tracking daily routines of elderly users through acoustic sensing: An unsupervised learning approach [10]

- **Preparation of Related Publications**

  - A combination of Chapters 4: Privacy Challenges for Acoustic Activity Recognition and Chapter 5: Privacy Preservation with Speech Removal is prepared for publication as a journal.

Chapter 2

Literature Review

2.1 Elderly Living at Home and Health Context

2.1.1 Introduction

The past few years have witnessed a remarkable rise in life expectancy, largely attributed to the rapid advancements in medical science. As a result, there has been a substantial increase in the number of individuals aged 65 and above [11]. Consequently, an increasing proportion of elderly people prefer to remain within their homes and communities; this phenomenon is commonly called "Aging in Place" [12]. The NHS has initiated "Virtual Wards", also referred to as hospital at home [2], which provide patients with hospital-level care from the comfort of their own homes. The objective of the "Virtual Wards" is to offer patients safe and convenient support at home instead of requiring them to go to the hospital. In addition to home caring, leveraging smart home technologies can play a pivotal role in significantly enhancing the quality of life for these individuals by promoting independent living and facilitating regular communication with family members and caregivers [13].

The integration of software and hardware components in smart homes enables the collection and analysis of data, allowing for continuous monitoring of various aspects such as health status, behaviour patterns, and overall quality of life among elderly individuals. This capability helps mitigate potential risks and facilitates communication between elderly, their relatives, caregivers, and medical professionals [11].

Implementing smart home technologies can bring a range of practical benefits for addressing the individual needs of ageing populations. These technologies offer a means of preventing risky situations, for example, fall detection and social isolation, and they also hold considerable importance for healthcare as a whole [14, 15]. With the potential to enhance overall well-being and improve quality of life, these technologies represent a promising solution for individuals with disabilities and those in all age groups.

In the context of ageing populations, smart home technology can significantly support elderly individuals who choose to live independently [16]. Depending on the elderly person's health status, they may require additional assistance and caregiver support over time. Smart home technologies can aid in the detection and analysis of health events, ensuring that appropriate medical assistance and caregiver support are provided to frail elderly individuals and those with disabilities or health issues who may not be able to live independently for extended periods of time. Moreover, smart homes can also assist in daily tasks such as medication reminders, scheduling doctor appointments, and even assisting in household chores such as turning off the oven or cutting off the water, promoting a safer and more comfortable living environment for elderly individuals. Aside from ensuring the health and safety of elderly individuals, smart home technologies also contribute to their independence and overall well-being. The ageing population and the increasing number of elderly individuals living alone necessitate implementing smart home technology to meet their specific needs [17].

2.1.2 Living at Home

Research has shown that most older adults prefer to age in the comfort of their homes and remain there for as long as possible [18]. This phenomenon, known as "Ageing in Place," is also called Ageing at Home [19–21], living with autonomy and independence accurately summarises the key elements of healthy, active, and successful ageing [22].

Considering this, ageing is not only about living longer but also living better with a higher quality of life. An example is "More Years, Better Lives", a European programme promoting research on the Potential and Challenges of Demographic Change [23].

It is crucial to create solutions that facilitate their lifestyle while prioritising their safety and overall health. An elderly individual suffering from any health condition might need a healthcare specialist to stay with or visit them regularly.

2.1.2.1 Healthcare Specialists

Healthcare specialists are crucial in providing medical care and support to the elderly living at home. Their main job is to assess the health status of the elderly, create personalised care plans, and monitor their progress. They are also there to support the elderly with daily tasks, administer medications, and provide medical treatments as needed. However, due to the physiological and cognitive dysfunction and reduced physical flexibility of the elderly, those who live alone face many challenges [24]. Some studies have shown that the incidence of accidental injuries and fatalities among older adults living alone at home is high [25], highlighting the importance of ensuring their safety and well-being in their home environment.

There are generally two main types of caregivers for the elderly, those who provide care within the individual’s home and those who visit periodically as a caregiver. The caregiver is expected to be able to monitor the health and well-being of the elderly individual and respond accordingly when necessary. In addition to providing care, the caregiver is responsible for monitoring the senior's medication, bathroom habits, and overall health status. To ensure a happy co-existence between the elderly and their caregiver, it is important to establish a healthy and friendly relationship.

Understandably, if the caregiver visits periodically or a family member may not be available to be with the elderly all day, every day of the week. This is where smart systems can be incredibly useful for the caregiver. These systems can help the caregiver and family members remotely monitor the daily activities, health status, and overall well-being of the elderly individual.

2.1.3 Daily Routine

Understanding individuals' daily routines is crucial to providing tailored support and timely assistance. It is important to acknowledge that the daily routines of older adults often differ drastically from those of younger people due to various factors, such as obligations, needs, cultural practices, and personal interests. Any part of people's daily routine might differ drastically based on their age and obligations.

The routine can also differ on the individual level based on the time of year or changes in lifestyle. Some daily routines will probably stay consistent throughout the year, such as brushing teeth in the morning or having a shower before bedtime. Understanding which parts are consistent and which parts are random during their routine is key for tracking routines with the aim of health monitoring.

2.1.3.1 Daily Routines of the Elderly

Regarding elderly individuals living independently, constant and random activities can be observed in their routines. It is acknowledged that certain parts of the daily routine are very random and are not worth tracking. On the other hand, parts of the daily routine can be very consistent [26], such as the morning or evening routine. These constant routines are usually carried out before or after bedtime when preparations for the day are made or when going to bed. Taking prescribed medication, preparing breakfast, and carrying out various personal hygiene tasks are typically involved in these routines.

2.1.3.2 Tracking Daily Routines for Health Monitoring

By identifying the regular aspects of an elderly person's daily routine, deviations from those patterns can be detected through continuous acoustic sensing. Valuable insights into cognitive decline can be provided by integrating health monitoring through continuous sensing in a smart home setting [27], making it a helpful asset for elders who live independently.

2.1.4 Dementia

There are currently 47 million individuals worldwide living with dementia, and it is projected that by 2050, this number will rise to 131 million [28]. Dementia is a pathological condition that disrupts the brain's normal functioning, leading to a gradual and sustained decline in cognitive abilities [29]. This decline manifests in various ways, including memory loss, language difficulties, attention deficits, orientation challenges, impaired judgement, and planning deficits [30]. Dementia is a complex and progressive neurodegenerative disease that requires specialised medical attention and care.

Dementia is a degenerative condition of the brain that can result from various factors, including brain disease, injury, or medical conditions originating from other body parts [29]. It should be noted that some diseases leading to dementia begin in the brain itself, such as Alzheimer's disease, which affects around 5 million individuals in the United States [28]. Moreover, the risk of developing Alzheimer's disease increases with age. Typically, dementia arises when the brain is affected by two or more common age-related diseases, such as Alzheimer's and stroke.

As age is experienced, the possibility of developing dementia becomes more common, although it may not be the same for everyone. The beginning of dementia usually commences in the seventh or eighth decade of our life and progressively deteriorates over time [28]. The first sign of dementia is often a decline in short-term memory, which the person affected may notice, as well as their family and friends. This might result in losing track of where the elderly have put things around the house or forgetting recent conversations. In dementia, these issues occur often, maybe multiple times a day, and for many months or years, which can harm our daily activities. Additional symptoms may include withdrawing from hobbies or social gatherings, anxiety or depression, difficulty performing routine tasks such as remembering to take medications, and difficulty with directions, such as when driving [30].

It's important to have a system in place to monitor any changes in daily routines for individuals living with dementia. These changes can indicate a decline in cognitive abilities and help ensure the appropriate care is provided. Since behavioural changes may occur gradually, consistently tracking and monitoring the condition's progression over time is crucial.

2.2 Smart Solutions for the Elderly

2.2.1 Smartphones

Modern smartphones are equipped with a wide range of sensors that enable them to measure various aspects of health and activity. These sensors include high-resolution cameras, GPS, accelerometers, gyroscopes, magnetometers, ambient light sensors, and microphones. Motion sensors such as the accelerometer and gyroscope can identify daily activities [31] and any falls that may occur [32]. Researchers have developed algorithms and models using smartphone sensors to accurately recognise walking, climbing stairs, sitting, and standing [33]. This allows for continuous monitoring and assessment of physical activity levels, which is important for maintaining overall health and detecting any changes or declines in mobility [34]. Overall, the various sensors present in smartphones make them valuable tools for tracking and monitoring different aspects of health and well-being.

According to reports, elderly individuals aged 65 or above are prone to experiencing falls within the confines of their own homes. More specifically, it has been observed that around 65% of women and 44% of men tend to fall within their usual place of residence, while around 25% of men and 11% of women have experienced falls in their gardens [35]. Such statistics highlight the importance of ensuring homes are safe and secure environments for the elderly. The fall detection problem is addressed by smartphones, as they are equipped with an accelerometer, gyroscope sensor, and cameras, and communication can be done through WiFi or cellular network at the moment a fall is detected.

Some studies concentrate on the elderly population, aiming to spot instances of falls during their daily activities [36]. The researchers used the smartphone accelerometer data and developed two algorithms to detect falls. The first one used a threshold method, similar to the previous example. The second algorithm utilised a kNN classifier was used to identify falls.

The features extracted to train the classifier are:

- Average Absolute Acceleration Variation (AAMV)
- Impact Duration (IDI)
- Highest Value of Acceleration (MPI)
- Peak Duration (PDI)
- Activity Level of a Window that Contains the Impact (ARI)
- Average Acceleration of free-fall Stage (FFI)
- Number of Steps (SCI)

- Skewness of acceleration Segment (SKEW)
- Kurtosis of Acceleration Segment (KURT)
- Interquartile Range of Acceleration Segment (IQR)
- Power of the Impact (POWER_IMP)
- Standard Variation of the Impact (STD_IMP)
- Square of the highest coefficient of Continuous Wavelet Transform (CWENERGY)
- Number of Peaks in the Continuous Wavelet Transform (CWPEAKS)

All features listed above are in the time domain except the CWENERGY and CWPEAKS, which are in the frequency domain. The kNN classifier with $k = 7$ was trained with 10-fold cross-validation of the MobiFall dataset [37]. The MobiFall dataset was collected using a smartphone and includes trials from 57 subjects. It covers four types of falls and nine different ADLs, totalling more than 2500 trials. After conducting experiments, the authors found that training a kNN classifier was more effective than using a threshold model.

2.2.2 Wearables

Wearable devices like smartwatches, wristbands and glasses can collect and upload physiological data throughout the day. Smartwatches can be worn on the wrist during the day and night, and they can collect continuous sensory data. Wearables provide opportunities to improve quality of life and present information to the user, like text notifications and urgent information, more conveniently than a smartphone [38]. Wearables became popular because of their convenience, health features, and activity tracking. Some wearables, for example, smartwatches, can also provide sleep-tracking capabilities because of the possibility of wearing them overnight [39].

Another very interesting and important capability of wearables is fall detection. In this review [40], the authors mentioned as a possible lifesaver. There is a lot of potential in using technology to prevent falls and protect against their negative consequences. However, currently, there is no satisfactory solution, even if cost is not a factor. The survey [40] concluded that the main problem is that it is difficult to understand the difference between actual falls and activities of daily living with the data and methodologies available up until now. The authors of the survey also mentioned that there is a need for further research to develop a large and comprehensive real-world database of falls that can be used to train advanced machine-learning algorithms.

2.2.2.1 Smartwatches

Most activity recognition algorithms recommended are adopting inertial sensors in smartphone devices or other bodily worn sensors [41]. On one hand, smartphones are not easily worn or placed on the human body other than a pocket or customised body-worn mounts. On the other hand, other wearables, like smartwatches, equipped with sensors can be easily worn on the body during the day and night, but they don't have a big screen to interact with as smartphones have. In the meantime, smartwatches are getting more advanced year by year, with the particular opportunity to give highly accurate activity recognition with great user interaction.

Embracing deep learning for activity recognition by smartwatches is an understudied problem [42]. The author used one of the most prominent deep learning methods, Restricted Boltzmann Machines (RBM), that had never been used to tackle activity recognition problems on smartwatches. They mentioned that commercial smartphone systems from Baidu, Amazon, and Google use deep learning remotely on the cloud and then send the results back to the device. They also showed that full RBM-based activity recognition can be used on state-of-the-art smartwatch hardware.

In another study [43], smartphone-based and smartwatch activity recognition was compared. The author demonstrated that smartwatch-based activity recognition can identify activities that the smartphone cannot, such as eating.

To achieve fall detection on wearable devices, like a smartwatch, the accelerometer sensor on the device is used alongside a threshold or machine-learning based algorithm. For the threshold-based approach, a point is set based on the wrist's movement, where it is possible to differentiate a daily activity movement under the threshold from a sudden fall movement above the threshold [44]. The machine learning approach will need to have some data collected from a daily routine and label them as non-fall and some fall activities that will be called falls. The traditional process of training a machine learning model will be needed, and the newly trained model should be able to identify if a new fall will be positive or negative.

2.2.2.2 Earables

Earables are also part of the wearable family. Earables are wireless earbuds that can play music and are used for phone calls. They are also equipped with sensors that detect movements, such as accelerometers and gyroscope sensors. Earables such as eSense [45] provide an in-ear multi-sensory stereo device that could help our understanding of a range of human activities in a non-intrusive manner. Earables work non-intrusively and can bring opportunities for new and effective applications in providing solutions for physical and mental well-being and cognitive assistance [46].

2.2.2.3 Active Noise Cancelling Headphones

In the earable category, headphones have recently advanced in isolating the audio signal playing from the headphones from the background sounds. Headphones come in different sizes and shapes. In-ear headphones usually benefit from the noise isolation effect because they block any sounds from the environment because of their shape where they are shaped to fit in the ear canal. Another technique used is

Active Noise Cancelling (ANC) in headphones [47]. Depending on the environment in which one uses the headphones, the ANC might be needed, especially in loud environments where the background sounds overtake the audio signal from the headphones.

ANC headphones work by having a microphone on the outside of the headphones to record environmental sounds [48]. In a real-time sound synthesiser, the ANC headphones can synthesise the audio signal from the audio source with an opposite signal from the environmental sounds to mute the audio from the environment. This results in an enhanced favoured audio signal without any background sound from the environment. The ANC technology has become popular in recent years in consumer electronics because smartphone users listen to music, podcasts or audiobooks in usually crowded areas while commuting, like trains or open office space areas [49].

Applications

According to Choudhury R. [50], the realm of earable computing is an up-and-coming area that deserves our attention and consideration. An earable platform already sits well in society, where humans have been using headphones to listen to music and podcasts or take phone calls for more than a decade now, and the popularity of wireless earbuds keeps growing [51]. Choudhury R. presented the challenges of earphones, such as energy, discomfort, privacy, and health, but they also emphasised in their paper some research problems that earables can solve [50]. Some research problems are motion and activity tracking, such as breathing problem detection or in-mouth motion detection. Earables hold enormous potential in applied research where other smart devices like smartphones or smartwatches will not be suitable for [52].

2.2.3 Smart Homes

Smart Homes were developed when IoT, the Internet of Things, started to gain popularity. Smart home technology allows for the integration of various electronic devices and appliances to create a network ecosystem within a house. All electronic devices are connected to the internet and equipped with a small computer or a microcontroller, so they can communicate with each other and be controlled remotely by the homeowners. In recent years, smart home technology has gained traction due to its potential benefits in improving convenience, energy efficiency, and security for homeowners [53]. A growing interest has recently been in utilising smart home technology for health monitoring purposes [11].

Smart Home systems can collect data using multiple sensors and use them to train predictive models for various health and comfort reasons. For example, Google Nest leverages artificial intelligence to dynamically adjust home temperature by considering occupants' presence or absence at different times. Using AI technology, Google Nest aims to optimise energy usage and enhance overall household comfort levels based on real-time occupancy data [54].

2.2.3.1 Sensors in Smart Home Environments

In smart homes, it is possible to have the space and the energy capacity to accommodate many sensors. There are different sensors for different applications, as seen in several assistive smart home projects identified in [55]. These smart home projects are trying to bring comfort and automation for the elder resident, along with cognitive and physical health checks, in a non-invasive manner. Table 2.1 presents the universally used sensors according to [55].

Smart homes that use home-based approaches are more efficient in using sensing devices compared to video-based and wearable device-based approaches. Smart home sensors are noninvasive, meaning they do not disrupt or interfere with people's normal lives while monitoring various aspects effectively.

Table 2.1: Ambient Sensors Used In Smart Environments

Sensor	Measurement	Data Format
Passive Infrared Motion Sensor	Motion	Categorical
Active Infrared	Motion/Identification	Categorical
Radio Frequency Identification	Object Information	Categorical
Pressure	Pressure on Mat, Chair, etc	Numeric
Smart Tiles	Pressure on Floor	Numeric
Magnetic Switches	Door or Cabinet, Opening or Closing	Categorical
Ultrasonic	Motion	Numeric
Camera	Activity	Image
Microphone	Activity	Sound

Unlike wearable devices that individuals must wear on their bodies/wrists or video surveillance systems that invade privacy, home-based approaches provide a more seamless and non-disruptive solution for continuous monitoring within a smart home environment [56].

2.2.3.2 Activity Recognition

Applying activity recognition within smart homes is very important as it paves the way for personalised and context-aware smart living. By accurately identifying different types of activities in a household environment, smart home systems can understand what is happening inside the home environment and what activities are taking place. Monitoring human activities can provide valuable insights into individuals' health, well-being, and daily routines in smart homes [57].

Activity recognition in smart homes is a vital component of assistive living, as it allows for the monitoring and analysing of residents' daily activities. Monitoring and analysis can provide valuable insights into individuals' health, well-being, and daily routines in smart homes [57].

Barger et al. developed a distributed passive infrared sensor system that collects data on a person's activities within the home environment [56]. This system enables the detection of movements and thermal signatures, which can be used to determine and analyse various activities such as sleeping, eating, and moving around the house. A novel approach is proposed for home-based activity recognition, which involves using a radial basis function neural network (RBFNN) in conjunction with a localised stochastic-sensitive autoencoder (LiSSA) method. The main aim of this technique is to accurately detect and classify different activities being performed within a home environment. An autoencoder (AE) was proposed to extract useful features from the binary sensory data by converting sensory input data into continuous inputs to extract improved levels of hidden information.

The proposed method was assessed using four binary home-based activity recognition datasets: OrdonezA [58], OrdonezB[58], Ulster [59], and activities of daily living data from van Kasteren (vanKasterenADL) [60]. The results of this evaluation showed that the proposed method achieved the best performance, with accuracy rates of 98.35%, 86.26%, 96.31%, and 92.31% on the four datasets, respectively.

2.2.3.3 Vital Measurements

In addition to activity recognition, smart homes can collect vital measurements for the elderly. These measurements include health-related data such as heart rate, blood pressure, and body temperature. This data can be collected using sophisticated instruments, as mentioned in [55]. Creating a consistent schedule in which senior citizens can consistently keep track of their vital signs is essential in guaranteeing their safety and overall health.

2.2.3.4 Assistive Living

Smartwatches have the ability to detect falls, providing peace of mind in assisted living scenarios [61]. The biggest issue facing fall detection systems and other assistive living systems is the need for accurate and reliable detection without generating false alarms, as these can lead to unnecessary anxiety and distress for both the individual and caregivers. Also, for the elderly population, especially if they suffer from dementia or other cognitive impairments, it may be difficult for them to manage their smartwatch, charge it and put it back on when it's fully charged. A multi-modal solution in smart care homes incorporating multiple sensors and technologies, such as wearable devices, floor sensors, and video monitoring, can provide more accurate and reliable human activity detection capabilities [62]. These sensors can detect sudden changes in movement or pressure and trigger an alert to caregivers or emergency services.

2.2.3.5 Daily Routine Recognition

Smart home systems have developed rapidly, thanks to advances in current sensor technologies, and they can help elderly people live safely and independently at home. Most studies done for routine daily recognition have manually labelled activities and trained supervised models to identify daily routines. In this study [63], the authors explored the possibility of using unsupervised learning methods. For this deployment, the authors used:

- Motion Sensor
- Humidity/Temperature Sensor
- Accelerometer Sensor installed on the bed
- Power sensors are plugged into power outlets
- Reed switches attached to doors/windows
- Circuit meter sensors installed in the meter box (for monitor energy usage)

Using the Markov chain to model a resident's motion patterns at different times of the day and identify clusters of daily routines at a large scale. Afterwards, for every cluster, the model looks deeper into the room-level discovery of activities and routines so the author can understand different indicators of functional decline by the elderly residents and suggest the appropriate mediation.

2.3 Challenges of Current Methods

2.3.1 Wearable Challenges

There are many benefits for the elderly to adopt new technologies because these technologies can help them live more independently in their own homes [64]. These new technologies can also improve the elders' health status; for example, the elderly can track their steps and hit their health goals daily. These wearable technologies can also track their heart rate history or electrocardiogram (ECG) and share this information with their doctors or family for heart-related issues.

From the wearable category, smartwatches were the most advanced [65]. The main reason smartwatches are superior is because they are equipped with an LED display and an array of sensors, including an accelerometer, gyroscope, proximity sensor, and GPS. These smartwatches are widely available for purchase and offer a wealth of capabilities to help users track their activity levels, monitor their heart rate and use them for navigation.

Eurostat, the statistical office of the European Union, has reported that a staggering 87% of citizens aged 75 years and above in the EU have not yet utilised the internet. This highlights a significant digital divide in the region that needs to be addressed to ensure equal access and opportunities for all age groups [66]. For many elderly individuals, the concept of using wearable technology can be difficult to accept and adjust to [67]. The idea of having to wear a device on their person may seem intrusive or unnecessary to some, and it can be a challenge to convince them otherwise. Additionally, there may be concerns about the practicality or effectiveness of these gadgets that need to be addressed.

Apple invented the first iPhone in 2007 [68], which greatly contributes to how people communicate and function daily. The older population lived most of their lives without a smartphone connected to their smartwatch and smart home gadgets. It seems like the people that didn't grow up around these technologies are more reluctant to use them because during their whole life, they didn't have them and they didn't need them [67]. Based on their experiences, their life with no gadgets was fine.

Recent studies have investigated the adoption of wearable technologies, particularly smart watches represent the most popular type of wearable devices, as their popularity has grown during the last few years. Chuah et al. explored the factors influencing adoption, including the role of usefulness and visibility [69]. Users of smartwatches are divided into two categories: those who buy for fashion and those who buy for functionality. This creates a new term, 'fashnology', which is a combination of fashion and technology. On the negative side, it has been observed that smartwatches are often designed to cater to the general population, thus neglecting the elderly population [70].

As the elderly need support, either from a carer or technology, there is a gap between the technology that can be very useful and how the elderly can use or cannot use the technology. Wearables especially need extra care, in terms of charging and ensuring it's up to date and connected with a smartphone or the internet. All these inconveniences can accumulate and add stress to the elderly with health and mobility issues rather than providing any health benefits. It's important to consider the impact of these inconveniences and find ways to mitigate them in order to improve the overall well-being of elderly individuals.

One of the primary concerns in assistive living with wearable devices for older adults is their inconvenience. This issue poses a challenge within the application domain, as wearable device usage is not well-received by this demographic [67]. Thus, a further investigation with alternative methods of assistive living was needed, with the purpose of being unobtrusive and not relying on wearable devices. Wearables brought a new dimension to the research and application domain in collecting physiological data that was not possible before. Smartwatches have proven more effective in gathering data continuously throughout the day and night than smartphones. The compact size and convenience of the watch allow for continuous monitoring of various metrics, such as heart rate, steps taken, and sleep patterns, without the need to constantly hold the device in hand. Meanwhile, smartphones may not always be available or easily accessible, making it difficult to track and collect data consistently.

As per observations, it has been noted that the physical and mental health of the elderly population is highly diverse. While some individuals seem to be able to carry out their daily activities with ease and generally feel good, others may be grappling with debilitating conditions such as dementia that hinder their ability to lead a normal life. As researchers, caution needs to be exercised regarding the technology that is being forced on a vulnerable population that struggles to navigate their daily lives.. The solution can also be individualised technology based on their technical knowledge, ability to remember to do daily tasks and mobility function. If they are willing to use a new technology and how they feel after using it for some time, they can be willing but not interested in using it daily because they didn't want or can form a new habit around that technology.

It is important to be cautious when introducing wearable technology to individuals with varying levels of dementia. Due to forgetfulness or skipping daily activities, such as brushing their teeth, these individuals may struggle to remember to charge their smartwatch or other wearable devices. This can prove the device useless and only serve as a cosmetic accessory rather than a functional tool.

2.3.2 Labelling Challenges

In a home environment of a single occupant, activities could be detected through multi-sensory data, wearable sensors or microphones. Having a pipeline for activity detection, where data is collected and the appropriate methodology is prepared to accurately detect the activities of home occupants, is vital. Depending on the application domain, different kinds of data will be needed. For example, continuous accelerometer data from a wearable device with timestamps and labels indicating whether a fall occurs or not is required in fall detection. This issue is classified as binary classification, and true or false labels are necessary to train a specific machine learning model to identify when a fall happens. In a smart home environment, an activity classifier is needed to interpret the participant's behaviour inside the house. A supervised classifier must be trained based on the participant's activities to determine which activities occur within the home. Data from multiple sensors must be collected and labelled for this application. Great results can be achieved with multi-modal classification, but labels are essential.

Labelling or annotating is a method to distinguish what is ground truth for a specific activity or event. Ground truth is the fundamental truth, distinguished by an observer or the participant doing the action or event. Labelling can be done during the data collection as the participant or an observer will set the labels and start and stop times of each activity or event, but it can also be done remotely if cameras are installed, and an observer can go through the videos and label based on what is captured on the cameras. In the context of smart care homes, cameras can be installed as part of the system. These cameras are utilised to gather labelled data through the process of annotating specific activities that take place within the home. An annotator will go through the video footage and label it after the data collection is done [71].

It is acknowledged that manual labelling is expensive and laborious for both the participant and the observer. This can lead to frustration and discomfort, especially if a lot of data needs to be collected. On the other hand, while cameras can be useful for smart home labelling, they can also be obtrusive and invasive. In some cases, house occupants may prefer alternative smart solutions that are less noticeable and intrusive. It's important to consider each occupant's needs and preferences when determining the best solution for their home.

With the new advancements in deep learning, deeper and more sophisticated models are created, but with disadvantages. These deep learning models will be much more accurate than traditional models in many problems, but more data and computing power are needed. That means more time labelling the data manually and a lot more expensive computing tasks. This poses a challenge in activity tracking for assistive living, and new solutions must be invented.

Using multi-modal methodologies for assistive sensing can be intrusive and computationally expensive. Working with multiple sensors for one task can also be overly complicated.

2.3.3 Limited Data

Labelled data are needed to train a supervised machine learning model [72]. The process of labelling data can present challenges or prove to be impossible in certain scenarios. One such example is in smart home settings where occupants may not feel comfortable being followed by an observer that notes down their daily activities or does not accept the installation of cameras due to privacy concerns. This reluctance can make it difficult to collect the necessary data for analysis. Additionally, elderly individuals may be hesitant to wear devices that track their movements, and even if they do, there is a chance that they may forget to recharge the device, leading to gaps in the data. Due to these challenges, alternative solutions must be explored to ensure accurate data collection.

In order to achieve a fully functional activity recognition smart home, a feasible approach is to install a network of multi-sensors throughout the house. This enables us to accurately identify the location of individuals within the home, providing insight into their movements and activity patterns. By utilising a humidity sensor, for example, it is possible to detect when an individual is taking a shower, while fobs can be implemented to track visitors' comings and goings. Additionally, audio sensors can be utilised to further enhance the understanding of the environment, distinguishing between activities such as walking, watching television, or talking on the phone [63, 73–75].

Having access to a large dataset is critical for creating effective machine-learning models for activity recognition. The better the training of specialised models for each participant based on their unique activity patterns, the more available data there is. However, building a system that is capable of collecting and processing an extensive amount of data is a time-consuming and expensive endeavour that is not feasible for the average researcher.

As a result, the focus is on finding a solution that is both cost-effective and efficient, utilising limited data sources to achieve accurate activity recognition. This will require careful consideration and planning to ensure that the system is designed to capture the most relevant data in order to provide accurate insights into an individual's daily activity patterns.

Audio sensing has proven to be an effective way of extracting contextual information from a home environment. By analysing audio signals, it is possible to determine whether someone is present in the house, whether they are walking around, cooking, or watching TV [76]. The value of audio sensing lies in its ability to serve as a single modality solution for a wide range of applications. Nonetheless, the challenge with limited data is how to build robust systems without compromising accuracy. A common approach to address the issue of limited data is to leverage transfer learning to utilise existing general pre-trained models using related data to train new models, thereby minimising the need for extensive data collection efforts [77]. Transfer learning holds immense potential for developing accurate and successful audio sensing systems, creating a more efficient and reliable way of gathering contextual information in a home environment.

Utilising a generalised pre-trained model, such as VGGish, a deep learning model pre-trained on a large-scale audio dataset for sound recognition, and a small amount of labelled data specific to the activity tracking in the assistive living domain, researchers can use transfer learning methodologies to fine-tune the model and achieve good performance [78].

2.4 Audio Sensing

2.4.1 Types of Audio Sensing

It is possible to extract some contextual information from inside the house environment using audio sensing. Any change in the occupant's daily routine can also be understood with audio [10]. It is essential to identify the available audio-sensing methods and their potential applications in developing better solutions for elderly people living at home. Acoustic Scene Classification and Acoustic Event Detection are two types of audio sensing. Acoustic scene classification identifies the context of a particular environment, while Acoustic Event Detection detects small temporal events within an audio stream [79]. Combining multiple audio events, like "cupboards open/close", "kettle", and "washing dishes", can form an audio scene, which can be called "kitchen" for this example.

2.4.1.1 Acoustic Scene Classification

Recognising various acoustic environments from recorded audio signals is an active research field that has received significant attention. Acoustic Scene Classification (ASC) is the task of recognising the acoustic environment based on the recorded acoustic signal, for example, "office" or "park" [80]. Recent results demonstrate that the state-of-the-art acoustic scene classification can outperform humans on the same task [81]. ASC algorithms have matured and are already in real-world application scenarios. However, one of the main challenges in this domain is the difficulty of collecting well-curated datasets. One of the most popular methodologies to apply ASC in new environments is using a pre-trained model [82].

Acoustic Scene Classification can be beneficial in a smart home environment focused on the elderly, as the system will be able to detect the environmental sounds of each room and how they change. A great example is when the elderly have visitors, the environmental sounds change. Another example of environmental sounds can change when people are in a room with a TV playing in the background or when people are having dinner and talking to each other. Understanding the environmental sounds of the smart home where the elderly live is crucial, as it is possible to extract a lot of information based on their habits or unfortunate incidents. ASC can also be combined with Acoustic Event Classification for a more detailed understanding of the environment and the specific events that are happening. The current state-of-the-art of ASC relies on deep learning, transformer models or a fusion of both [83]. The deep learning models that are commonly used are Convolutional Neural Networks (CNN), Long Short-Term Memory Networks (LSTM), and more recently, attention mechanisms have been applied to transformer models. Those techniques will be described in the Deep Learning Applications section.

2.4.1.2 Acoustic Event Detection and Classification

Acoustic Event Classification (AEC) is defined as the recognition of individual sound events in an audio signal, for example, "walking", "boiling water" or "typing on keyboard". AEC and detection is an active field, with a majority of work focusing on improving the overall performance of the models [84–88]. Very popular solutions of AEC are common in surveillance, including security, healthcare and wildlife monitoring [79].

According to [89], Acoustic Event Detection (AED) is a harder task than Acoustic Event Classification (AEC), and there are three popular solutions for the AED problem. The popular solution will be to do an AED at a continuous long audio signal, and when the AED captures a timestamp of an event of interest, the AEC will run for classification of that specific window. As stated by [89], the first popular solution is not suggesting using an AED but running an AEC on a sliding window on the continuous audio signal. The second solution suggests using the Automatic Speech Recognition (ASR) framework, which is popular in Text to Speech or other Speech-related problems. ASR uses Gaussian Mixture Models followed by Hidden Markov Models to model frame-wise features and their distributions [90–92]. The third solution suggests a Regression approach, where a random regression forest needs to be trained on labelled data with specific acoustic event categories. The authors from [89], proposed their own solution where the AED will happen with a verification step before moving to the AEC. The verification step will be a trained AEC that confirms this activity of interest in the continuous audio signal before moving to the AEC. This solution seems to have better overall results than the combined solution of AED and AEC.

An elderly smart home system that detects real-time events can greatly benefit the habitat. One great example is falling among the elderly, which is a quite common incident. According to the World Health Organisation, the highest number of fatal falls occurs in individuals who are 60 years old or older [93]. Detecting falls in an elderly home environment and getting immediate support can be crucial and a robust smart home system with AED and AEC can be useful. According to [94], using audio, the system could detect falls with 98.72% accuracy, resulting in the carer or a family member being notified in case of a fall.

2.4.2 Traditional Methods of Machine Learning

2.4.2.1 Introduction

Machine learning (ML) is a problem-solving method by machines that underwent some training/learning by human programmers [95]. Artificial neural networks have recently become very popular because of their success in obtaining much better results on hard problems compared to previous approaches [96, 97]. Machine learning has been applied to large language models [98], computer vision [99], speech recognition [100], medicine [101] and many more.

Smart Home technologies, in combination with Machine Learning, have become popular in recent years. Most applications of Machine Learning in Smart homes are focused on comfort, convenience, security or enjoyment [102]. In a smart home for the elderly, Ambient Assisted Living (AAL) could be used, which can help us understand patterns of the elderly's daily life and their quality of life could be improved with health suggestions or prevent injuries from a fall [103].

The World Health Organisation (WHO) predicts that by 2050, the elderly population over 65 years of age will outnumber children under 14 due to an increase in life expectancy [104]. In addition, approximately 15% of the world's population experiences various disabilities. Of these individuals, 110-190 million adults face significant functional difficulties [105].

Remote health monitoring is becoming a need in elderly care giving. Physiological signals can be monitored in real-time through remote monitoring without interrupting the daily activities of individuals, making it an unobtrusive and ubiquitous method. Remote monitoring, especially for the elderly who want to stay in the comfort of their home, is possible, thanks to the smart home technology [3].

If the focus is only on Acoustic Event Detection and Classification, it is clear that ML is the most popular method that gives great results compared to other methodologies [106]. After some advances in Machine learning, deep learning started to become more and more popular, but with the disadvantage of needing more data points and bigger and better computer systems. ML is still powerful on small tasks, but there is a need for some feature extraction and data preparation for the ML model.

2.4.2.2 Feature extraction

According to [107], feature extraction is one of the most important steps for audio signal processing. Audio signal has many features which are not all needed for audio processing. In the field of audio signal processing and pattern recognition, various methods for audio classification have been proposed based on system performance evaluation. These methods differ regarding the selection of classifiers and the number of acoustic features involved. Based on the decomposition, the extracted features are classified into temporal, spectral, and prosodic features.

Acoustic Event Classification plays a major role in audio signal processing and pattern recognition, enabling AED and classification applications. It involves accurately classifying the selected feature vectors into a suggested class. To tackle classification problems different types of classifiers are employed to tackle classification problems, including manual labelling, which is time-consuming, and supervised, unsupervised, and semi-supervised learning algorithms [108].

One of the most common ways to prepare a training dataset from audio is to transform the audio dataset into a time-frequency representation, such as Short Time Fourier Transform (STFT), which gives a richer representation of information, based on time, frequency and amplitude of the audio signal. STFT was developed mainly according to mathematical rules that led to the transition to linear frequency scales [109]. Because the human perception of frequency is not linear, mel scale is used. A human ear will identify the difference between 2 low pitches, 50 Hz and 150 Hz, but when it comes to 800 Hz to 900 Hz, even if the distance is the same, the human ear will distinguish a much smaller difference in pitch. Mel scale is a perceptual scale of pitches judged by listeners to be the same in distance from one another.

Sound analysis tasks often benefit from hand-crafted features, although few have been developed specifically for sound scene and event analysis. Existing features are typically adapted from other tasks rather than tailored for this specific problem. One such example is the use of mel frequency cepstral coefficients (MFCCs) in the sound scene and event analysis, which has become a well-established approach [102]. MFCC is a feature extraction technique, usually used in Speech Recognition and audio processing.

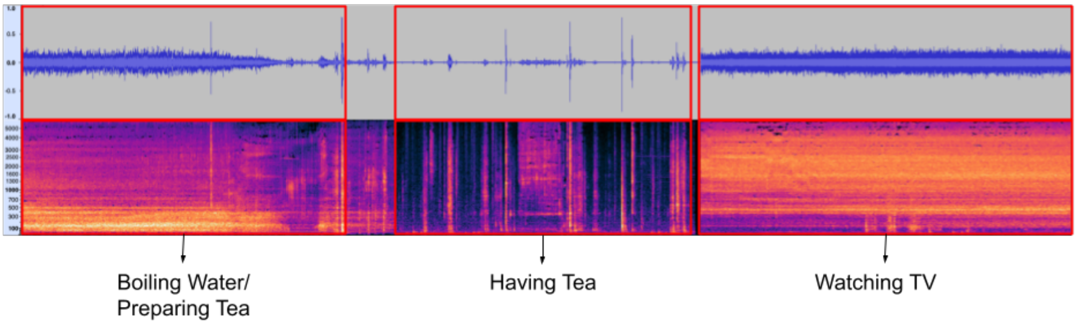


Figure 2.1: Audio activities and their spectrogram representation

MFCCs are a set of coefficients that capture the shape of the power spectrum of a sound signal. First, to capture the MFCCs of the sound, the sound needs to be transformed into the frequency domain using a technique like Discrete Fourier Transform (DFT) or Short Time Fourier Transform (STFT). Afterwards, the mel-scale needs to be applied for the sound frequency approximate to the human auditory perception [110]. Cepstral coefficients are derived from the mel-scaled spectrum at the end. In figure 2.1, audio from three activities, "Boiling Water/Preparing Tea", "Having Tea" and "Watching TV", is shown in both raw waveform and mel-scale spectrogram.

2.4.2.3 Machine Learning Models

In machine learning, there is a plethora of methods that could be used to train our model. The two subcategories of machine learning are supervised and unsupervised learning. For supervised learning, the data need to be labelled so that the model can name each event or activity through the data and teach the model what activity is correlated to specific data.

On the other hand, unsupervised learning is characterised by the absence of labels for specific data used to train the model. With unsupervised learning, a larger amount of data may be required; however, the identification of recurring patterns in the data is the primary goal. The exact names of specific events or activities will not be known, but changes between activities and patterns can be distinguished using an unsupervised model.

In a smart home environment for the elderly, having a supervised machine-learning model with a fusion of sensors is more common [73, 111]. These studies equipped the participant with wearable devices, accelerometer sensors, and various motion sensors in the house, including acoustic sensors (microphones). Using feature extraction and feature selection, they distinguished the more significant features that train Support Vector Machines (SVM) [3, 103]. SVM is a powerful supervised model that performs well in supervised classification and regression.

Support Vector Machines (SVM) Overview

The primary objective of an SVM is to find the optimal hyperplane that separates data points from different classes in a high-dimensional space [112]. The hyperplane is a decision boundary that separates different classes. The SVM algorithm identifies the hyperplane that maximises the margin, which is the distance between the hyperplane and the nearest data points from each class (these points are called support vectors). Support vectors are the critical elements of the dataset that lie closest to the hyperplane. They are essential because if they were removed, the position of the hyperplane would change. Thus, they have a significant influence on the model.

SVM can handle non-linearly separable data using a technique known as the kernel trick. The idea is to transform the data into a higher-dimensional space where a linear separation is possible [113]. Common kernel functions include linear, polynomial, and radial basis function (RBF) kernels. The choice of the kernel can significantly affect the performance of the SVM model.

During the training phase, SVM attempts to solve an optimization problem to find the hyperplane that offers the best separation while maximising the margin. Once trained, the SVM model can classify new data points by determining on which side of the hyperplane they fall. The class corresponding to that region is assigned to the data point.

Hyperparameter Optimisation is crucial for the performance of an SVM [113]. Here are some key hyperparameters to consider:

- **C (Regularisation Parameter)**: The C parameter controls the trade-off between achieving a low training error and a low testing error, which helps to prevent overfitting. A small value of C allows for a wider margin but may result in more misclassifications. In contrast, a larger value of C tries to minimise training errors, potentially leading to a narrower margin and overfitting.

- **Kernel Function**: The choice of kernel function determines how the SVM maps the input features into higher-dimensional spaces. Common kernel functions include:
  - **Linear Kernel**: Suitable for linearly separable data.
  - **Polynomial Kernel**: Useful for datasets with polynomial relationships.
  - **Radial Basis Function (RBF) Kernel**: Effective for non-linear problems, it transforms the data into an infinite-dimensional space.
  - **Sigmoid Kernel**: Less commonly used, this kernel can mimic the behaviour of neural networks.
- **Gamma**: This parameter is specific to the RBF kernel and affects the influence of individual training examples. A small gamma value means a broader decision boundary, while a high gamma value results in a more complex boundary that can lead to overfitting.
- **Degree**: This parameter is used with the polynomial kernel to control the degree of the polynomial. It allows you to adjust the complexity of the model, with higher values enabling the SVM to fit more complex relationships.

Tuning these hyperparameters is essential to optimising the performance of an SVM model. Techniques such as grid search or randomised search, often combined with cross-validation, can help identify the optimal hyperparameter values for a specific dataset.

Random Forests Overview

Another popular supervised methodology is Random Forests [114], which combines the output of multiple Decision Trees [115], a more traditional machine learning model, to reach a single result. Since the Random Forest consists of multiple Decision Trees, describing the decision tree algorithm will be helpful first. Decision trees start with a fundamental question, for example "Is anyone home?", afterwards you can ask a series of questions to get to the decided answer, "Are the lights off?" or "Is the water tap running?". These questions make up for the decision nodes in the tree, splitting the data that are more homogeneous in terms of their target class/answer. Information gain is the metric of homogeneity, which is a measure that quantifies how much the entropy of the dataset is reduced because of this particular split on a feature (water running, lights on/off). Each question directs us to the final answer or decision, which would be chosen by the leaf node. A decision tree calculates the information gain for each potential feature it could split on. The feature that results in the highest information gain is chosen to create the decision node. The feature with the highest information gain leads to the greatest reduction in entropy after the split. This means the resulting subset, after the split of the data, is more homogeneous in their decision/class.

When using a decision tree to analyse data, observations that meet certain criteria will follow the "Yes" branch, while those that do not will follow an alternate path. The goal of decision trees is to find the most effective way to divide the data into smaller subsets, and typically, the Classification and Regression Tree (CART) algorithm is used to train them. Metrics, like mean square error (MSE), can be used to evaluate the quality of the division [115]

Decision trees are a frequently utilised supervised learning technique, but they may encounter issues such as bias and overfitting. Nevertheless, the random forest algorithm can generate more precise predictions when multiple decision trees form an ensemble, especially when the trees are uncorrelated with one another. Some classifiers, such as decision trees, include ensemble learning methods, and their predictions are combined to determine the most popular result. The two most popular ensemble methods are bagging (bootstrap aggregation) and boosting. In 1996, Leo Breiman introduced the bagging method, which involves selecting a random sample of data from a training set, allowing individual data points to be selected more than once. Several models are then trained independently on the generated data samples, and depending on the task type (classification or regression), the predictions' majority or average provides a more accurate estimate [116]. This technique is frequently used to decrease variance in a noisy dataset.

The random forest algorithm is a method that uses both bagging and feature randomness to create an independent forest of decision trees. Feature randomness generates a random subset of features, also known as feature bagging or "the random subspace method" [114], which ensures that the decision trees are not strongly correlated. This is a crucial difference between decision trees and random forests. Instead of considering all possible feature splits, random forests only select a subset of those features. In the "Is anyone home?" example, the questions I ask to make the prediction may not be as comprehensive as someone else's questions. Considering all possible variations in the data could reduce the risk of overfitting, bias, and overall variability, resulting in more accurate predictions.

K-Nearest Neighbours Overview

Another simple supervised learning classification algorithm is the k-nearest neighbours algorithm, known as KNN or k-NN. KNN is a non-parametric classifier trained on the class labels assigned based on a majority vote and is frequently represented around a given data point [117]. In a binary classification, in which K is equal to two, for example, "Is anyone home?", if the majority voting is more than 50% towards "Yes", that means this data point is "closer", based on the "voting" to the "Yes" training data. The percentages will vary based on the number of classes that are had. For example, if five classes are available, which means k is equal to five, more than 20% will be looked at and so on. Regarding audio classification, the KNN algorithm was used in combination with a pre-trained deep learning model to create a more robust model. Embedding features were extracted using the VGGish model [118] from the audio, and afterwards, those embedded features were used to train the KNN algorithm.

The aim of the KNN algorithm is to identify the nearest neighbour of a given point by assigning a class label to that point. To achieve this, a distance metric must be defined that shows us the nearest neighbour. A distance metric is an objective score summarising the relative difference between two data points.

The most popular distance matrices are

- Euclidean Distance [119]

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

- Hamming Distance [120]

$$d_H(p,q) = \sum_{i=1}^{n}|p_i - q_i|$$

- Manhattan Distance [121]

$$d_M(p,q) = \sum_{i=1}^{n}|q_i - p_i|$$

- Minkowski Distance [122]

$$d_{Mk}(p,q) = \left(\sum_{i=1}^{n}|q_i - p_i|^r\right)^{\frac{1}{r}}$$

The most commonly used distance is Euclidean distance. It measures a straight line between the new unknown point and the class-labelled point to get the distance. Hamming distance calculates the distance between two binary or string vectors, identifying the points where the vectors or strings do not match. Manhattan distance, also called taxicab distance, calculates the absolute value between two points. And lastly, Minkowski Distance is a combination of Euclidean and Manhattan distance metrics. Choosing distance metrics can be tricky, and sometimes, trial and error must be done to get the best for the specific problem.

2.4.3 Datasets

In the domain of smart home care homes, where the focus is audio sensing, there are a limited number of datasets that could be used. As mentioned before, it is challenging to collect data on a new smart care home environment because labelling is time-consuming and expensive. The alternative solution is to use public datasets that have been collected and labelled by other researchers for analysis. In a perfect scenario, data collection will take place have collected data from the selected home environment using multiple microphones and use those data to train our models for classification, voice activity detection, speech separation and further analysis. This scenario is not always plausible, and that's why the use of public assets is very popular in this domain.

2.4.3.1 Public Datasets

There are numerous audio datasets collected and utilised by researchers across various domains. At times, researchers may require a subset of these large datasets, such as the smart care home environment, to fit the needs of their indoor environment domain. AudioSet dataset [123] is based on videos that are uploaded on the YouTube website. For AudioSet the labelling is done based on the name and description of the videos without start and stop times, which is called weakly labelling. Using a dataset that is weakly labelled is not optimal because a video might have extra sounds that are not correlated to the title or subtitle of the video. Having a large dataset with weak labels is useful when researchers need to train a deep-learning model on a large amount of data. For supervised classification, hard labelled data are preferred but in this previous study [124], this research group used weakly labelled data to train an attention neural network and they had great results.

Training with the whole AudioSet dataset is not very practical, can take a lot of time and it can be computationally expensive. This is how the Freesound Dataset 50K (FSD50K) was created [125], a benchmark dataset based on a smaller number of classes and audio files as a subset from the AudioSet. The creators of the FSD50K also created a baseline for audio classification challenges based on their dataset, hoping that more researchers will use it and develop new methods for audio classification using FSD50K.

TUT dataset [79] is also created to be part of an Audio Classification Challenge, called DCASE, to bring researchers together and test their audio classification skills [126]. TUT Acoustic Scenes 2017 dataset includes audio recordings from 15 distinct acoustic environments. Each acoustic scene consists of 3-5 minutes of audio recorded in various locations. The acoustic scenes included in the dataset are bus, cafe/restaurant, car, city center, forest path, grocery store, home, lakeside beach, library, metro station, office, residential area, train, tram, and park. This is not an enormous dataset, compared to AudioSet, which is very useful for researchers to use and train their machine-learning models without needing multiple expensive graphics cards.

One of the most widely used benchmark datasets is ESC-50 [127]. It contains 2,000 labelled environmental audio recordings suitable for testing environmental sound classification methods. The dataset is composed of 5-second-long recordings, categorised into 50 semantically different classes, each containing 40 examples. These classes are further organised into 5 major categories.

Table 2.2: ESC-50 Classes

Animals	**Natural soundscapes water sounds**	**Human, non-speech sounds**	**Interior/ domestic sounds**	**Exterior/ urban noises**
Dog	Rain	Crying baby	Door knock	Helicopter
Rooster	Sea waves	Sneezing	Mouse click	Chainsaw
Pig	Crackling fire	Clapping	Keyboard typing	Siren
Cow	Crickets	Breathing	Door, wood creaks	Car horn
Frog	Chirping birds	Coughing	Can opening	Engine
Cat	Water drops	Footsteps	Washing machine	Train

Continued on next page

Table 2.2: ESC-50 Classes (Continued)

Hen	Wind	Laughing	Vacuum cleaner	Church bells
Insects (flying)	Pouring water	Brushing teeth	Clock alarm	Airplane
Sheep	Toilet flush	Snoring	Clock tick	Fireworks
Crow	Thunderstorm	Drinking, sipping	Glass breaking	Hand saw

One of the most important datasets for our project was the speech dataset. In a domestic environment, conversations are likely to take place where an elderly person lives alone or with a caregiver, whether on the phone, with neighbours through the window, or with visitors. Speech is considered a natural part of human life. However, concerns might be raised regarding continuous audio recording for a smart home, as discomfort may be felt by those living in the home about being recorded. This socio-ethical issue requires further discussion. The goal is to utilise speech, along with background noise from the domestic environment, for the detection and classification of activities. It is pursued to determine whether speech can be removed while still allowing for the detection of changes in activities for further analysis.

After evaluating multiple speech datasets, the LibriSpeech dataset [128] was particularly useful for our needs. It contains hundreds of audio recordings of books read by volunteers, spoken clearly and loudly. This dataset provides us with the opportunity to use speech from multiple speakers, both male and female, which can be synthesised with background audio for tasks such as speech classification, removal, and voice activity detection.

By combining this dataset with others, a robust multi-purpose dataset that includes information on the number of speakers, the number of days of activity, and the percentage of audio versus speech in terms of loudness (measured in decibels) can be created.

In audio recording scenarios, it is common to have foreground and background sounds. However, in a smart care home setting, where a microphone may be near a television or a bedside, the distinction between foreground and background sound may not always be straightforward. For example, the sound of speech may be considered background noise if it is spoken softly while there is a louder sound in the foreground.

2.4.3.2 Real Deployments

Using public datasets to research and develop machine learning models is a great achievement. However, having a real deployment that collects real audio data alongside other types of data is unmatched. Using these data, any researcher can develop a model that works in similar or the same environments where the data were collected. In a perfect scenario, a small data collection would take place at the same smart home where the machine learning model would be deployed and collect data from that smart home.

Another option is to collect data in a similar home environment and use that data to train the model before deploying it in similar home environments. However, different background sounds and sound reflections will always depend on the walls, flooring, carpets, hardwood flooring, or marble used in each environment.

Sound INterfacing through the Swarm (SINS) Dataset [103] is a real deployment that consists of daily activities in a home environment. During a week-long experiment, an individual resided in an environment without simulated predefined scenarios, and their activities were recorded, including when they were absent, had visitors, or were working on a computer. The recorded scenarios included being absent, such as getting groceries and going for a walk. Although there was no restriction on the activities being performed, the number of labelled activities was limited, as indicated in Table 2.3.

In total, five different rooms were annotated with 16 different activities. Figure 2.2 shows the floor map of the recording environment and each room and table 2.3

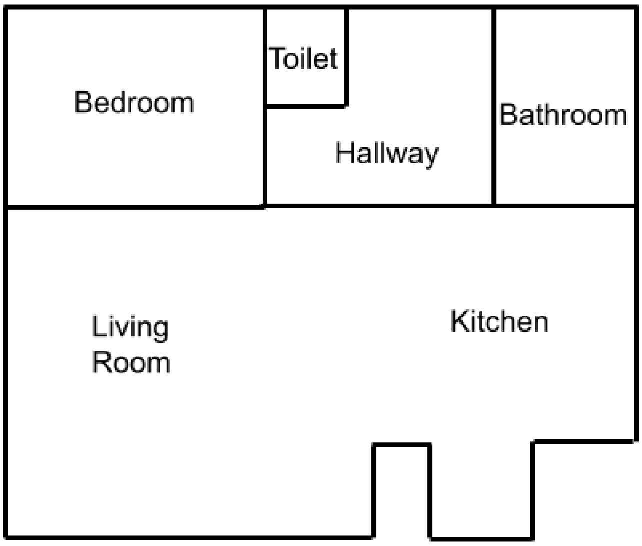


FIGURE 2.2: SINS Dataset Floor Map of the Recording Environment

lists each room's different activities. "Working" contains recordings of the person working on a computer, while "Other" represents the presence of a person not performing any activity listed in Table 2.3. Examples of recordings included in the "Other" activity are transitions between activities or the time between entering the room and starting an activity.

The dataset is strongly unbalanced, reflecting the imbalance of various activities in daily life. For example, activities such as "Absence" and "Watching TV" are 10 to 30 times longer in terms of total duration than the shortest activities, such as "Vacuum cleaning" and "Other." This ratio is even more significant for the other rooms. The acoustic sensor network setup consists of thirteen sensor nodes, each with four low-cost microphones, distributed across five rooms. The living room has eight nodes, the bathroom has one, the hall has one, the bedroom has two, and the toilet has one.

The SINS dataset annotation was performed in two phases. First, during data collection, the monitored participant used a smartphone application to annotate the activities while being recorded. The participant could only select activities listed in Table 2.3. The application was easy to use and did not significantly influence the transition between activities. Secondly, annotation software refined each activity's start and stop timestamps.

Table 2.3: Activities recorded in SINS Dataset for each room

Living Room	**Bathroom**	**Hallway**	**Toilet**	**Bedroom**
Phone call Cooking Dishwashing Eating Visit Watching TV Working Vacuum Other Absence	Drying with towel Shaving Showering Toothbrushing Vacuum Other Absence	Vacuum Other Absence	Toilet visit Vacuum Absence	Dressing Sleeping Vacuum Other Absence

Privacy-related aspects were taken into consideration during the data post-processing and sharing, as multiple people visited the home during the experiment, and the people's conversations during the "Phone call" activity were partially recorded. The database and annotation are publicly available [129], and all participants provided written informed consent.

Each audio channel was sampled sequentially at 16 kHz with a bit depth of 12. The acquired data was sent to a Raspberry Pi 3 for data storage. The data was stored in chunks of one minute and timestamped. Timestamps were obtained based on a Network Time Protocol (NTP) for rough synchronisation between the sensor nodes, with a sample accuracy of approximately 500 ms.

2.4.4 Deep Learning Applications

2.4.4.1 Introduction

Deep Learning is an advancement of Machine Learning techniques that aim to solve more supervised or unsupervised problems. The differentiating part of deep learning methodologies from machine learning is that they can be more complex in terms of computation and also the size of the dataset they can accommodate. Deep learning models can be trained on massive datasets to interpret audio and images, predict patterns and generate information [130]. This section will discuss various applications in deep learning related to acoustic sensing in a home environment and what methodologies were used.

Previously, multi-sensory approaches were used to support healthcare homes, aiming to classify activities in the home environment based on a fusion of sensory data collected from motion and smart devices [73]. The applications that were discussed in this section focus on audio as a modality and their ability to be used in a home environment to support the elderly. Most deep learning methodologies were implemented in Computer Vision and later adapted to work with audio data. The audio data are usually processed as a raw waveform or transformed into the frequency domain before applying the appropriate methods. The form of the data that is utilised by each methodology will be discussed. Convolutional Neural Networks (CNN) are being explored for automatic speech recognition, with the capability to predict the context of conversation through acoustic signals [131]. The CNN model is also employed for human activity recognition [132], allowing for the detection of various human activities, as well as for Acoustic Scene Classification to identify different environments. Long Short-Term Memory Networks (LSTM) and Transformers have also been utilised for Acoustic Scene Classification tasks [133]. Finally, the speech separation task will be discussed, which incorporates various deep learning methods, including AutoEncoders [134].

2.4.4.2 Speech Recognition

Speech recognition is a complicated task in which a deep learning model tries to define the context of speech in an audio stream. Previous attempts at solving this problem were made using a deep neural network. A deep neural network (DNN) is a feed-forward neural network with more than one hidden layer. Each hidden layer has a number of neurons, each of which takes all outputs of the lower layer as input, multiplies them by a weight vector, sums the results and passes it through a non-linear activation function such as sigmoid $\sigma(z) = \frac{1}{1+e^{-z}}$, tanh $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ or ReLU $f(x) = \max(0, x)$.

Abdel-Hamid, O. et al. [131] suggested deploying an alternative deep learning methodology for speech recognition on the TIMIT dataset [135] and large vocabulary voice search (VS) task [136] with better performance. Their proposed methodology uses Convolutional Neural Networks (CNN) instead. CNN is a type of neural network that differs from the standard version. CNN, instead of using fully connected hidden layers, utilises a unique structure consisting of convolution and pooling layers that alternate. The feature extraction of the audio signal that needs to be done for working with CNNs will now be described, along with an explanation of what the CNN method is. Afterwards, the datasets and their usage will be revisited.

To use the CNN for pattern recognition, the input data should be arranged as a set of feature maps that will be the input for the CNN. This method was borrowed from image processing applications where it is natural to arrange the input as a two-dimensional (2-D) array, which includes the pixel values at the horizontal x and vertical y coordinate indices. RGB (red, green, blue) values are three distinct 2-D feature maps for colour images. At both training and evaluation, CNNs run a small window over the input image, allowing the network's weights that look through this window to learn from various features of the input data, regardless of their absolute position within the input.

Moving back to speech recognition, To process speech using CNNs, the feature vectors need to be organised into feature maps. The input image can be considered a spectrogram without using RGB (red, green, blue). The authors [131] suggest that a single CNN input window contains a lot of context (9-15 frames) when it comes to speech. However, frequency's challenging due to the conventional use of MFCCs. The discrete cosine transform (DCT) used in MFCCs projects spectral energies into a new basis that may not retain locality. As shown in figure 2.3, the authors utilised log energy obtained directly from mel-frequency spectral coefficients (MFSC features) instead of MFCCs to address this issue. They also included their deltas and delta-deltas to describe each speech frame's acoustic energy distribution in different frequency bands.

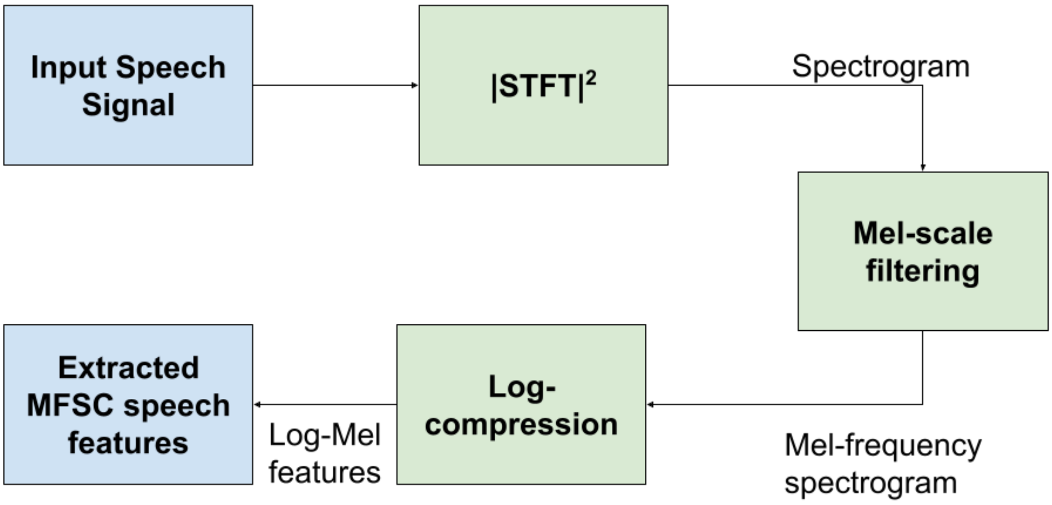


Figure 2.3: Extraction process of mel-frequency spectral coefficients (MFSC) features

The MFSC features are organised as one-dimensional (1-D) feature maps (along the frequency band index). For example, if the window contains 15 frames and 40 filter banks are used for each frame, 45 were constructed (i.e., 15 times 3) 1-D feature maps, each with 40 dimensions. As a result, a one-dimensional convolution will be applied along the frequency axis.

After the input feature maps are created, the convolution and pooling layers perform their respective operations in sequence to produce the activations of the units in those layers. The units of the convolution and pooling layers can also be grouped into maps similar to those of the input layer. In CNN terminology, a single CNN "layer" usually refers to a successive pair of convolution and pooling layers, as shown in the figure 2.4.

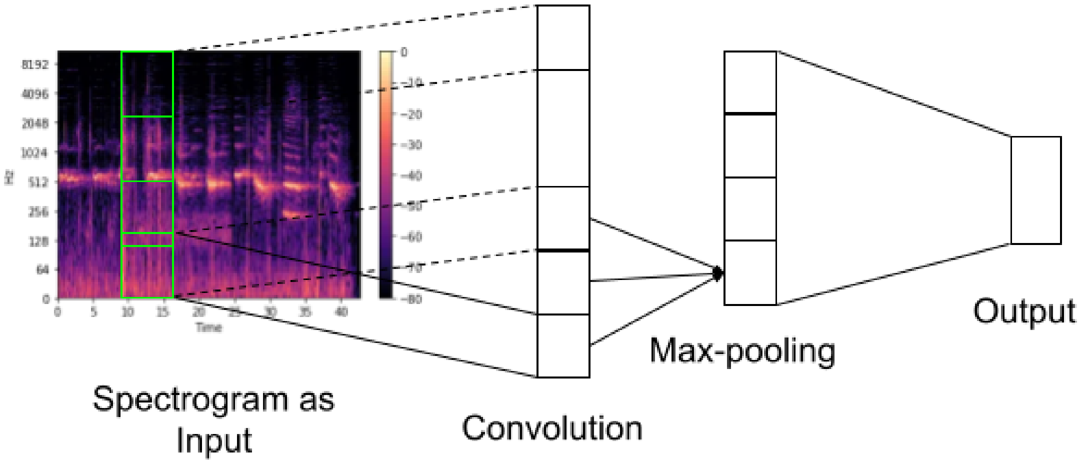


FIGURE 2.4: An illustration of the CNN applied to MFSC features of the speech signals.

The two datasets, TIMIT dataset [135] and large vocabulary voice search [136], use similar analysis methods. The audio signal is analysed using a 25-ms Hamming window with a fixed 10-ms frame rate. Speech feature vectors were generated by a Fourier-transform-based filter bank, which includes 40 log energy coefficients distributed on a mel scale, along with their first and second temporal derivatives. Each vector dimension of the speech data was standardised to have a mean of zero and a variance of one.

TIMIT training dataset consists of audio files from 462 speakers and a separate development set of 50 speakers. The development set was used to train all meta-parameters, including learning schedules and multiple learning rates. The test set consisted of 24 speakers and did not overlap with the development set. MFSC features were used in addition to the log energy feature per frame. The log energy was normalised to have a maximum value of one and then normalised to have a zero mean with unit variance over the whole training set. For this experiment, one convolutional layer, one pooling layer and two fully connected hidden layers on the top were used. The fully connected layers had 1000 units in each. The convolution and pooling parameters were pooling size of 6, shift size of 2, filter size of 8, and 150 feature maps for full weight sharing. The authors obtained minor differences in performance from the previous deep neural network methodology compared to using one convolutional layer. On the other hand, their results showed that using two convolutional layers tends to have great improvement. The overall performance of CNN configuration gave an 8% relative reduction in phone error rate compared to a deep neural network.

The voice search dataset contains 18 hours of speech data for vocabulary speech recognition. The deep learning model for this dataset utilised by CNN was established with two bigger hidden layers, each containing 2000 units. The initial 15 epochs were executed with a learning rate of 0.08, which was followed by 10 additional epochs with a decreased learning rate of 0.002. The CNN layer utilised limited weight sharing and had 84 feature maps per section. Its filter size was 8, the pooling size was 6, and the shift size was 2. The method Abdel-Hamid, O. et al. [131] showed performance improvement at 6% at word error rate from a standard deep neural network.

The methodology utilising CNNs for speech recognition was highly effective in accurately transcribing spoken words from audio streams. This success can be attributed to the use of appropriate feature extraction techniques, which allowed the system to process and analyse the audio data in a manner that enabled accurate recognition of speech patterns and phonemes. Overall, this approach proved to be a reliable and efficient means of converting audio input into text output.

2.4.4.3 Human Activity Recognition

Human Activity Recognition (HAR) was previously conducted with the assistance of various sensors, such as accelerometers and gyroscopes, which can be found in smartphones or wearables like smartwatches [137, 138]. In a previous study, it was observed that researchers utilised accelerometer data [139] collected from smartphones to build a deep learning model aimed at HAR for the elderly. Their architecture consists of feature extraction from the accelerometer data and then using a feed-forward neural network with 3 hidden layers for training and predicting human activity.

Using a smartphone or wearable device makes sense because these devices are very versatile and equipped with multiple sensors. Zerkouk et al. [140] utilised a public dataset of smartphone accelerometer data. They trained an Autoencoder-CNN-LSTM model to identify and predict abnormal behaviours of elderly individuals accurately. As the elderly population continues to grow, so does the need to monitor their health and well-being. While smartphones and wearables have become popular tools for tracking health data, they may not be the most suitable option for some adults, particularly those who suffer from cognitive deterioration such as dementia.

In these cases, alternative options for collecting passive data should be explored. For instance, the Internet of Things (IoT) is a very popular domain and can be explored further [141]. In an IoT or smart home environment, sensors are placed around the home and can provide insight into a person's daily activities, such as how much they move around, how often they go to the bathroom, or whether they eat and drink regularly. These data can then be collected and analysed to identify any changes or concerns requiring attention.

By using non-intrusive data collection methods, caregivers and healthcare specialists can better monitor the health and well-being of the elderly without disrupting their daily lives [142]. This can lead to earlier detection of health issues and more effective interventions, ultimately improving the quality of life for seniors and their families.

2.4.4.4 Acoustic Scene Classification

Acoustic Scene Classification had a great interest, and research communities, like the Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge [143], motivated researchers to get more involved in the complications of this task. In 2013, the DCASE Community released an acoustic dataset with one or more tasks as a challenge and in 2016, DCASE released a yearly challenge with open-source datasets for everyone to try to solve the task. Researchers worldwide can download the acoustic dataset and apply various forms of computation to solve the task.

An example research paper from DCASE 2016 was implementing a CNN model for acoustic scene classification [144]. This study used the TUT Acoustic Scenes 2016 dataset [79], comprising 15 audio classes. The audio classes are:

- beach
- bus
- cafe/restaurant
- car
- city center
- forest path
- grocery store
- home
- library
- metro station
- office
- park
- residential area
- train and tram

The recordings were gathered from various places, mostly in Finland. They used a sampling rate of 44.1 kHz and a resolution of 24 bits. Each location was recorded for 3-5 minutes and the original recordings were divided into 30-second parts. For audio pre-processing, the authors utilised Constant-Q-Transform (CQT) [145]. CQT is a time-frequency representation characterised by geometrically spaced frequency bins and equal Q-factors, which are the ratios of the centre frequencies to bandwidths for all the bins. The CQT is essentially a wavelet transform, which means that the frequency resolution is better for low frequencies and the time resolution is better for high frequencies. Using the TUT dataset, a 30-second input file results in 31 CQT excerpts of shape 80 bands × 82 frames. The CNN model was trained using the CQT excerpts of shape 80 x 82. Throughout the training process, multiple training examples are taken from segment-wise CQT extraction of all the files present in the training set during each epoch. The CNN layers were followed by a Max Pooling Layer, which performs sub-sampling of the matrices that are output after applying the the filter kernels for the CNN model.

Weights are learned in a filter kernel that has a specific shape. To apply these filter kernels to the input data, the authors used 32 filters. These filters are applied through convolution, which involves multiplying the filter kernel with an equally size portion of the input image. After this, the filter kernel window is moved sequentially over the input data, usually from left to right and top to bottom, producing an output of either equal size (when padding is used at the borders) or reduced by the filter length minus one on each axis (when no padding is used, and the filter kernel is kept inside the borders of the input).

The pipeline's purpose is to capture relationships between frequencies. To achieve this, it uses filter kernel sizes of 14x15 and sets the Max Pooling size to 1x17. As a result, the filtering step produces 32 matrices that are 67x68 in shape. These matrices are then "pooled" to create 32 matrices that are 67x4 in shape. By pooling the matrices, more information on the frequency axis is preserved compared to the time axis. The complete network has 4,126,223 parameters, which hints at the power of Convolutional Layers.

The Gaussian Mixture Model (GMM) classifier was used as the baseline system, which utilises MFCC audio features. These features are calculated by using frames of 40 ms with a Hamming window and 50% overlap. The system extracts 40 Mel bands, but only retains the first 20 coefficients, along with delta and acceleration coefficients, amounting to 60 values in total. To classify, the system learns one acoustic model for each acoustic scene class (GMM with 32 components) and employs a maximum likelihood classification scheme (expectation maximisation). The average classification of the authors' proposed method's accuracy over 4 folds is reported to be 72.5%.

Another methodology that was developed because of a DCASE task from 2019 was proposing to tackle the sound event detection task [146]. The authors chose to use a combination of CNN and Long Short-Term Memory Networks (LSTM) for this purpose. For this study, TAU Spatial Sound Events 2019-Microphone Array [147] Dataset was used. The TAU dataset was recorded with a four-channel directional microphone. This allows the dataset to gather sound data from all directions. The dataset consists of development and evaluation sets, containing 400 sound clips. Each clip has four channels of recording waveforms, along with its corresponding CSV file containing the recording of sound events that took place during the clip. The CSV file also includes the starting and ending time of each sound event, as well as the azimuth and pitch angle of the direction from which the sound source originated.

For the audio pre-processing, the authors extracted features at a sampling frequency of 32000Hz, used 1024 Mel filters and FFT points, 64 Mel bins, and divided each second into 64 units. They obtained the Log-Mel spectrum by calculating the STFT result.

An 8-hidden-layer ply comprising a CNN layer followed by a Max Pooling layer was utilised, followed by a dropout layer and an LSTM layer. Because of the CNN, it can understand spatial features and an LSTM is designed to handle sequential features. LSTM discovers time-dependent patterns, learning how features change from moment to moment. LSTMs achieve this through a gating mechanism consisting of a forget gate, input gate, and output gate. These gates regulate information flow within the LSTM cell. The forget gate determines what information to discard from the previous cell state, the input gate controls the extent to which new information is added, and the output gate modulates the information flow from the current cell state to the network's output. This architecture allows LSTMs to learn long-term dependencies in sequential data effectively.

The authors used an Adam optimiser and 50 epochs to observe the change in error rate. After 15 epochs, the trainer converged to a more stable situation. Their methodology of combining CNN and LSTM showed a detection accuracy of 85.2% and an F-1 score of 90.4%.

Deep Learning models, such as CNN and LSTM, showed great potential in the acoustic space. However, a massive labelled dataset is needed to train those models without overfitting. The Transfer Learning approach can be utilised for scenarios where collecting a lot of labelled data is not feasible. One example of a limited data scenario is a study done with crowd-sourcing sound data for automatic diagnosis of COVID-19, where users used a smartphone application to collect audio recordings [77]. The authors used transfer learning for their approach. A pre-trained model VGGish was used to extract embedding features alongside handcrafted features, for example, the temporal differential (delta) of the Δ-MFCC, Δ^2-MFCC: the differential of the delta of the MFCC (acceleration coefficients). Afterwards, the authors split participants into 80/20 disjoint train and test set, and they trained a Support Vector Machine (SVM) classifier with a Radial Basis Function (RBF) kernel based on the hand-crafted and features extracted from the VGGish pre-trained model with a Precision 80% and Recall of 72%.

2.4.4.5 Speech Separation

Speech Separation is a long-lasting challenge, where the goal is to identify one source of audio or speech from multiple sources [148]. This effect is called "the cocktail party effect", where multiple people speak simultaneously, but we, humans, can focus on one source and dim down the other audio sources. The speech

Separation task is a more advanced task nowadays when the complication of multiple speakers, from 2 to 8 speakers, was added in the mix, with multiple background sounds.

The research paper that got a lot of attention is called Attention is All You Need [149], where the authors discussed that the attention mechanism can be utilised for many deep learning tasks, not just Natural Language Processing. Afterwards, Subakan, C. et al. [150] proposed a new neural network called the SepFormer for speech separation based on Transformers instead of Recurrent Neural Networks. It uses a multi-scale approach to learn both short and long-term dependencies. The model has been trained on standard WSJ0-2/3mix datasets [151] and has achieved state-of-the-art performance. The SepFormer (Separation Transformer) model is primarily made up of multi-head attention and feed-forward layers. The dual-path framework was utilised, as introduced by DPRNN [152], and has substituted the RNNs with transformers. The proposed pipeline is capable of both short and long-term dependencies. The dual-path framework helps reduce transformers' quadratic complexity by allowing them to process smaller chunks of data.

Eliminating previous RNN models' recurrence and replacing it with a fully attention-based mechanism enables Transformers to avoid this bottleneck completely. By accessing the entire sequence simultaneously, Transformers can establish a direct connection between distant elements, making it easier to learn long-term dependencies [153]. This is why Transformers are becoming increasingly popular in speech processing.

Subakan, C. et al. [150] proposed the SepFormer model, which utilises a masking approach and consists of an encoder, a decoder, and a masking network. The encoder is fully convolutional, and the masking network employs two Transformers that are embedded within the dual-path processing block, as proposed in [152]. By utilising the masks predicted by the masking network, the decoder is able to reconstruct the separated signals in the time domain. SepFormer model is available for other researchers to download and use online through the SpeechBrain toolkit [154].

The novelty of the SepFormer from previous Speech Separation models is the attention mechanism. The benefit of the attention mechanism is understanding each section of the audio stream based on its surrounding audio sections. An example of how the attention mechanism works in natural language processing (NLP). Using the attention mechanism, the model can understand the meaning of a word based on the rest of the sentence. BERT, a transformer model that is utilised in NLP [155], uses a process of contextual analysis to understand the meaning of words in a sentence. In this example, "He got bit by Python," BERT analyses the relationship between each word in the sentence to grasp its contextual meaning. To understand the context of the word "Python", BERT examines the word in relation to all other words in the sentence. As shown in figure 2.5, BERT understands that "Python" in this sentence refers to something that can bite.

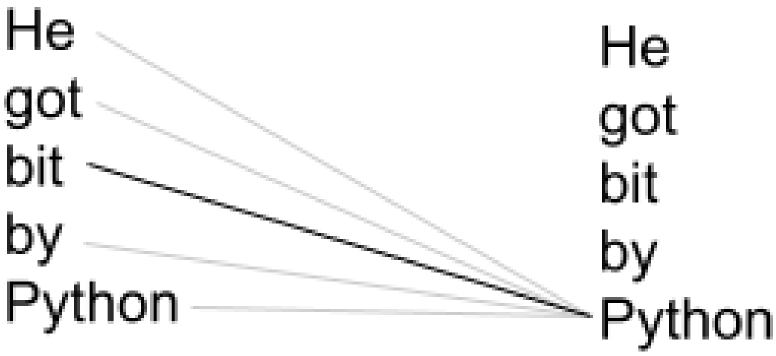


FIGURE 2.5: Example one: Relationship of word Python with the rest of the sentence

Another example, shown in figure 2.6, is "Python is a programming language", where the word "Python" has a different meaning and can be matched with the word "programming."

Regarding multi-head attention, the model goes through all the words and checks their relation with the rest of the sentence. Going back to SepFormer [150], as

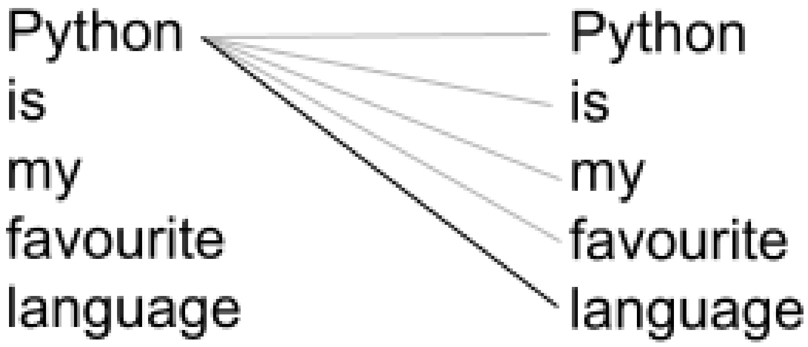


FIGURE 2.6: Example two: Relationship of word Python with the rest of the sentence

mentioned earlier, the model consists of an encoder, the masking network and the decoder. The encoder takes in the time domain audio signal as input, which contains audio signals from multiple speakers. Then, the model is trained on STFT-like representations using a single convolutional layer $h = ReLU(conv1d(x)$.

The second part is the masking network. It uses the encoded representations from the encoders and estimates a mask for the Ns speakers in the mixture. The encoded input h is normalised with layer normalisation and processed by a linear layer with dimensionality F. Then, the authors created overlapping chunks of size C by chopping h in the time axis with a 50% overlap. Afterwards, the authors denoted the output of the chunking operation with $h' \in \Re^{FxCxNc}$. C is the length of each chunk, and Nc is the number of chunks. h' is fed into the SepFormer block, which is the main component of the masking network. The output of the SepFormer block is processed by Parametric ReLU (PReLU) activation followed by a linear layer. Then, the representation goes through two feed-forward layers and a ReLU ($f(x) = \max(0, x)$) activation at the end to obtain the mask for each speaker.

The decoder ultimately employs a transposed convolution layer that shares the same stride and kernel size as the encoder. To decode, the decoder takes the element-wise multiplication of the mask of the source k and the output of the encoder h as input. The transformation of the decoder denotes the separate source k. This now is the pipeline SepFormer uses to separate multiple overlapping speakers from the audio signal.

SepFormer was trained with WSJ0-2mix dataset [151]. WSJ0-2mix is generated from the Wall Street Journal (WSJ0) audio set. The dataset consists of 30 hours of training data, 10 hours of validation data and 5 hours of test data at 8 kHz sampling rate. Each mixture in the dataset is artificially generated by randomly selecting different speakers from the corresponding set. The speakers are then mixed at a random relative signal-to-noise ratio (SNR) between -5 and 5 dB. The proposed architecture, SepFormer, demonstrates exceptional performance on the test set with dynamic mixing, achieving an improvement of 22.3 dB in scale-invariant signal-to-noise ratio (SI-SNR) and 22.4 dB in signal-to-distortion ratio (SDR) [156]. When utilising dynamic mixing, the SepFormer achieves a state-of-the-art performance, setting a new benchmark in the field.

Deep learning models, like SepFormer, are made to separate overlapping speakers into two audio signals that consist of speech data from each speaker. In a home environment, the most common situation is background noise with speech from the occupant. The task should focus on separating background domestic noise from the speech, and models like SepFormer that focus on multiple-speaker separation will not be the most appropriate. In the past, we've seen various applications, such as Speech Enhancement, which could be closer to our goal. Facebook Research group created a deep learning model called Denoiser [157], which eliminates all the background noises and enhances speech audio data that come from one speaker.

The Denoiser model [157] uses a DEMUCS architecture [158] adapted for speech enhancement. DEMUCS architecture is an Auto Encoder model that incorporates U-Net [159] like skip-connections. The model consists of a multi-layer convolutional encoder and decoder. The layers of the encoder and decoder are assigned numbers, starting from 1 and going up to L. For the decoder, layers on the same scale have matching indices with the encoder in reverse order.

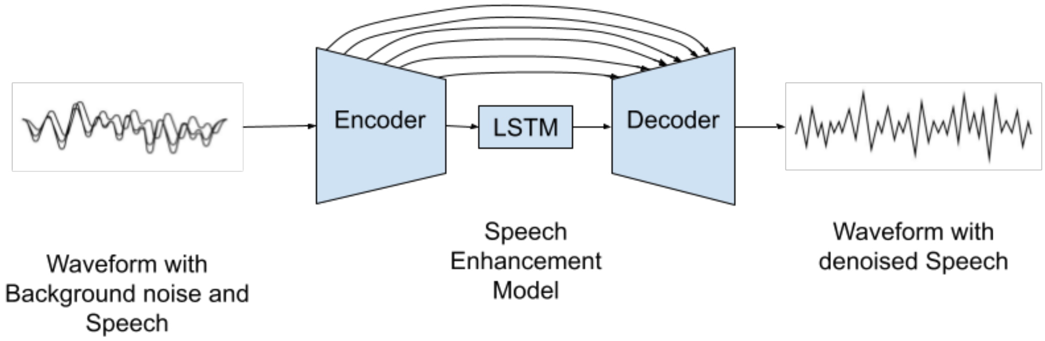

Encoder
LSTM
Decoder
Waveform with Background noise and Speech
Speech Enhancement Model
Waveform with denoised Speech

FIGURE 2.7: DEMUCS Architecture

Denoiser uses an audio signal in the waveform to remove noise from the speech signal and returns it back to the waveform in a single channel. The latent representation $E(x) = z$ is produced by the encoder network E, which takes the raw waveform as input. Each network layer comprises a convolution layer with a kernel size of K and stride of S, with $2^{i-1}H$ output channels. This is followed by a ReLU activation, a "1x1" convolution with 2^iH output channels, and finally, a Gated Linear Unit (GLU) activation [160] that converts the number of channels back to $2^{i-1}H$.

A sequence modelling R network takes the latent representation z as input and outputs a non-linear transformation of the same size, denoted as ẑ, $R(z) = LSTM(z) + z$. The LSTM network comprises two layers and $2^{L-1}H$ hidden units. A unidirectional LSTM ensures causal prediction, while a bidirectional LSTM is preferred for non-causal models. Both outputs are merged using a linear layer.

The final component in the system is a decoder network, denoted by D, which takes a corrupted signal's estimated latent representation, ẑ, as input and produces a clean signal estimate, ŷ. Each layer of the decoder network takes $2^{i-1}H$ channels as input, applies a 1x1 convolution with 2^iH channels, then a GLU activation function that outputs $2^{i-1}H$ channels, and finally a transposed convolution with a kernel size of 8, stride of 4, and $2^{i-2}H$ output channels with a ReLU function. The output of the final layer is a single channel with no ReLU. A skip connection links the output of the i-th layer of the encoder to the input of the i-th layer of the decoder.

The DEMUCS model was trained for 400 epochs on the Valentini dataset [161] and 250 epochs on the DNS dataset [162] by the authors. The L1 loss between the predicted and ground truth clean speech waveforms was used. Additionally, the authors added the STFT loss for the Valentini dataset. The STFT loss was given a weight of 0.5. The authors used the Adam optimiser with a step size of 3e-4, a momentum of $\beta1 = 0.9$, and a denominator momentum $\beta2 = 0.999$. For the Valentini dataset, the original validation set was used, and the best model was kept. For the DNS dataset, the authors trained the model without using a validation set and kept the last model. The audio was sampled at 16 kHz.

The input was normalised by its standard deviation before feeding it to the model, and the output was scaled back using the same factor. The process audio had a frame size of 37 ms and a stride of 16 ms. Random shift and Remix augmentation shuffle and SpecAug [163] are used for audio augmentation. Lastly, the DEMUCS model, a state-of-the-art architecture developed for music source separation in the waveform domain, could be used for a casual speech enhancer. With state-of-the-art results on standard Valentini benchmark. By incorporating additional data augmentation techniques during the training phase, the performance of the speech enhancement model can be improved. An Automatic Speech Recognition model performance can be improved using this methodology in noisy conditions.

Various deep learning applications that come close to achieving our goal were mentioned, but they are not exactly tailored to it. For instance, SepFormer [150] can separate multiple speech signals that overlap with each other to identify the context of each speech. On the other hand, the Denoiser model [157] uses an architecture commonly used in music to separate different audio sources. The denoiser model's purpose is to eliminate any background noise from speech audio data, which improves the accuracy of Automatic Speech Recognition. However, in our case, our objective is to enhance the background sound and extract speech from the audio signal, ensuring that the occupant's privacy is maintained in an acoustic environment.

2.5 Conclusion

This literature review delved into the health context of elderly individuals living independently at home. Additionally, it investigated how dementia can impact their daily habits and routines. To enhance the quality of life for the elderly, smart solutions that utilize various digital devices, including smartphones, wearables, and smart home devices, were presented.

Following this, the challenges faced by current methods when caring for the elderly living at home were examined. Issues surrounding wearable devices and the collection of labelled data were identified. Furthermore, the difficulties associated with curating large datasets with labels were highlighted.

Lastly, different types of audio sensing techniques, machine learning methodologies, datasets, and deep learning applications that can be used to address these challenges were explored. This literature review provided a comprehensive summary of the existing research on the topic. Areas needing more focus were identified, such as monitoring changes in activities in an acoustic environment and investigating privacy concerns related to the acoustic environment. The objective of this work is to enhance the health and well-being of elderly individuals who live independently in their homes.

Chapter 3

Datasets

3.1 Introduction

We are currently exploring the curation of datasets specifically for our research on Tracking Daily Routines and Speech Preservation. To evaluate the effectiveness of the proposed system, it is essential to collect data that reflects the dynamics of daily life. Public datasets are required as well as data from real-world deployments that can accurately replicate the home environment where our system will be utilised. This will help ensure that the system operates efficiently in real-world conditions.

For this experiment, two types of audio data were used:

- Background Data
- Speech Data

Background data from indoor environments is required to ensure accurate analysis. This data can vary significantly from other audio data due to quieter atmospheres and more distinct sounds, including footsteps on different types of flooring or the opening and closing of doors and cupboards. Speech data collected from multiple speakers is also required to investigate privacy preservation regarding speech.

3.2 Public Datasets

For our study, various datasets are being utilised to accurately represent a real-world deployment where the system will be used. This section provides a detailed explanation of the data gathered from public datasets. This will offer a better understanding of the nature of the experiments and the significance of the findings that will be presented.

- For background sounds:
  - Audio Set [123]
  - Freesound Audio Tagging 2019 dataset [164]
  - ESC-50 [127]
  - SINS [129]
- For speech data:
  - LibriSpeech [128]

The Audio Set [123] was created following the success of ImageNet [165]. The Audio Set was created using the audio of videos uploaded on YouTube. It aims to fill the gap in data availability between image and audio research. The dataset uses a 6-level hierarchical structure, for example, "Sounds of things" → "Vehicle" → "Motor vehicle" → "Emergency vehicle" → "Siren" → "Ambulance (siren)". Audio Set consists of over 2 million audio clips of 632 audio classes. Segments of 10-second audio clips were selected for labelling using searches based on metadata, context such as links, and content analysis.

The labelling process for Audio Set is carried out by analysing the metadata, like title and description of videos, without considering the start and stop times. This type of labelling is known as weak labelling. However, it may not always be the best approach, as videos can have sounds that are unrelated to their title or subtitle. Despite this limitation, having a large dataset with weak labels can prove helpful for training deep-learning models when a large amount of data is required. Typically, hard-labelled data is preferred for supervised classification. However, in a previous study conducted [124], weakly labelled data was used to train an attention neural network, yielding promising results. Training using the entire Audio Set dataset can be impractical, time-consuming, and computationally expensive. Therefore, smaller benchmark datasets like ESC-50 [127] were created to address this issue. These datasets have fewer classes and audio files, which were selected as a subset from the Audio Set.

The Environmental Sound Classification (ESC-50) dataset [127] is a collection of 2,000 labelled environmental audio recordings extracted from a human-labelled dataset called FreeSound 50K [125] and is one of the most widely used benchmark datasets. FreeSound 50K is a subset of AudioSet, which was mentioned earlier. The ESC-50 dataset is suitable for testing environmental sound classification methods and consists of 5-second-long recordings sampled at 16 kHz, categorised into 50 different classes, each containing 40 examples. These classes are further organised into 5 major categories, which are Animals, Natural Soundscapes, Human Non-speech Sounds, Interior Sounds, and Exterior Noises. You can refer to Table 3.1 for more information on the different groups of classes in the dataset.

Table 3.1: ESC-50 Classes

Animals	**Natural soundscapes water sounds**	**Human, non-speech sounds**	**Interior/ domestic sounds**	**Exterior/ urban noises**
Dog	Rain	Crying baby	Door knock	Helicopter
Rooster	Sea waves	Sneezing	Mouse click	Chainsaw
Pig	Crackling fire	Clapping	Keyboard typing	Siren
Cow	Crickets	Breathing	Door, wood creaks	Car horn
Frog	Chirping birds	Coughing	Can opening	Engine
Cat	Water drops	Footsteps	Washing machine	Train
Hen	Wind	Laughing	Vacuum cleaner	Church bells
Insects (flying)	Pouring water	Brushing teeth	Clock alarm	Airplane
Sheep	Toilet flush	Snoring	Clock tick	Fireworks
Crow	Thunderstorm	Drinking, sipping	Glass breaking	Hand saw

The LibriSpeech Corpus [128] only includes speech audio and contains about 1000 hours of recorded audio, sampled at 16 kHz. Each audio clip represents a book chapter, but if the chapter is longer than 30 minutes of audio recording, it is split into multiple files. It is worth mentioning that this dataset is a part of the larger project called LibriVox, which aims to create a comprehensive digital library of public domain books. The LibriSpeech corpus was explicitly designed to cater to the needs of researchers and developers working on speech recognition technologies and has been used in various applications ranging from speech-to-text transcription to speaker identification.

A subset of the LibriSpeech Corpus was needed as part of the research. A dataset consisting of 20 speakers was explicitly required. Since the full dataset was too large, one chapter of 20 books was randomly chosen, resulting in speech data from 20 different speakers being collected for 30 minutes.

3.3 Real Deployments

Public datasets consist of audio data that were not collected through real deployment scenarios. Some data came from a large dataset captured from YouTube, while others were recordings made by volunteers reading books. However, real deployments are different. In these scenarios, microphones are installed in participants' homes and data are collected as they go about their daily routines. Collecting this type of dataset can be time-consuming and burdensome for both participants and researchers.

Moving forward, two types of deployments will be discussed. The first deployment, collected by the European project SINS [166], has been granted access for use in this study. The researchers created the second deployment, which involved two stages of data collection, with some amendments made between the two stages. It is assumed that the data received from these deployments will correspond to the data that will be obtained when the proposed system is deployed in a new home.

3.3.1 Sound INterfacing through the Swarm (SINS) Dataset

In recent years, there has been a growing interest in developing smart environments that cater to the needs of the elderly population [166]. These environments aim to enhance the quality of life for the elderly in various aspects, such as safety, security, comfort, and home care. However, traditional wearable systems or devices that require carrying have not been found to be practical by elderly users [167]. Therefore, a passive smart system that can monitor and respond to the needs of the elderly without requiring any physical interaction should be more suitable for this task.

In order to obtain a dataset from a passive smart system that is deployed in the real world, the Sound INterfacing through the Swarm (SINS) Dataset [103] was created. This dataset is derived from a real-world deployment that captures daily activities in a household environment.

Throughout a week-long experiment, an individual lived in a home environment without any predefined activity scenarios. Instead, the person's actual activities were recorded, including times when they were absent, had visitors, or were working on a computer. The recorded scenarios included being absent (e.g. getting groceries and going for a walk). Although there was no restriction on the activities being performed, the number of labelled activities was limited, as indicated in Table 3.2. In total, 16 different activities were annotated in five different rooms. Table 3.2 lists the different activities for each room. Most of the activities are self-explanatory, except for "Working" and "Other." "Working" contains recordings of the person working on a computer. The activity "Other" represents the presence of a person when not performing any activity listed in Table 3.2. Examples of recordings that are included in the "Other" activity are transitions between activities or the time between entering the room and starting an activity.

The dataset is strongly unbalanced, reflecting the imbalance of various activities in daily life. For instance, activities like "Absence" and "Watching TV" last 10 to 30 times longer than shorter activities like "Vacuum" and "Other." This imbalance is even greater for the other rooms. The acoustic sensor network setup consists of thirteen sensor nodes, each with four low-cost microphones, distributed across five rooms. The living room has eight nodes, the bathroom has one, the hall has one, the bedroom has two, and the toilet has one.

The annotation was performed in two phases. First, during data collection, a smartphone application was used by the monitored participant to annotate the activities while being recorded. The participant could only select activities listed in Table 3.2. The application was easy to use and did not significantly influence

Table 3.2: Activities recorded in SINS Dataset for each room

Living Room	**Bathroom**	**Hall**	**Toilet**	**Bedroom**
Phone call Cooking Dishwashing Eating Visit Watching TV Working Vacuum Other Absence	Drying with towel Shaving Showering Toothbrushing Vacuum Other Absence	Vacuum Other Absence	Toilet visit Vacuum Absence	Dressing Sleeping Vacuum Other Absence

the transition between activities. Secondly, the start and stop timestamps of each activity were refined by using annotation software.

Privacy-related aspects were considered during the data post-processing and sharing, as multiple people visited the home during the experiment, and the people's conversations during the "Phone call" activity were partially recorded. All participants provided written informed consent, and the database and annotation are publicly available.

The sampling for each audio channel is done sequentially at 16 kHz with a bit depth of 12. The acquired data is sent to a Raspberry Pi 3 for data storage. The data is stored in chunks of one minute and timestamped. Timestamps were obtained based on a Network Time Protocol (NTP) for rough synchronisation between the sensor nodes, with a sample accuracy of approximately 500 ms.

3.3.2 Data Collection

In addition to the public dataset, our system was evaluated through a small-scale collection of real-world sounds. For the first phase of this data collection, 10 participants were asked to record their morning routine using their smartphones over a period of 5 days. The data was collected through the AudioHive App [168], a purpose-built application developed for this study. The app allows participants to label their activities. Activity durations and sequences varied across participants, with some activities being of short duration (e.g., an average of 2 minutes for "brushing teeth") and others being much longer (e.g., an average of 10.6 minutes for "having coffee"). Similar to the public dataset, the labels were only utilised as part of the system evaluation process. A second phase of data collection is planned to further enhance our findings.

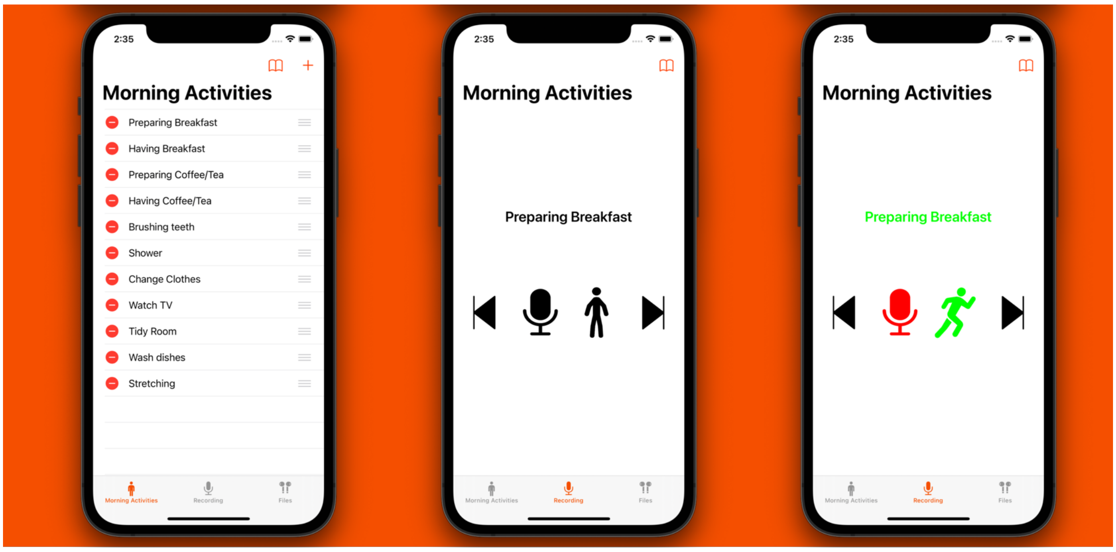


FIGURE 3.1: AudioHive 2.0: Data Collection App

The AudioHive: Passive activity Sensing Data Collection for Assisted Living was approved by the Central Research Ethics Advisory Group (CREAG) at the University of Kent.

This study aims to simulate a real-world deployment scenario in which elderly people are at home, going through their daily lives, and the smart home device records environmental audio data for analysis. For the data collection, an App was developed 3.1, which participants can download on their iOS or Android devices to complete the data collection. The App was purposely created for data collection, so the user interface was designed to be simple and easy to use.

The App was developed naively on both platforms, with Swift Language being used for iOS devices and Java Language for Android devices. Participants were asked to download the App on their smartphone devices and fill out a list of

Table 3.3: Activities recorded using AudioHive App with the average length of each activity

Activity	**Length of Activity (minutes)**
Having Breakfast	8
Preparing Coffee/Tea	4
Having Coffee/Tea	10
Brushing teeth	2
Showering	10
Watching TV	18
Having lunch	14
Making/preparing dinner	16
Eating dinner	11
Washing dishes	5
Playing board games	10
Hoovering	8
Washing face + creams	6
Getting ready	3

usual activities that are performed during their morning routine at home. Examples were provided to ensure clarity regarding the expected labels for this data collection. Some of the examples proposed include "Preparing Breakfast," "Having Breakfast," "Preparing Coffee/Tea," "Having Coffee/Tea," "Brushing teeth," "Shower," and "Watching TV."

In the second phase of data collection, 12 healthy adults, 6 males and 6 females participated. The participants were asked to select around four daily activities that they would like to track and manage on the app. They were given the option to modify the list of activities on the app. This allowed them to choose their specific activity when recording their data.

The data collection took place over a period of 7 days, and the app allowed participants to label their activities. The activity duration and sequences varied across participants, with some activities being short (e.g. an average of 2 minutes for "brushing teeth" or "preparing coffee") and others being much longer (e.g. an average of 8 minutes for "having coffee"). In table 3.3, there is a complete list of audio activities recorded by the participants and an average of each recording. The labels were only used as part of the system evaluation process, similar to the public dataset.

During the second data collection phase, participants were asked to speak during the activities if the activity allowed it. They were advised to discuss non-personal scenarios, explain their activity, or read from a provided script. The AudioHive on iOS recorded M4A audio files, and the AudioHive on Android recorded WAV files. Both platforms were recorded at a 16 kHz sampling rate and stored temporarily on the participant's device. The App created a CSV file with the appropriate timestamp information, and they were uploaded to the server through the App. After the data were uploaded to the server, any M4A audio files were converted to WAV files for further analysis.

3.4 Dataset Synthesis

The appropriate datasets for our methodology will be developed starting with speech data from the LibriSpeech Dataset [128]. The LibriSpeech dataset, consisting of audio-recorded books by volunteers, is massive. Data from 20 speakers was selected, representing a chapter from 20 books. After an examination of face-to-face social networks [169, 170], it was found that the average number of people met in daily lives is about 20, leading to the decision to create a multi-speaker dataset for this study instead of using just one speaker. A subset of the speech data was created with a 30-minute audio file from each speaker. The audio files from the LibriSpeech Corpus are 30 minutes long, each with a sampling rate of 16 kHz.

Some adjustments were made to the speech sound levels in our datasets. The speech dataset of speakers reading a book in front of a microphone will be used. As a result, the sound level of these recordings varies from those recorded in an acoustic environment where the microphone may be placed near a TV or on a side table away from the speaker. There are also complications in a home environment, where the occupier speaking might face the other direction while looking outside the window. These are factors to consider while developing these datasets.

The decision was made to synthesise speech data with the proposed datasets in order to incorporate speech data at various speech volume levels. Overall, speech data will be synthesised with both the ESC-50 and SINS datasets. The speech volume levels that are being incorporated include:

- 40 % speech volume level
- 60 % speech volume level
- 80 % speech volume level
- 100 % speech volume level

To prepare the audio dataset for this part, PyDub [171] was used, a Python library that allows us to create a script to load each audio file, an audio file from the background dataset and an audio file from the speech dataset, alternate them accordingly and then export the edited audio file into a selected directory.

First, a list of directories was created that needed to be populated with audio data. Our focus is to keep the quality and length of the background files intact and alternate accordingly the speech audio files. Here is a list of the directories that were created that correspond to our 2 suggested datasets:

- SINS
  - SINS with Speech (40% sound level)
  - SINS with Speech (60% sound level)
  - SINS with Speech (80% sound level)
  - SINS with Speech (100% sound level)
- ESC-50
  - ESC-50 with Speech (40% sound level)
  - ESC-50 with Speech (60% sound level)
  - ESC-50 with Speech (80% sound level)

  - ESC-50 with Speech (100% sound level)

Our next step involves utilising a Python Script that sequentially loads one file at a time from the background dataset, as well as a file from the speech dataset. To ensure variety, the speaker selection for each file is randomised, allowing us to include all 20 speakers from every audio file in our speech dataset. Using PyDub, a Python library, it is possible to crop the speech audio file according to the length of the background file. Additionally, a record of the speech audio files was kept that have already been used to avoid any duplicate overlaps with other background sounds. Once the audio files were synthesised, three more synthetic audio files were created, each with 80%, 60%, and 40% sound levels, respectively, based on the overall sound level of the specific audio file. Finally, all four alternations were exported to their appropriate directories.

Chapter 4

Tracking daily routines of elderly users through acoustic sensing: An unsupervised learning approach

4.1 Introduction

There is a growing demand for the development of assistive technologies that can provide support for elderly people living with dementia. As the specific condition can progressively lead to a significant deterioration of people's ability to function without support, there is great value in the design of a system that can monitor their condition unobtrusively and alert carers when there are signs of significant cognitive decline.

A vital sign that can indicate deterioration of cognitive abilities for people with dementia is a progressive difficulty following their typical daily routines, where specific regular tasks, such as having a meal, can be skipped or repeated in short intervals. In this work, the design of a system for the passive tracking of daily activities is explored to detect diversions from regular routines.

The well-documented limitations of deploying wearable technologies to support people living with dementia [55] are considered in this work. The feasibility of employing acoustic sensing technologies as the primary modality for monitoring daily activities is explored. The wide adoption of smart assistant devices like Alexa [172] and Google Home [54] has motivated the development of similar products tailored for elderly people (e.g., MiiCube [6]). These digital technologies provide the opportunity to utilise sound as a rich sensing modality for providing assistive applications. The feasibility of using acoustic sensing to identify diversions from the typical daily routines of elderly people with dementia will be examined. The significant challenge of collecting datasets with appropriate labelling of the activities performed is acknowledged, and an unsupervised approach to tracking daily routines through sound is investigated.

This preliminary work presented an architecture for the unsupervised tracking of daily routines through sound and detecting changes in the regular sequence of activities or the “skipping" of particular activities as part of the daily routine. The general idea behind our approach is to consider a small set of sound sequences as representative patterns of typical activities. The detection of diversion from these patterns is performed through the combination of mapping sounds into multi-dimensional embeddings through the use of a pre-trained model (VGGish [118]), the reduction of the dimensionality of the produced embeddings through a novel approach presented in this chapter and the application of Dynamic Time Warping as a pattern matching technique. This chapter presents preliminary system performance results, evaluated through synthetic data using public datasets of sounds within domestic environments and data collected through a controlled study with healthy participants. Using the public dataset, the system achieves 74% precision and 93% recall and using our collected dataset, the system obtains 99% precision and 95% recall.

4.2 Motivation

By 2020, 50 million people were estimated to be affected by dementia worldwide [173]. Furthermore, the number of cases is rising by about 10 million annually [174]. In the face of these projections, there is an increasing need to provide technology that can observe their daily lives and the progress of their condition while reducing the need for unnecessary involvement of carers.

In this work, the feasibility of using voice assistants within the homes of people with dementia is considered to detect diversions in their daily routines, which could be seen as a sign of cognitive decline. Considering example technologies like MiiCube [6], there is an opportunity to work with large acoustic datasets collected from the living environments of older users. However, the nature of these deployments makes it highly challenging to collect any ground truth information about the specific activities people perform.

The main objective of this work is to develop a system that can "learn" the typical activity patterns of users in an unsupervised way, using sound. Although the typical day of any person is not predictable, there are certain times that most people follow regular routines, for example, morning routine, mealtime, or bedtime routine. The proposed system is viewed as a passive sensing tool to detect diversions within those regular routines. The system should identify diversions from the "typical" pattern by detecting skipped activities not performed in the typical order. Using only sound signals as input, the main idea is to transform acoustic signals into a low-dimensional time series indicating when the user switches from one activity to the next.

The proposed system does not need to identify the exact activities involved and, therefore, does not rely on training over labelled data. Instead, by transforming acoustic signals into low-dimension time series, to apply time series pattern matching techniques like Dynamic Time Warping (DTW) [175]. DTW is an algorithm

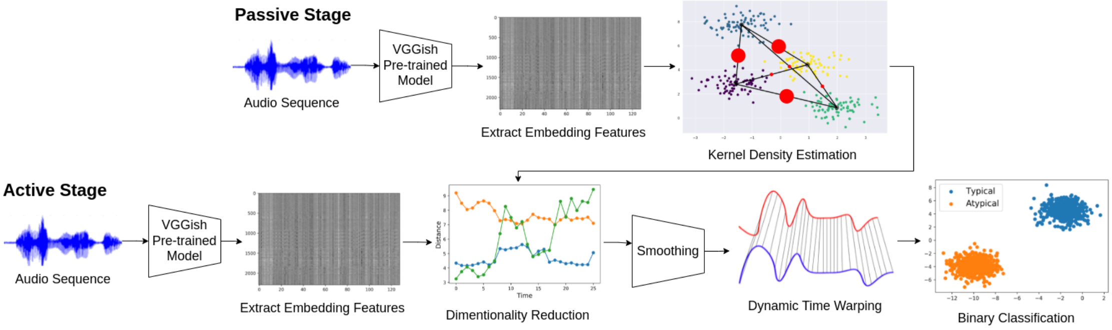


Figure 4.1: Pipeline: The process aims to analyse a stream of sounds captured over a long time window where the user performs a typical routine (e.g. morning routine) and compares the sound patterns with previously captured "normal" routine for that individual. The process does not rely on identifying the specific activities and, therefore, does not require labelled data for training.

that compares two temporal sequences that might not match perfectly, for example, detecting the walking pattern through accelerometer data at various speeds. DTW is commonly applied in gesture recognition [176].

4.3 Methodology

The overall architecture of the proposed system can be found in Figure 4.1. The system operates in two stages: the *passive stage* aims to collect acoustic data over a period of time in order to "learn" the sound characteristics of typical routines for each user; the *active stage* is where the system analysing acoustic data aims to detect diversions from the typical routines. The general pipeline involves extracting features from the acoustic signal, applying a technique for dimensionality reduction, smoothing, and a final stage of pattern matching using DTW.

The proposed system was developed and evaluated using two different datasets. The Freesound Audio Tagging 2019 dataset [164] is a large public dataset of a range of sounds, typically used for Acoustic Scene Classification (ASC) tasks. The dataset is labelled with ground truth, but the labelling is only used in the evaluation process. In addition to the public dataset, the system was evaluated through a small-scale collection of real-world sounds, described in 3.

4.3.1 Feature Extraction

The approach followed for extracting features from acoustic signals is influenced by typical methods applied for acoustic scene classification [133]. Traditionally, features on sound signals are extracted by transforming the signal into the frequency domain. However, since the publication of the VGGish model in 2017 [118], the acoustic sensing community has increasingly explored the application of transfer learning to extract sound features that can be used for acoustic scene classification.

The VGGish model is a pre-trained CNN network that has been trained over the AudioSet [123] dataset. The model takes as input frames of 975ms of audio data, converts them into a spectrogram, applies the Mel-Frequency Filter Banks, and uses this as input for training for acoustic scene classification. Applying VGGish for direct activity recognition in new environments can produce poor results, as the model is trained over a fixed set of labels. Instead, the VGGish model is used as a feature extraction method [118] by stripping the last layer of the model and utilising the generated 128-dimensions feature vector produced by the model as a multi-dimensional embedding representing features that can be used for acoustic sensing.

In this work, a similar methodology for feature extraction was employed. Our assumption for the VGGish produced embeddings is that the model is trained to provide high discrimination between sounds of different activities and closer similarity between sounds of the same activity. A preliminary investigation through the AudioSet dataset has validated that assumption. Indeed using Euclidean distances over the embedding vectors between audio samples of similar activities (e.g. cooking) are significantly smaller than the distances across activities (e.g. cooking vs bathroom). Significant variations with respect to the generated embeddings were also noticed for certain sound types. Indeed, sounds within the same activity can vary over short periods. As the VGGish model operates over frames of less than a second, these variations are manifested as changes in the embedded vectors produced for the same activity.

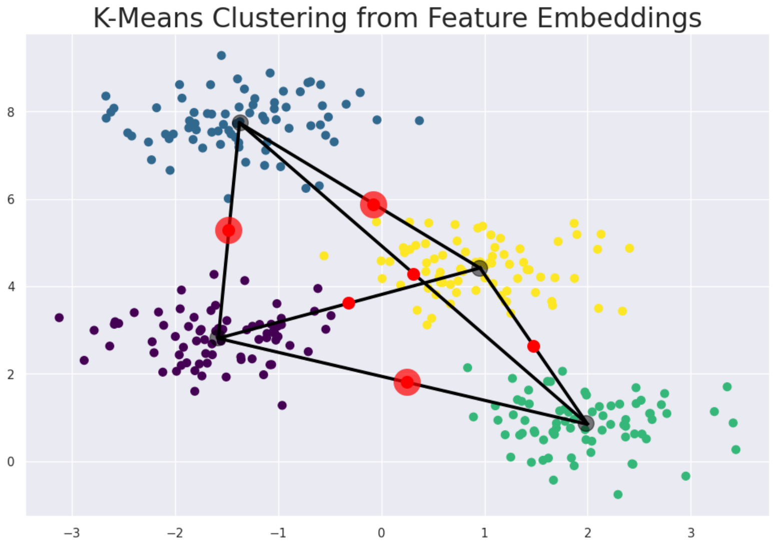


Figure 4.2: Identifying Reference Points (illustration): K-means were applied over the dataset of domestic sounds. The red dots represent the mid points between cluster centres. The 3 points with the lowest density were selected.

A smoothing function was incorporated over the generated embeddings to address the issues with high variability within the sound embeddings before they were forwarded to the next stage. Specifically, long activities (such as cooking or eating) are identified by aggregating the produced embeddings over a window of 1 min. Essentially, the pipeline operates over a 1 min sliding window (50% overlap), where the VGGish embeddings are averaged to produce a single 128-dimension embedding for each 1 min window.

4.3.2 Dimensionality Reduction

DTW is a common technique for pattern-matching time series. However, it is not well suited for time series with a high number of dimensions, as the complexity increases exponentially with each additional dimension. DTW is commonly applied to series with up to three dimensions.

DTW was applied in one or multiple dimensions by other researchers in the past. When it comes to multi-dimensional spaces, DTW could be computationally expensive. It is noted in the literature [177] that applying DTW with two dimensions achieved better results than the conventional methods. Another observation from

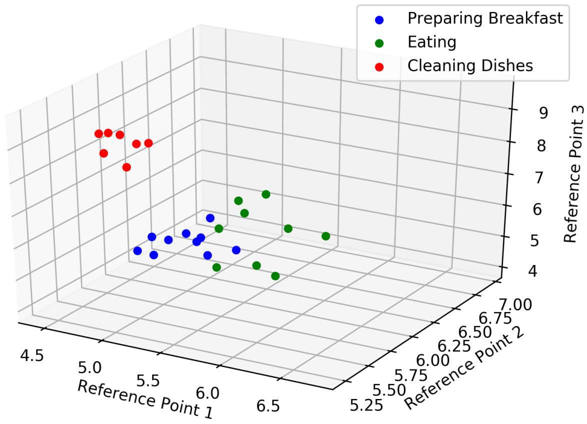


FIGURE 4.3: 3D visual representation of distances between reference points and activities. Each colour represents one activity. Each data point on the plot represents the distance in 3D from the reference points and the activity, showing that each activity can be differentiated from other activities.

the literature was applying DTW in computer vision tasks using the three dimensions of an image (Red, Green and Blue) [178], or three dimensions extracted from the MFCC coefficients from the audio signal [178, 179]. Three dimensions were chosen for the dimensionality reduction process in this work based on observations from the literature, which indicates they would be sufficient for the task without causing computational difficulties.

As DTW is intended to be applied over a time series of sound signals, the transformation of the 128-dimensional embeddings into a significantly reduced dimensionality vector is deemed essential. Following the observation made regarding Euclidean distances between embeddings, a small set of fixed reference points within the 128-dimensional embedding hyperspace is considered for use, with the distances from these reference points being utilised as a new lower-dimensional feature vector. In the proposed system, a reduction of the dimensionality of the embeddings into 3-dimensional feature vectors is aimed. Each value within the 3D vector should be comprised of the distance of each sound sample from three identified reference points within the embedding space.

Selecting the appropriate reference points is essential to ensure that the calculated distances will still have a discriminatory effect in separating sounds of different activities. This is a key part of the "passive stage" of the overall system. Over the "passive stage" a range of domestic sounds are collected over multiple days. As the VGGish model is trained over a larger range of sounds (including a variety of outdoor acoustic scenes), the distribution of domestic sounds within a specific household was expected to be relatively sparse when embedded into the 128-dimension space. Therefore, the aim was to identify three reference points within that sparse space to help differentiate activities through their Euclidean distance.

Our approach for selecting the reference points is based on the following rationale:

- Reference points should not be within dense areas of sound activity in the embedding space: The rationale is that a dense area can contain representations of a specific activity, and a reference point within that space can have a highly discriminatory effect for sound samples of the same activity.
- Reference points should be "near" the areas in the embedding space where domestic sound activity is located: This way, estimated distances from sounds of different activities can have a more discriminatory effect.
- The three reference points should be far from each other: It is crucial to minimise any correlations between the three dimensions produced.

The approach that satisfies these requirements is the following: A set of domestic sounds collected during the "passive" stage was utilised. The VGGish model transforms the dataset into a 128-dim data series of embeddings. A sequence of k-mean clustering operations was performed over the embeddings of all the domestic sounds of the household, with $k \in \{3, 4, 5, ...\}$. k cluster centres were produced for each clustering operation, which is assumed to be within dense areas in the embedding space. Using the cluster centres, the midpoint within each pair of them was calculated: a total of $k(k-1)/2$ candidate points for each clustering (Figure 4.2).

These points were considered candidates that can be "near" the areas of activities but potentially within low-density space. The density for each candidate point was calculated using the Kernel Density Estimation (KDE) function fitted over the dataset of domestic sounds. Finally, the 3 candidate points with the lowest density were identified as potential reference points.

This process will generate a triplet of points for each clustering operation. The last step is to select the triplet with the maximum distance between the candidate reference points. For each triplet, the surface area of the triangle produced between the three points was calculated as an estimator of the distance between them. The final selection includes the triplet produced by one of the clusters with the highest surface area.

The identification of the three reference points allows the application of dimensionality reduction over the 128-dimension embeddings. Specifically, it is possible to calculate the Euclidean distance of each sound sample from the three identified reference points. For example, in Figure 4.3, you can see a sample of the distribution of three domestic activities after dimensionality reduction. Although there is a significant loss of information from the reduction to a 3D vector, the proposed approach can still maintain the discriminant characteristics of the specific features.

4.3.3 Pattern Matching with DTW

Dynamic Time Warping (DTW) [180] is an algorithm that calculates the dissimilarity or distance between two time series while allowing the warping (compression or expansion) of the time axis in order to find the best alignment of the two series. Specifically, DTW calculates the distance between each possible pair of points within two time series. Through these, the cumulative distance matrix is calculated, and the ideal warping path is identified that minimises the distance between the two series. When working with multi-dimensional time series,s the multivariate DTW [176] algorithm has been successfully used for gesture recognition using 3D-activity time series.

In this work, the use of DTW was considered a form of identifying activity sequences similar to the typical routine of the specific user, and flag sequences were considered atypical. After the dimensionality reduction process, a time series of sound signals mapped into a sequence of 3D vectors can be used for activity sequence pattern matching. Different sequences of daily activities can be used to identify the “typical” daily pattern of activities. Similarly, new sound sequences can be classified using DTW based on their similarity with the typical activity patterns.

When monitoring a particular “routine” of activities (e.g., morning routine, bedtime routine), a set of sound sequences collected over a few days was considered the user’s typical pattern and is set as the *training* set. Calculating DTW between the sample of sound sequences from the *training* set allows us to estimate a threshold θ as the maximum DTW distance between sequences of the training set. When tracking new sound sequences, the average DTW distance of that sound sequence was calculated with all sequences of the *training* set. If the average distance is lower than θ the particular sequence is considered “typical”, and when the distance is higher it is considered “atypical”.

4.4 Results

The performance of the proposed model is evaluated using two datasets: the public Freesound Audio Tagging 2019 dataset [164] and datasets collected through the AudioHive app [168] by 10 users over 5 days. The Freesound Audio Tagging dataset contains 297,144 samples of audio data, which are accurately labelled with the activities they represent. For this work, only a subset of the dataset containing domestic sounds was selected. Similarly, the AudioHive dataset was manually labelled by the participants, with activities representing their morning routines.

4.4.1 Synthetic Sequences

In order to evaluate the model, a sufficient dataset of both "typical" and "atypical" sequences of activities was needed. Such sequences were synthesised by combining data from the original datasets and stitching sound samples of different durations from various activities.

In particular, three new datasets were generated by stitching activities from the AudioHive dataset and three new datasets from the Freesound Audio Tagging 2019 dataset.

- Typical sequence: A sequence $S_t = a_1, a_2, ..., a_n$ of n activities selected from the complete set of activities within the dataset, to represent a typical set of activities performed in a household, e.g. preparing breakfast, cleaning dishes, etc. For each activity a_i within the sequence, the distribution of each duration as recorded by our participants was estimated. Based on this distribution, samples of variable length within the range of two standard deviations from the mean duration for each activity were generated.
- Reordered sequence: Using the same S_t, random re-orderings of the set of activities were produced. These sequences contain the same activities as those in the typical sequence but in randomly mixed order.
- Missing activity sequence: Using the S_t sequence generated a set of activities where one of the a_i activity is removed from the sequence.

Through this process, 30 acoustic signals of “typical” sequences, 30 re-ordered sequences representing the ”atypical” set, and 30 sequences with a missing activity were generated using the public dataset. Next, the equivalent was done with the AudioHive dataset; 30 acoustic signals of “typical” sequences, 30 re-ordered sequences, and 30 missing activities using the AudioHive dataset were created.

For the training of the DTW algorithm, a random subset of 10 typical activities was selected to estimate the acceptable range of DTW distances to classify a sequence as “typical”. The validation set consists of the remaining 20 “typical” sequences and 60 “atypical” sequences.

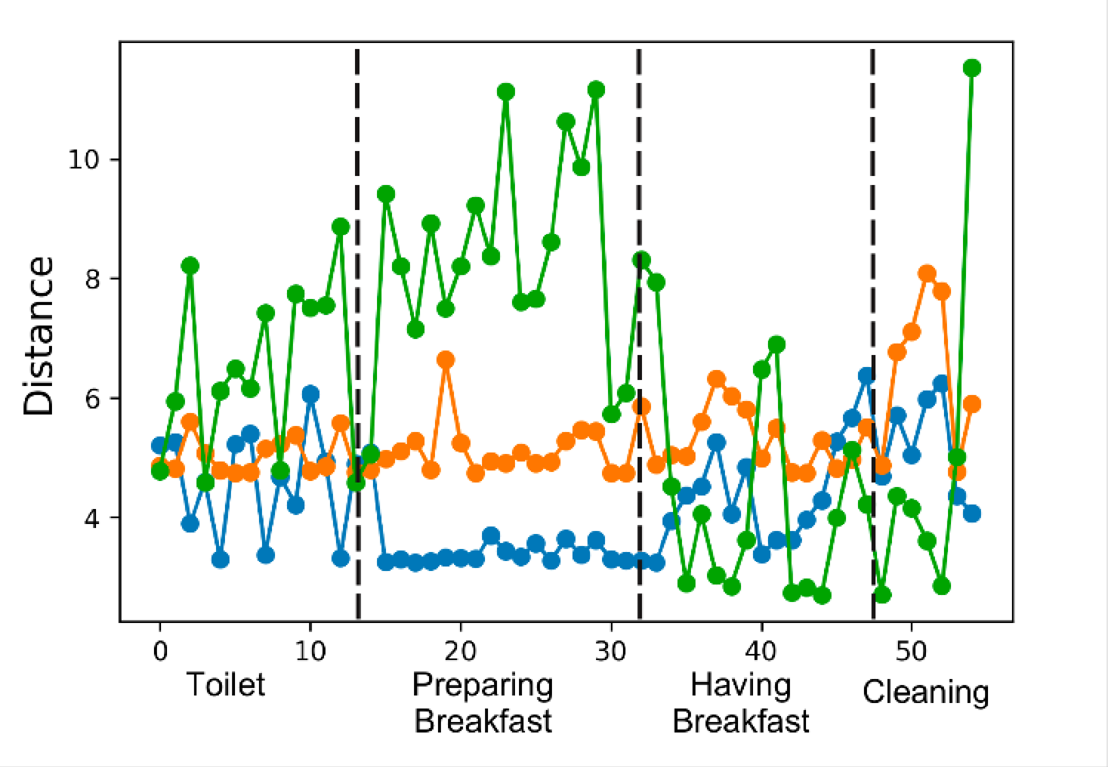
The “passive” stage of the system involves identifying appropriate reference points for dimensionality reduction. These are selected for each environment and the set of domestic sounds in each dataset. Table 4.1 illustrates the outputs of that process over a series of clustering steps. For each of these, the midpoint between all cluster centres was calculated, selected the three midpoints with the lowest density, and calculated the surface area between them. In this case, midpoints from $k = 7$ cover the largest surface area.

Table 4.1: Results of the selection of reference points using clustering. k=7 generates points with the widest distance between them.

Clusters (k)	Avg. Density 3 midpoints	Surface area
k = 3	20.50	1.47
k = 4	48.25	1.71
k = 5	11.24	2.52
k = 6	34.17	1.43
k = 7	8.027	2.73
k = 8	-11.57	1.10
k = 9	-12.78	1.22
k = 10	20.50	1.08

During the “active” stage, the system uses these reference points to reduce the dimensions of any 1min sound sample. Figures 4.4, 4.5 shows samples of the produced 3-d vectors for different sequences. It can be observed visually that sequences of the “typical” pattern demonstrate similarly shaped time series, whereas those of “atypical” behaviour do not. This indicates that the time series transformation into a 3D model produces suitable results for DTW pattern matching.

The algorithm using the *validation* set was evaluated, consisting of 20 correct, and 60 wrong activity sequences for each dataset. The results are shown in Table 4.2. The results are also compared with the effects of a similar system that relies on a more traditional dimensionality reduction technique using principal component analysis (PCA). As shown the proposed algorithm can achieve very high performance. It was noted that the performance is particularly high for the real-world collection through participants. The reason for this high performance is that the “passive” stage analyses the patterns of the sounds that are generated within each household. This leads to a transformation that is tailored to the sound patterns produced by each participant. Instead, the public dataset consists of activity sounds from a range of different environments grouped together.



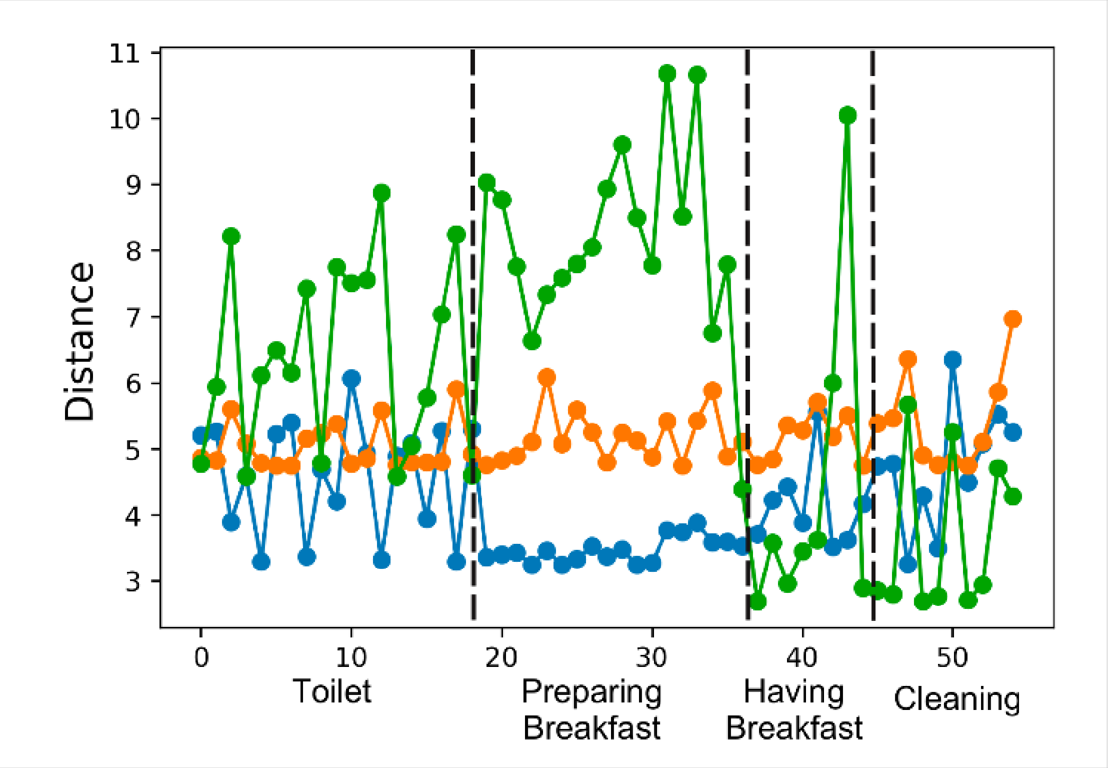
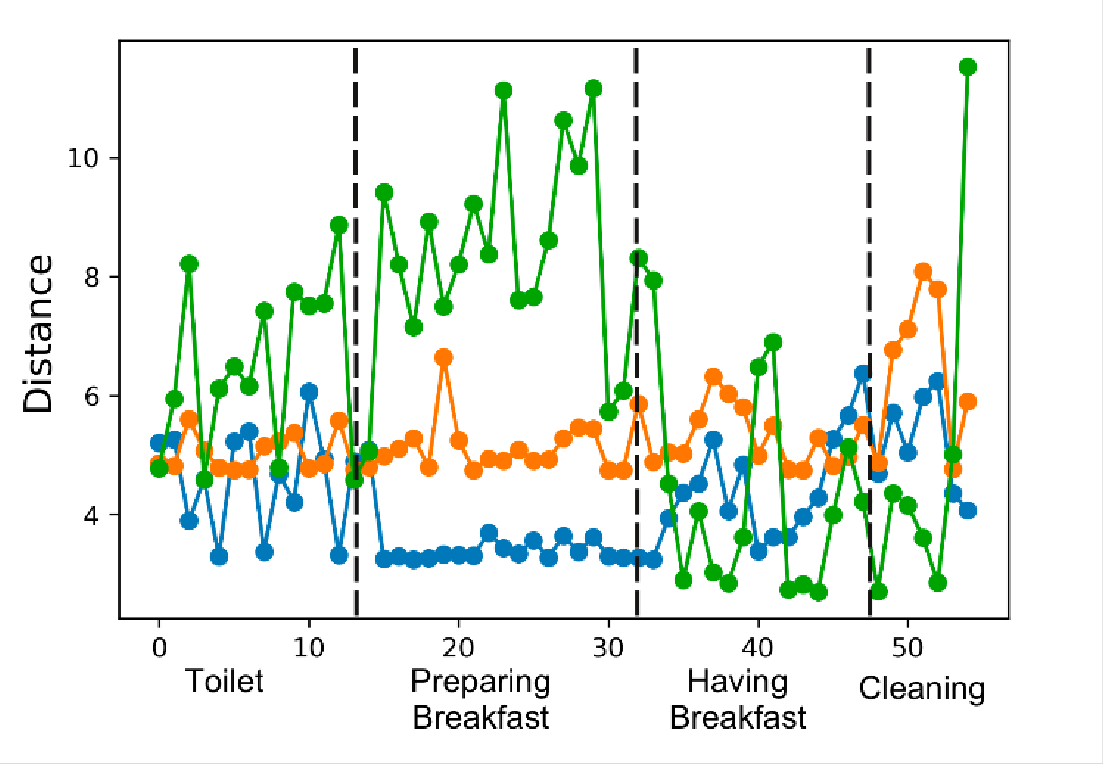


FIGURE 4.4: Visual representation of distances from the three reference points and audio samples from collected data. The three reference points and centroids from new data were set from the extracted embedding features and dimensionality reduction. For both sequences, the classes are known and labelled. This could also be done in an unsupervised manner but labels were available at this point. Both sequences have the same pattern of activities and it is assumed that with the use of Dynamic Time Warping (DTW) it could be detected as a "typical" sequence of activities.



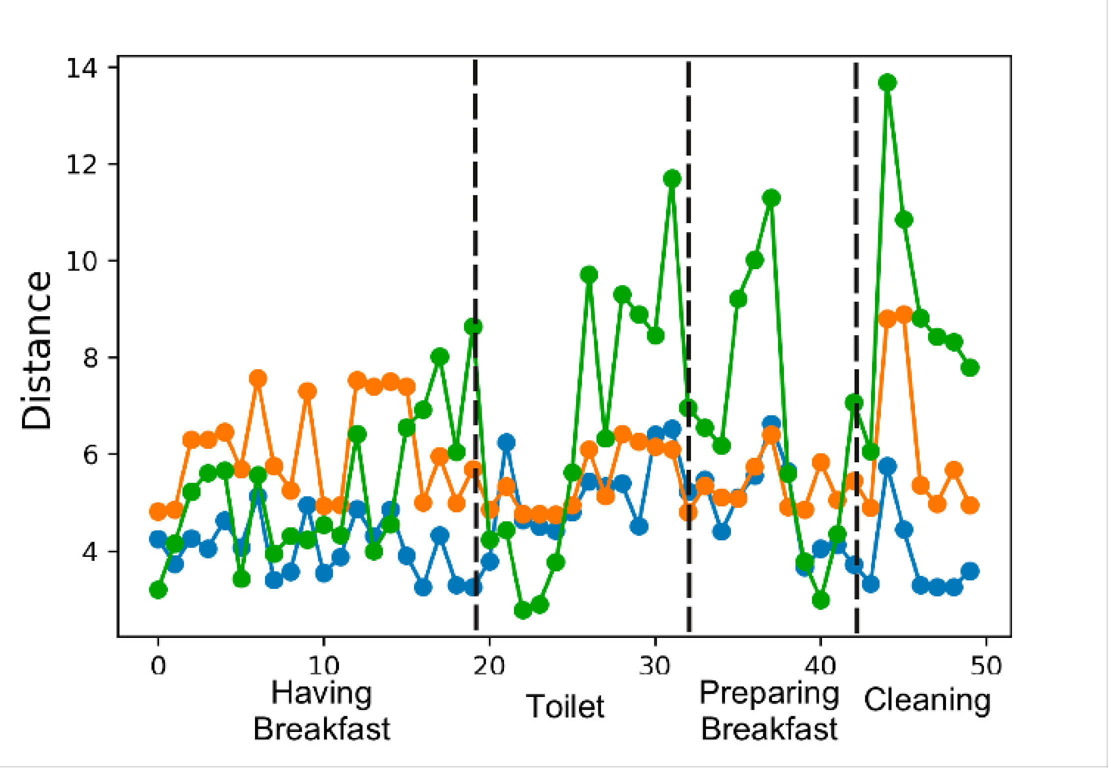


FIGURE 4.5: Visual representation of distances from the three reference points and audio samples from the collected data. The three reference points and centroids from new data were set from the extracted embedding features and dimensionality reduction. For both sequences, the classes are known and labelled. This could also be done in an unsupervised manner but labels were available at this point. In this occasion, the activity sequence is different, the labels are provided and they are visually dissimilar. The assumption here is with the use of DTW will distinguish the sequence as "atypical" or heterogeneous.

Table 4.2: Performance of proposed clustering method, and comparison with baseline (PCA)

Method	Precision	Recall
Clustering (AudioHive Dataset)	99%	95%
Clustering (Public Dataset)	74%	93%
PCA (Public Dataset)	59%	80%

4.5 Conclusions

This chapter presents a novel technique for the unsupervised tracking of changes in daily routines using acoustic sensing. This work aims to develop a system that can detect significant changes in the daily routines of people with dementia.

The proposed system relies upon the VGGish model to generate embeddings of sound samples. A novel dimensionality reduction technique transforms the acoustic signal into a 3D time series of features. DTW is then used to match different patterns of activity sequences. The system's evaluation through synthetic data achieves a precision of 99% and a recall of 95%.

This research phase will explore the potential privacy challenges associated with an acoustic activity recognition environment. While a system that can detect changes in an acoustic environment can be beneficial, specific challenges are also posed. Privacy, in particular, is considered a crucial issue to be investigated. It is understood that privacy concerns can be detrimental to the overall acceptability of the system from the user's perspective. Therefore, this aspect will be thoroughly investigated to ensure the system is helpful, fair, and safe for everyone involved.

Chapter 5

Privacy Challenges for Acoustic Activity Recognition

5.1 Introduction

Acoustic sensing in smart homes offers benefits like monitoring and anomaly detection, helping to enhance security by detecting unusual sounds, such as glass breaking [7]. However, privacy concerns arise with continuous audio recording, particularly regarding private conversations. There is a need for effective anonymisation techniques and privacy-preserving algorithms to safeguard data while utilising the benefits of the acoustic context. Anonymisation techniques involve removing or altering personal data so individuals cannot be identified, while privacy-preserving algorithms are designed to protect sensitive information. As highlighted by Corti et al. [181], it is essential to balance data collection with the occupants' right to privacy and freedom of speech.

5.1.1 Case Study

The case study in appendix 8 highlights a company that uses voice-activated devices and multi-sensory data for elderly care. The team encountered both challenges and opportunities in smart home deployments. Their project, ADAPTIVE (AI-based Dementia Assistive & Passive Technology for Non-Invasive Elderly Care), utilizes audio sensing to monitor changes in occupants' gait.

5.1.2 Planning Our System

As a first step in exploring the privacy concerns derived from real-world audio sensing, a baseline privacy-preserving system was aimed to be developed. The proposed system is designed to detect speech audio signals from an audio recording and to remove them completely. Based on the case study discussed earlier, it is noted that labelled data are limited and hard to collect in a real-world deployment; thus, public datasets and synthetic data will be depended upon for this system. A small and limited public dataset is to be worked with. The proposed system will focus on detecting speech segments, muting those segments, and conducting audio event detection, as shown in figure 5.1, to showcase the audio classifier's ability after the data is obfuscated. The ADAPTIVE project eventually adopted this approach to address the challenges regarding the deployment of conversational audio data recording.

However, a disadvantage of this system is that it may also remove any background sounds along with the speech audio signal. The system aims to identify any decline in the overall Acoustic Activity Sensing after removing chunks of the audio due to speech privacy concerns. This is assumed to be the basic system that can initially solve the Privacy Regarding Speech in a smart home environment.

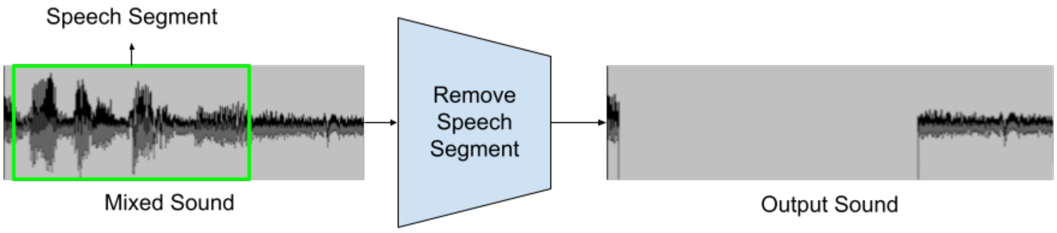


Figure 5.1: Plan of our System

The proposed system will follow these steps:

- Detect of Speech Segment
- Mute the sections that contain speech so there is no detectable speech in our dataset
- Audio Event Detection

5.2 Challenges and Approach

As previously mentioned, the challenge is posed by the difficulty of collecting data due to limitations in the environment. Cameras or human observers cannot be relied upon to label audio data, so Public Datasets must be used to synthesize a simulation of a smart home device's environment.

The data synthesis process takes into account the typical setup of a smart home, where the speaker and microphone are often located far apart. In a home environment, the speaker may not be close to the microphone or may even be positioned on the opposite side of the room. Therefore, four datasets have been created that utilise varying volume levels of speech audio. The data was synthesised to mimic a real-life smart home setting.

Here is a list of the proposed datasets:

- ESC-50 with added Speech (40% speech volume level)

- ESC-50 with added Speech (60% speech volume level)
- ESC-50 with added Speech (80% speech volume level)
- ESC-50 with added Speech (no alternations to speech volume levels)

Typically, the audio signal is converted to the frequency domain for audio sensing, and features are extracted from the frequency representation. The usual approach to training an audio classifier is to convert the audio waveform of t seconds to a spectrogram representation. For this task, a log-mel spectrogram was used, which represents the sound frequency regarding how a human ear perceives sound. The log-mel spectrogram offers a superior resolution for lower frequencies compared to the spectrogram. This is due to its use of the mel frequency scale. Mel-Frequency Cepstral Coefficients (MFCC) is another option using frequency representations for sound, which is used in speech or voice recognition tasks [182] and other acoustic recognition tasks [182]. MFCCs are coefficients which represent the spectral characteristics of an audio signal. They are obtained from the mel spectrogram but are processed further to extract significant information.

Using the spectrogram representation, a deep learning approach can achieve great results in audio classification. Deep learning methodologies were widely utilised for the acoustic sensing tasks from the annual Detection and Classification of Acoustic Scenes and Events (DCASE) competition. DCASE [143] is a community of researchers which announces yearly, with its first edition in 2013, several acoustic sensing-related tasks alongside public datasets for researchers to use. In recent years of this competition, Convolutional Neural Networks (CNNs) have been utilised for audio classification and have been very successful [133]. CNNs were combined with Recurrent Neural Networks (RNN), also called CRNNs and produced state-of-the-art results [183]. The CRNN model was trained with TUT Sound Events Synthetic 2016 [79] consists of approximately 9 and a half hours of audio data.

However, the recent literature shows an emphasis on the use of Attention Layer within their Deep Learning model. After the publication of the paper Attention is All You Need [149], more researchers are implementing the proposed new technique that also works for Audio Classification [184]. The attention mechanism functions in a similar way to how humans focus their attention on a single sound or activity amidst many sounds/activities in a room. The attention layer processes the entire audio sequence at once, and the model computes an attention score for each element of the input sequence.

Transfer learning is a machine learning technique that involves using knowledge gained from a pre-trained model to help improve the performance of a model on a different but related task. This technique has become increasingly popular in various tasks, especially when only a small amount of domain-specific data is available and where general data can be used from a public dataset. With transfer learning, a pre-trained model can be fine-tuned on a new dataset, reducing the need for large amounts of labelled data and improving the model's overall performance. This technique has been successfully applied in various domains, including computer vision, natural language processing and speech recognition [185].

A common approach for acoustic scene classification is utilising a pre-trained model for generating embedding features. For example, VGGish is a commonly used pre-trained model [118], which can extract a 128-D embedding feature from each 0.96-second audio signal sampled at 16 kHz. The embedded extracted features could then be used to build a machine-learning model for acoustic scene classification.

One final challenge was preserving privacy, given the system's location and continuous recording within a residential environment where people live and interact. Concerns regarding the confidentiality of occupants and their family members or carers were raised. To mitigate this issue, an approach was developed to detect and eliminate recorded audio sections containing speech. While this method effectively removes potentially sensitive content, the accuracy of the audio classifier may also be impacted. This chapter explores the consequences of this approach and determines its efficiency.

5.3 Methodology

5.3.1 Overview

The real-life scenario of smart care homes does not allow for a large labelled dataset to be collected, and the gathering of labelled data is often time-consuming, expensive, and sometimes infeasible. Public datasets have been created by other researchers and can be utilised for this study. Additionally, our own datasets were synthesised using two public datasets, where speech was synthesised with background audio to illustrate the complications associated with smart home environments that involve continuous audio recording for audio-sensing purposes. To demonstrate the impact of this research, a combination of public datasets was used, followed by testing of the system on data collected using the smartphone app that was created, as mentioned in Chapter 3. The study aims to detect and remove conversation or speech sounds in a smart home environment, as shown in figure 5.2.

The pipeline consists of several steps, including synthesising the dataset, using a Voice Activity Detection algorithm to detect speech, removing those audio segments that contain speech, and then analysing the remaining audio signal using an Acoustic Activity Detection or other audio analysis techniques.

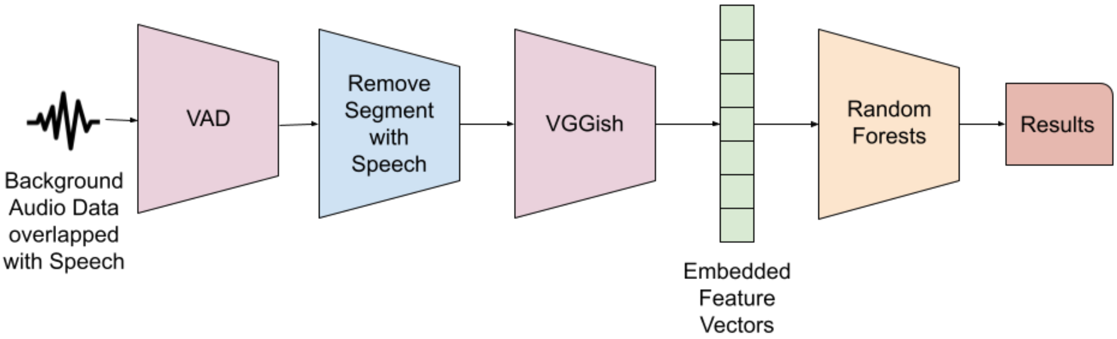

Background Audio Data overlapped with Speech
VAD
Remove Segment with Speech
VGGish
Embedded Feature Vectors
Random Forests
Results

FIGURE 5.2: Overview of this Methodology

5.3.2 Voice Activity Detection

Voice Activity Detection (VAD), also known as speech activity detection, is a technique used in speech processing to detect the presence or absence of human speech. VAD is one of the main pieces of the pipeline. Silero VAD [186] was used for this step as it showed impressive results on public datasets compared to other VAD models, like WebRTC [187] and SpeechBrain [154].

Silero VAD's architecture is based on convolutional neural networks and transformers. The model was trained on a large collection of audio recordings containing speech and non-speech segments from open SLR [188]. The authors implemented this PyTorch neural network with multi-head attention (MHA) and utilised the Short-time Fourier transform (STFT) as a feature. MHA-based networks can learn from multiple input representations simultaneously, making them extremely powerful in tasks that require integrating information from different sources. This has made them a popular choice for various applications such as natural language processing and computer vision. The authors believed that combining MHA and Short-time Fourier transform features would lead to a highly effective neural network, and their results proved promising.

Silero VAD [186] is a PyTorch model that supports audio inputs with sample rates of 8 kHz and 16 kHz. However, the model can also process audio inputs with sample rates of 32 kHz and 48 kHz by resampling them to 16 kHz through slicing. In addition to this, Silero VAD supports three different chunk sizes: 30, 60, and 100 ms. The model has been trained in more than 100 languages and has been shown to generalise well. Lastly, it's worth mentioning that each chunk of audio takes approximately 1 ms to process on a single CPU thread. The model outputs a probability between 0 and 1 to indicate if speech is present in the audio segment. The threshold for speech detection is set at 0.5. The threshold was selected by the creators of Silero because they did not have any reason to be biased in detecting speech or not, and a cut-off in the middle was a natural selection.

Audio data that may contain speech can be input using Silero VAD [186], and based on the output of the model, post-processing can be performed using Python to determine the timestamps of each section that includes speech. As shown in Figure 5.3, the total amount of speech detected and the exact moments at which it was detected will be identified through this process.

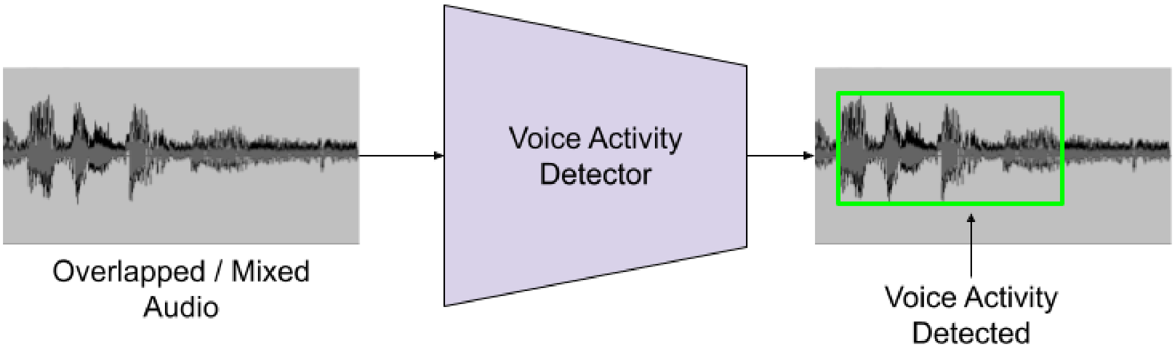


FIGURE 5.3: Voice Activity Detection

5.3.3 Remove Audio that Contains Speech

In the current stage, the output data generated by the Voice Activity Detection (VAD) model were utilised to eliminate the audio data containing speech without altering the overall duration of our audio dataset. This process is illustrated in figure 5.4. The audio signal output will be the same length as the original, but it will be removed if any audio has speech.

Silero VAD is trained on data with a sampling rate of 16 kHz. Silero VAD uses a 30 ms non-overlapping sliding window to predict whether there is speech in the audio. A window of 30 ms was found to be too short to contain any meaningful speech content. To ensure the accuracy of the pipeline, only audio where speech was detected by Silero VAD for at least 1 second was removed.

A Python library, the same one used to synthesise the audio data, called PyDub [171], was utilised to split the audio into three parts. The audio clip before the speech, the section containing the speech, and the remaining audio clip.

The speech section was silenced using PyDub, and finally, all three sections were stitched together.

This step is crucial in preparing our audio data for further processing, as it will remove any unwanted speech audio while preserving the privacy of the dataset regarding speech.

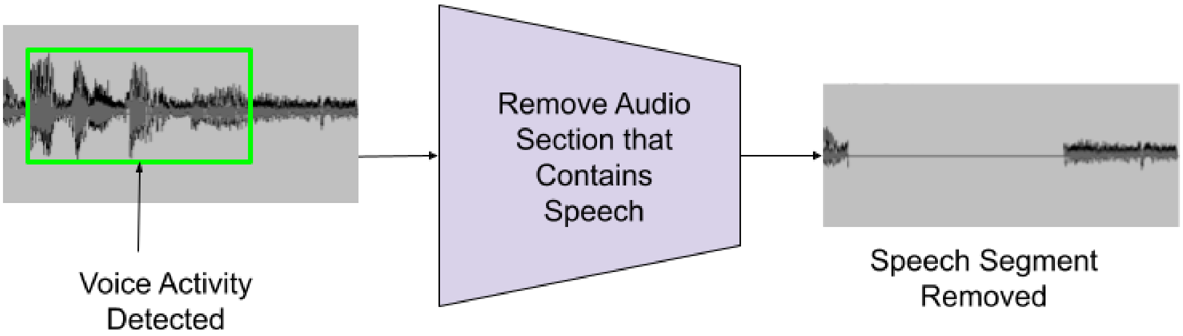


FIGURE 5.4: Remove Audio Signal that Contains Speech

5.3.4 Acoustic Event Detection

Based on our requirements for this system, acoustic activity classification techniques were utilised that do not demand a massive amount of labelled data [133]. With the advancements of the VGGish model in 2017 [118], the acoustic sensing community has been exploring transfer learning as a means of feature extraction for acoustic scene classification.

The VGGish model, shown in figure 5.5, is a pre-trained CNN network that has been trained on the AudioSet dataset [123]. AudioSet is a large YouTube dataset, a preliminary version of what later became YouTube-8M [189] where audio was extracted from YouTube Videos and was human labelled based on a generic set of labels. AudioSet Ontology contains 632 audio event categories with approximately 2 million audio files, ranging from less than a second to several minutes per audio file. The median audio file duration is about 10 seconds.

The VGGish model has a VGG-like architecture and uses stacked Convolutional layers with pooling layers. This deep model is capable of capturing increasingly

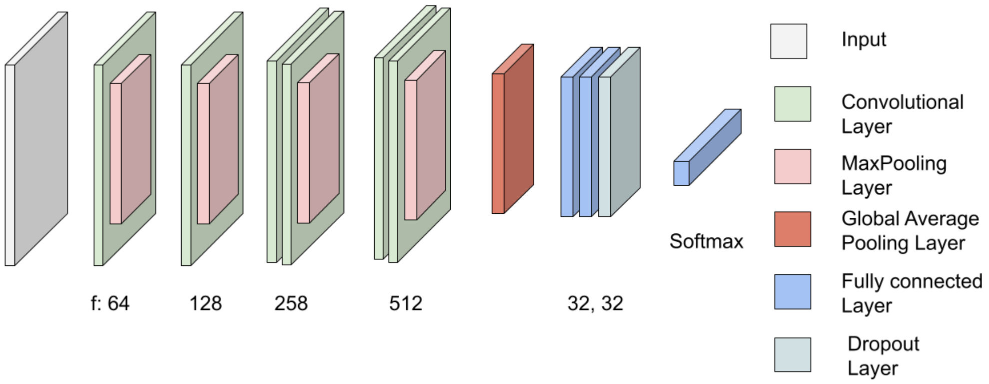


FIGURE 5.5: VGGish-based audio baseline model's architecture. The size of the feature maps (f) of each convolutional and fully connected layer are shown below each block of operations.

complex features from the input data. Initially designed for image recognition, VGG can be adapted to other domains like audio. Compared to previous models, VGG has a significant number of layers and depth. This innovation proved effective in learning robust features. Despite its depth, VGGish mostly uses small 3x3 Convolutional filters, achieving efficiency without overly complex units.

The VGGish model is adapted to work with spectrograms and is developed in Python using the Tensorflow framework [190]. It takes a spectrogram patch of 96 frames x 64 mel channels as input. This patch corresponds to 96 consecutive columns from a spectrogram that is calculated with 10ms frames (based on 25ms windows), which covers approximately 0.96 seconds of input per patch (0.975s if you include the whole of the 25 ms window at the end). Additionally, VGGish is a pre-trained network that was trained on these 96x64 input spectrograms.

Audio features were extracted from our custom datasets by using the appropriate spectrograms of size 96 x 64, with a sliding window of 25 ms overlap and a 10 ms step. Then, the output was used as input for the VGGish model, which returned a feature embedding vector of 128 dimensions per window of 0.975 seconds. The extracted feature set was used to train a Support Vector Machine (SVM) model. Other researchers commonly used the SVM model in audio classification, such as detecting COVID-19 from crowdsourced respiratory sound data [77].

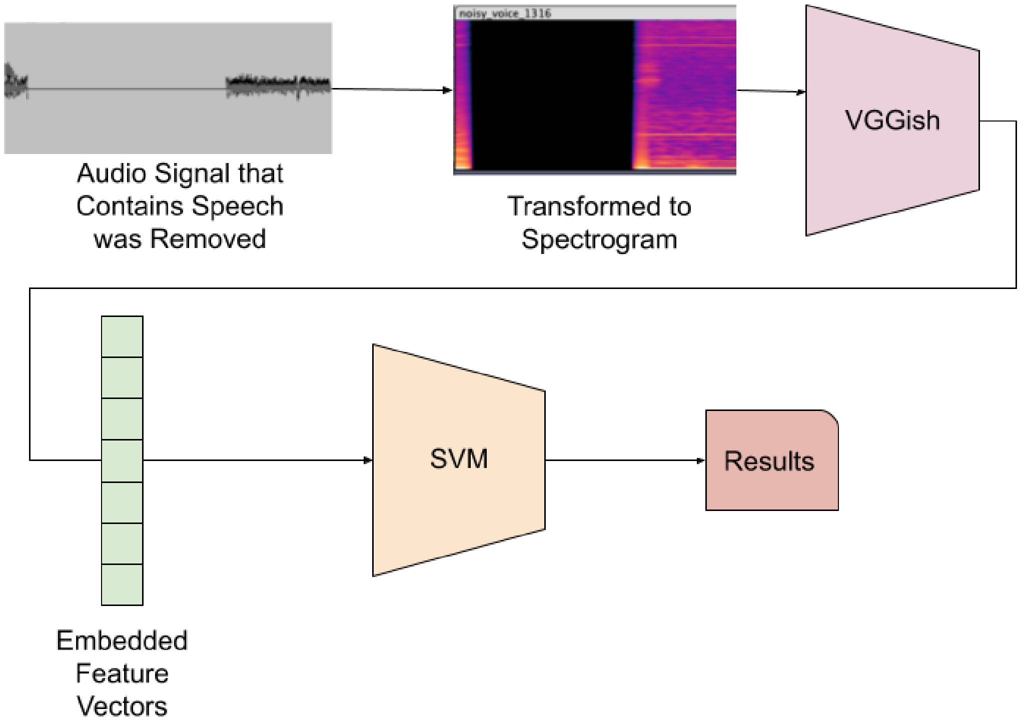

Audio Signal that Contains Speech was Removed
noisy_voice_1316
Transformed to Spectrogram
VGGish
Embedded Feature Vectors
SVM
Results

FIGURE 5.6: Acoustic Activity Classification on the Audio Signal without Speech

The extracted embedding features from the VGGish model were utilised to train the SVM model. The available data was divided into Train and Test data, with a 70/30 split, using Stratified Shuffle Split from SciKit Learn [191]. The purpose of this split is to have an even data size per class, which can be challenging to achieve with data collected from the wild or from a dataset like AudioSet Ontology.

The data was first normalised using L2 normalization, resulting in a range from 0 to 1. The Support Vector Machine (SVM) model was then trained using SciKit Learn and fine-tuned to achieve the best possible results. A parameter range, including C, $gamma$, $kernel$, and $degree$, was defined for this purpose. The parameters were tested using Grid Search, which refitted the model with the best parameters. Finally, the classifier was set up with the optimal parameters obtained from the grid search, where $C = 10$, $degree = 5$, $gamma = 1$, and a polynomial kernel was used.

In conclusion, audio classification with transfer learning techniques like VGGish and SVM is a promising methodology in the application domain where labelled data are scarce; for example, the training dataset was less than 2 hours of audio data. The accuracy of the models depends on various factors, like the quality of the data, feature extraction techniques, and the algorithm used for classification. With the help of modern tools like TensorFlow and SciKit Learn, researchers and developers can achieve high accuracy in audio classification tasks, which can be used for tasks like speech recognition, music genre classification, and environmental sound classification.

5.4 Results

The audio signal was transformed into spectrograms after the speech was removed for the proposed methodology. The 128-D embedding features were extracted using the VGGish pre-trained model. An SVM model was subsequently trained, and results were produced. The methodology, shown in figure 5.6, aims to detect speech in the audio signal, eliminate sections that contain speech, and apply audio classification to understand the effect of the methodology that seeks to preserve speech.

The Acoustic Activity Classification (AAC) and Voice Activity Detection (VAD) were used as metric tools to understand the benefits of our methodology. The results are presented in table 5.1 in the first two columns (Precision and Recall) alongside VAD before our methodology was applied and in the last two columns (Precision and Recall) after the methodology was applied. After the methodology was applied, VAD could not detect any conversation.

The acoustic event detection system that is used operates on audio signals that have undergone a pre-processing stage in which speech has been removed. As a result, some gaps in the audio signal may be observed, causing the classifier to label the audio segment as "silence". It is essential to consider this aspect when the system's performance is analysed and any potential shortcomings that need to be addressed are identified.

Table 5.1: Performance of AAD, VAD, and Proposed Methodology's AAD Results after the Speech Segments were Removed on ESC-50 Datasets Mixed with LibriSpeech.

ESC-50 Datasets	**Precision**	**Recall**	**VAD**	**Precision (Post-Proposed Method)**	**Recall (Post-Proposed Method)**
Clean (No Speech)	85%	83%	0.6%	84%	82.1%
Added Speech (40% speech volume level)	75%	72%	6.80%	50.01%	39.75%
Added Speech (60% speech volume level)	73%	71%	7.2%	40.97%	41.32%
Added Speech (80% speech volume level)	70%	68%	17.19%	29.12%	32.56%
Added Speech(100% speech volume level)	69%	65%	6 7.5%	41.32%	29.58%

Additionally, it should be noted that the speech section in the audio signal is removed to preserve the privacy of individuals present inside the smart home. This is an important consideration, as it ensures that sensitive information is not recorded or transmitted by the system while still allowing for the detection of relevant acoustic events.

When speech or conversation is present in the background audio signal, it can negatively impact the performance of the Audio Classifier. This is because the presence of speech can alter the sound spectrum, making it difficult for the classifier to identify and classify the audio accurately. However, if the speech sound levels are lower, there may be an improvement in the performance of the audio classifier. This is because, at lower sound levels, the background sound can dominate the sound spectrum, which can make it more distinguishable and easier for the classifier to identify. Additionally, when speech is at lower sound levels, the Voice Activity Detection (VAD) may not be able to detect the speech, which can also contribute to an improvement in the audio classifier's performance. Overall, the presence of speech in the background audio signal can have varying impacts on the performance of the audio classifier, depending on the sound levels of the speech and the ability of the VAD to detect it.

5.4.1 Evaluation with Collected Data

A data collection was conducted to evaluate the performance of our system. The primary objective was to demonstrate how the performance metrics of our system align with the results obtained from synthetic datasets. Data was collected in home environments where participants performed daily routines to achieve this goal. Participants were also encouraged to speak when their daily activities allowed, with the goal that speech would be removed by the proposed methodology.

The audio data collected from the 12 participants was stored anonymously in individual folders for each participant, along with a CSV file containing all the information about the activities being conducted and which participant was involved, based on the input data from the App. A Python script was then run to read the timestamps from the CSV files, checking if the audio files existed and if the timestamps had not been set by accident, for example, if they showed a duration longer than 1 second. It was also verified that the audio file type was WAV; M4A files created by the iOS device were converted to WAV, ensuring that all files were single-channel rather than stereo or multi-channel.

Afterwards, a new dataset was created using the labelled data; each audio file was represented as an audio class, e.g., shower, preparing breakfast. Once this process was completed, a manual check was conducted to identify any duplicate audio classes that were merged together. Additionally, any audio classes containing less than 1 second of audio data were discarded. All new audio files pertaining to the original participant were stored in the same folder. The aim of this pre-processing section was to establish a structure for the collected data similar to that of the ESC-50 dataset structure [127].

Once the collected data have been prepared for analysis, the proposed pipeline will be proceeded with, consisting of the following steps:

- Detect speech
- Silence the audio segments that contain speech
- Extract embedding features from the processed dataset
- Split embedding features into train and test datasets
- Deploy audio classifier on the embedding features

The aim of our pipeline is to eliminate any speech in a smart home environment and preserve the occupant's privacy regarding speech. Speech segments were detected using Silero VAD [186], and these segments were silenced using PyDub library [171]. Afterwards, 128-D embedding features are extracted using the pre-trained model VGGish [118]. VGGish has been trained on audio data with a sampling rate of 16 kHz and a duration of 0.975 seconds. Although our dataset also utilises a 16 kHz sampling rate, longer audio durations had to be employed for each embedding due to the nature of the data. The 0.975-second window is considered too short to distinguish the differences between some activities, such as washing dishes and having a shower. An attempt was made to train an SVM with features extracted from each 0.975 second, but good results were not achieved.

VGGish expects a spectrogram of 96 x 64 as input, which covers approximately 0.96 seconds of audio signal at 16 kHz sampling rate and returns a 128-D embedding feature. Our solution was to use an overlapping sliding window of 5 seconds of audio with a 1-second step, in which VGGish return five 128-D embedding features. The mean of those five embedding features was calculated, resulting in one 128-D embedding feature for a 5-second audio clip. The extracted feature set was then used to train a Support Vector Machine (SVM) model, and the results were produced.

Table 5.2: Performance for AAD (Precision and Recall) and VAD after removing the audio sections that contain Speech from the collected data.

Collected Data	**Precision**	**Recall**	**VAD**
Collected Data with Speech	81%	87%	15.2%
Processed by the Proposed Methodology	56%	55%	0%

The dataset collected through a smartphone app was used to evaluate the methodology. As shown in Table 5.2, the first row represents the Original dataset, where acoustic activity recognition was applied using an SVM and Voice Activity Detection (VAD) before processing. The second row details the results after the proposed pipeline was applied. An Audio Classifier achieved a precision of 81% and recall of 87%, while Voice Activity Detection identified 15.2% of the audio signal as containing speech from the collected dataset (without speech removal).

After the proposed methodology was deployed, speech from the collected data was diminished, indicating that the goal of eliminating any speech audio data that could raise privacy concerns for the occupants of the smart home was achieved. However, a disadvantage observed in the results, which were similar to those of the synthesised dataset, is that the proposed methodology unfortunately deteriorated the accuracy of the audio classifier. Any additional analysis of the processed data will likely have limited success because of the impact of the current speech removal method.

Our methodology aimed to remove speech and highlight the constraints of our existing pipeline. Silencing audio sections containing speech can protect the privacy of smart home occupants regarding speech. However, it also limits the system's primary objective of audio analysis.

Developing a system to leverage audio-sensing technology has been valuable in understanding the possibilities and limitations of this approach. However, given that continuous audio recording in private homes can make some individuals uneasy, it is clear that a solution that can filter out speech while preserving background noise will be critical in advancing the audio-sensing field.

5.5 Conclusions

The proposed methodology for removing speech from audio data that is recorded in a smart home environment found that the system can effectively eliminate speech. However, the process has significant shortcomings. According to the results, removing speech segments entirely can negatively impact the Audio Classifier's performance. It is to be emphasised that the objective of collecting data is to use audio as a modality for sensing, and removing every segment that contains speech limits our ability to analyse the audio data. This can result in undesirable results.

Therefore, to ensure that the system works efficiently, several considerations need to be considered. Firstly, the system should be designed in such a way that it can work effectively within single-occupant households. Secondly, due to the limited availability of labelled data, the system should incorporate transfer learning techniques to improve its accuracy and efficiency. Thirdly, the system must be able to effectively remove speech from audio data recorded in a smart home environment without hampering the performance of the audio classifier. Fourthly, the system should be designed with safety in mind, ensuring that it does not collect or store any speech audio data. Lastly, the system should be scalable, allowing it to adapt easily to new environments.

To summarise, our next study should have these requirements:

- The system should be designed to work efficiently with a limited number of people in a house.
- The system should incorporate transfer learning techniques to improve its accuracy and efficiency because of limited labelled data.
- The system should be able to effectively remove speech from audio data recorded in a smart home environment.
- The system should be designed with safety in mind, ensuring that it does not collect or store any speech audio data.
- The system should be scalable, allowing it to be easily adapted to new environments.

Chapter 6

Privacy Preservation with Speech Removal

6.1 Introduction

This research mainly focuses on passive sensing using audio. This is because audio is a versatile modality that detects various activities in a smart home scenario. Instead of installing multiple sensors, a microphone can be used with machine learning or deep learning algorithms to create a system that can detect activities or changes of activities [10]. This method can be useful in a home environment, where voice-assistant devices are installed and microphones are already in place for audio recording. This might also bring privacy concerns, especially regarding private conversations. In this chapter, we developed a system that is a step forward to privacy preservation from chapter 5 with a more sophisticated and data driven way to eliminate any private concerns regarding speech.

6.1.1 Case Study

The case study 8 highlighted a company using voice-activated devices and multi-sensory data in smart homes for the elderly. The researchers gained insights into the challenges and opportunities from the AI-based Dementia Assistive & Passive Technology for Non-Invasive Elderly Care (ADAPTIVE) project, which uses audio sensing to monitor changes in occupants' gait during real-life deployments.

6.1.2 Challenges

In their case study 8, the researchers found that collecting and labelling data is not a feasible or sustainable option due to the nature of the environment. However, they noted that it could be feasible to implement manual data labelling over a short period, for example, 2 to 4 weeks, for any new deployment.

Another challenge they encountered during the case study was related to data collection in elderly homes. Specifically, when recording private conversations between carers or family members and the elderly, visitors felt uncomfortable being recorded. In each home where the system was installed, a single-occupant household, the participants' input indicated a small number of regular visitors, well below 20 per household, including family and carers.

As mentioned, these are the challenges they have:

- Maintain Acoustic Activity Recognition
- Limited Number of Labelled Data
- Limited Number of People Visiting

6.1.3 Requirements

The case study, a project focused on a different task, allowed the researchers to gain a better understanding of the challenges associated with deploying a smart home environment and its limitations. Similar to previous studies [192], they were able to train an audio classifier using a deep learning model. However, they recognised that for real-world deployment, labelled data were limited, necessitating reliance

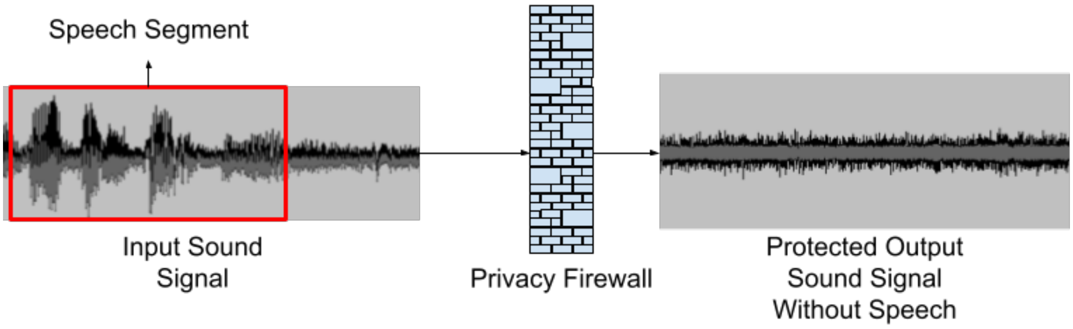


Figure 6.1: Required Plan

on a small amount of data collected from the environment and manually labelled. A common approach in acoustic sensing, when faced with limited audio data, is to utilise transfer learning [78, 193]. The authors employed a larger pre-trained model to extract audio features and subsequently developed a shallow model using these features, which yielded promising results.

It is important to remember that the number of regular visitors, which includes family, friends, and carers, is small, typically fewer than 20 people. This is a crucial factor to consider when assessing the privacy of the occupants and visitors. Based on previous research on face-to-face social networks [169, 170], the authors indicate that the number of the most common individuals encountered in daily environments ranges from 4 to 20.

In order to preserve privacy while using audio as the main modality, it is necessary to remove speech from the audio signal while maintaining the background sound. As shown in figure 6.1, a privacy firewall is required to prevent any speech audio data from being used in any audio analysis from the home environment. Removing speech data while the background audio stays intact can support the development of assistive living systems without any privacy concerns, while activity data can be collected and analysed.

These are the requirements were concluded from the previous chapter:

- The system should be designed to work with a limited number of people in a house.

- The system should incorporate transfer learning techniques to improve its accuracy and efficiency due to the limited labelled data available.
- The system should be able to effectively remove speech from audio data recorded in a smart home environment while keeping the background intact.
- The system should be designed with safety in mind, ensuring that it does not collect or store speech audio data.
- The system should be scalable, allowing it to adapt easily to new environments.

6.2 Methodology

6.2.1 Overview

The objective of our methodology is to create a pipeline that removes speech while maintaining the background audio in order to be used for acoustic activity classification. The overall methodology has to be broken down into several compartments, and here the bigger picture will be described alongside the steps needed to develop and achieve this methodology. The methodology is broken down into three steps. The first step is capturing audio data, which will happen in an audio-sensing environment. The second step is applying our proposed methodology which works as a Privacy Firewall to protect the privacy of the occupants regarding speech. The third step is a performance analysis of the data. In an audio sensing environment, where audio is collected, there might be multiple audio sensing analyses, and our proposed methodology can be integrated into those systems to preserve privacy for the occupiers. Some examples for further analysis can be footstep analysis or sleep analysis through audio.

In this study, an Acoustic Activity Classification (AAC) and a Voice Activity Detection (VAD) method are being applied for the "performance analysis" step. During this analysis, two aspects are focused on. The first aspect concerns how well privacy is preserved by the pipeline (e.g., the removal of speech), and the second aspect relates to the usability of the output for AAC. AAC is used to understand the activities occurring in the audio scene, while VAD is used to determine the extent of speech present in the recorded audio. The choice to utilise AAC and VAD results as metrics is made in order to demonstrate the improvement of the proposed methodology.

To complete the proposed methodology, these three tasks need to be achieved:

- Acoustic Activity Classification in real deployment and with the public dataset.
- Assessing the impact of speech on Acoustic Activity Classification from the real deployment and public dataset.
- Evaluating the feasibility of removing speech while retaining background activity sounds.

Our proposed system will be integrated into passive acoustic environments and prioritise privacy by removing audio-recorded speech data before undergoing further analysis. During the evaluation, an audio classification analysis will be applied, but it is anticipated that any type of audio analysis will be viable once the audio data has passed through the privacy firewall.

The methodology consists of three tasks, each designed to achieve a specific goal. In the first step, a thorough understanding of how an audio classifier can work with real deployment data and its accuracy is aimed for. To accomplish this, data from a real deployment will be used and fed to the audio classifier. The results will then be analysed to determine the accuracy of the classifier. This task will provide the baseline performance of the classifier and help in identifying any areas where improvements can be made in the next step.

In the second task, speech will be introduced to the data from the real deployment. This will allow for the analysis of the effect that speech has on the performance of the baseline audio classifier. The results of this step will be compared with those of the first step, and any differences or discrepancies arising due to the introduction of speech will be identified.

The third and final step of the methodology involves creating, training, and evaluating a deep learning model that can effectively remove any speech data from the sound signal for privacy preservation while improving the accuracy of the audio classifier. The data to be used for this deep learning model will be synthesised. For the previous two steps, data from a real deployment alongside data from a public dataset was used. The same datasets will be utilised for this step, with the addition of speech data.

The speech audio data is aimed to be introduced to the background audio at various sound levels in order to demonstrate the different speech sound levels that naturally occur in a real acoustic environment. Where the person who talks is not directly speaking in front of the microphone, but the occupier speaks in multiple locations in the room, for example, while sitting on the sofa or walking around the house. For this purpose, four sound levels of speech audio will be synthesised with background audio from public datasets:

- 40% Speech Sound Level
- 60% Speech Sound Level
- 80% Speech Sound Level
- 100% Speech Sound Level

In the Speech Removal section, an explanation will be provided about how data are synthesised. The insights gained from the first two steps will be used to develop a deep learning model that can remove only speech sounds while keeping the background sounds intact. This model will be extensively trained and evaluated to ensure that it meets our requirements and effectively removes unwanted speech data from the sound signal.

Acoustic Activity Classification (AAC) and Voice Activity Detection (VAD) will allow us to identify performance improvement or decline in each step. These are valuable tools in the application domain of an acoustic environment where data are not always clean or labelled.

6.2.2 Stage 1: Acoustic Activity Recognition

In this first stage of our methodology, data from a real deployment and a public dataset will be experimented with, as mentioned in Chapter 3. The data from the real deployment are labelled and categorised in SINS dataset [103]. The SINS dataset was collected by researchers at the University in Leuven, Belgium. This dataset was a home environment replicating a smart care home equipped with cameras and microphones alongside other sensors, further explained in the Datasets section. The SINS dataset is the closest to a real environment deployment. The other dataset, ESC-50 was used for comparison, a public dataset with 20 classes of environmental data. It is popular in the acoustic domain and is commonly used as a benchmark dataset.

6.2.2.1 Baseline

At this stage, the aim is to establish a performance baseline for Acoustic Activity Recognition with the suggested datasets. As pointed out in the requirements, collecting labelled audio data in our target environment is challenging, resulting in limited annotated data. Deploying cameras with this system is considered a privacy concern, and having an individual observe the activities of the smart home occupier while at home is not feasible in this scenario. Therefore, transfer learning will be employed for our acoustic activity recognition Machine Learning (ML) model as a typical approach employed when there are limited datasets from a target environment [133].

Transfer Learning in ML consists of two parts. The first stage is utilising a pre-trained model that is trained on a large dataset. It is possible to extract embedding features from our smaller dataset using the pre-trained model. For this example, the VGGish [118] model was selected, which was trained with Audio Set [123] and

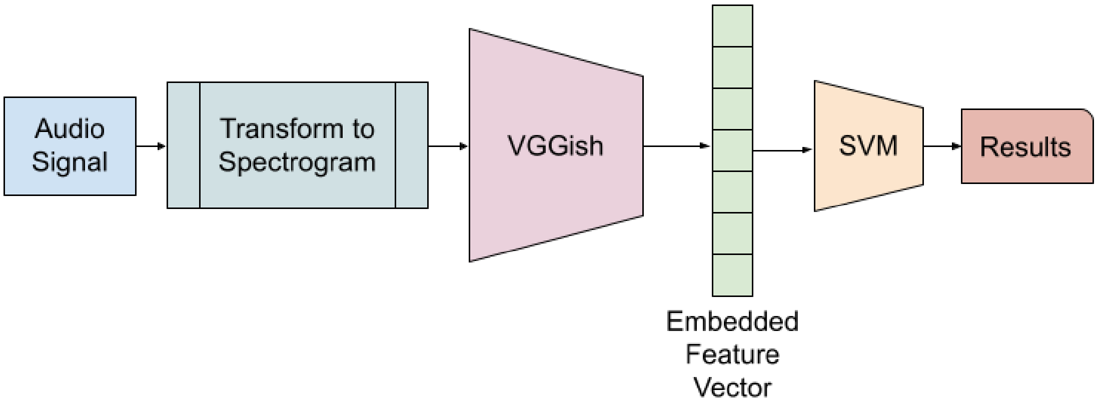


FIGURE 6.2: Acoustic Activity Classification with VGGish

is a popular model to use in the acoustic classification domain [194]. As mentioned in the Dataset section, Audio Set is an enormous collection of 10-second audio clips from videos updated on YouTube, and metadata is used to label those audio clips.

The second stage of transfer learning is to build our ML classifier using the embedding features extracted from the VGGish pre-trained model using the data from the suggested dataset. On the first occasion, the SINS dataset was used; on the second, the ESC-50 dataset was used. This methodology, as shown in figure 6.2, can be very useful as it is possible to adapt the depth of a deep learning model VGGish and specialise it to a specific domain or environment. For example, in a home environment, the labelled data is limited and hard to capture.

To get into the technicalities of the proposed pipeline, first it is needed to establish the pre-trained model that is used. VGGish [118] is the pre-trained model commonly used to extract embedding features from audio. VGGish is a Convolutional Neural Network (CNN), that is based on the VGG architecture. VGG was originally proposed by the Visual Geometry Group at the University of Oxford [195] for image recognition tasks. However, VGGish is specifically adapted for audio signal processing tasks, particularly feature extraction and classification.

While VGG was designed to operate on 2D data (images), VGGish converts the 1D audio waveform into a 2D image (audio spectrogram) where one axis represents time and the other represents frequency bands. VGGish model consists of

4 convolutional layers followed by 2 fully connected layers. After each convolutional layer, there is a max pooling layer, which reduces the spatial dimensions of the feature maps while retaining the most important information. VGGish is a prevalent model in the audio classification domain used in various applications.

The proposed pipeline utilises the VGGish pre-trained model and follows these steps:

- Transformation of the audio signal to Spectrograms
- Use of VGGish to extract embedding features
- Aggregation of embedding features
- Data preparation for ML classification
- ML Classification

First, some pre-processing is taking place on our audio files based on the VGGish model requirements. The VGGish model takes a spectrogram patch of 96 frames x 64 mel channels as input. This patch corresponds to 96 consecutive columns from a spectrogram that is calculated with 10ms frames (based on 25ms windows), which covers nearly one second of input per window (0.975s if you include the whole of the 25 ms window at the end). The VGGish model was trained on these 96 x 64 input spectrograms, so the data are prepared in a similar manner to our suggested dataset.

The appropriate spectrograms of size 96 x 64 were utilised to obtain embedding features from the suggested datasets, with a sliding window of 25 ms overlap and a 10 ms step using the Librosa library [196]. These spectrograms were then used as input for the VGGish model, which returned a feature embedding vector of 128 dimensions per window of 0.975 seconds. These extracted embedding features were used to train a Support Vector Machines (SVM) model. As mentioned in chapter 5, SVMs were commonly used in addition to VGGish models, as was noted by other researchers in the audio sensing domain [193].

Our dataset contains a diverse range of audio file lengths. To process these data, a sliding window of 0.96 seconds was employed, which equates to 15360 frames for the 16kHz sampling rate audio file. This sliding window has a step length of 0.48 seconds or 768 frames, which is a 50% step.

After the spectrograms were produced, the spectrograms from each dataset were fed separately, with the data from the real deployment (SINS) and the data from the public dataset (ESC-50) being used. A 128-D embedding feature vector was produced for each 0.96-second audio clip from the data. Due to the nature of the data and the goals to be achieved, it was determined that a 0.96-second clip is too short to capture certain activities, such as preparing breakfast. After multiple experiments, it was decided that every 10 embedding feature vectors coming from the same audio class would be aggregated. The aggregation was calculated using a non-overlapping sliding window to find the mean of 10 embedding features from each audio class. This method was employed to capture embedding features from approximately 5 consecutive seconds of audio.

The data, embedding features and labels were combined to train the SVM using the information extracted from the VGGish model. The data were divided into Train and Test datasets, with a 70/30 split using Stratified Shuffle Split from SciKit Learn [191] to ensure that the data size per class was equal.

To normalise the data, L2 normalisation was applied, which resulted in the data ranging from 0 to 1. Then trained the SVM model using SciKit Learn [191] and fine-tuned it to obtain the best possible results. For hyper-parameter optimisation, a parameter range was defined, including C, $gamma$, $kernel$, and $degree$, to achieve this. Grid Search was used to test the parameters, which refit the model with the best parameters. Finally, the classifier was set with the optimal parameters obtained from the grid search, where $C = 10$, $degree = 5$, $gamma = 1$ and using a polynomial kernel.

In conclusion, transfer learning techniques like VGGish and SVM offer a promising methodology for audio classification in domains where labelled data is scarce. The accuracy of the models depends on various factors, such as the quality of the data, feature extraction techniques, and the model used for classification. With modern tools like TensorFlow [190] and SciKit Learn [191], researchers and developers can achieve high accuracy in audio classification tasks, which can be used for tasks such as speech recognition, music genre classification, and environmental sound classification.

Moving forward, the abilities of an acoustic activity classification model using the proposed datasets have been understood. The development of this methodology was guided by the scarcity of labelled data in a home environment. Additionally, the privacy of the occupant is acknowledged, necessitating an understanding of how much of the recorded audio contains speech. As shown in figure 6.3, the Silero Voice Activity Detection model [186] will be used to detect any speech present in the data.

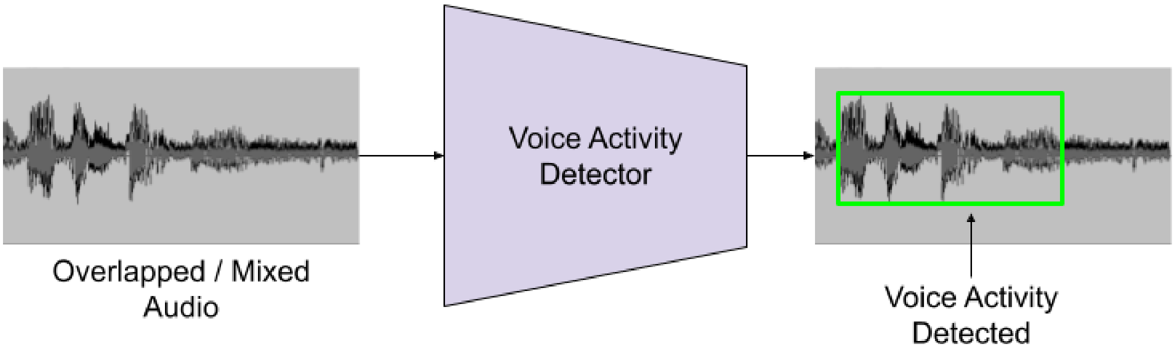


Figure 6.3: Voice Activity Detection

Silero VAD was trained on a large collection of audio recordings that contained both speech and non-speech segments from Open SLR [188]. Silero is a neural network with multi-head attention (MHA). MHA-based networks are popular in various applications, such as natural language processing and computer vision. The authors believed that combining MHA and Short-time Fourier Transform as features would result in a highly effective neural network, and their results proved to be promising.

Silero [186] is a PyTorch-based model that can handle audio inputs with sample rates of 8 kHz and 16 kHz, developed by the S Team. However, the model can also process audio inputs with sample rates of 32 kHz and 48 kHz by resampling them to 16 kHz through slicing. Additionally, Silero supports three different chunk sizes: 30, 60, and 100 ms. The model was trained in more than 100 languages and has shown good generalisation. Each chunk of audio takes approximately 1 ms to process on a single CPU thread. The model outputs a probability between 0 and 1 to indicate if speech is present in the audio segment. The creators of Silero set the speech detection threshold at 0.5. The threshold of 0.5 was selected by the creators of Silero because the researchers did not have any reason to be biased in detecting speech or not, and a cut-off in the middle was a natural selection.

Silero can be used to input audio data that may contain speech, and based on the model's output, it is possible to perform post-processing using Python to determine the timestamps of each section that includes speech. The process will identify the total amount of speech detected, as shown in Figure 6.3. Silero uses a 30 ms non-overlapping sliding window to predict if there is speech in the audio. However, a window of 30 ms was found to be too short to contain any meaningful speech content. To ensure the accuracy of the pipeline, speech was recognised in the audio, where speech was detected by Silero for at least 1 second. Afterwards, the total of detected speech was calculated if the speech detected exceeded the 1-second mark, and the results were presented in a table as a percentage of speech relative to the length of the audio files from the two datasets.

6.2.2.2 Results

Progress has been made in the implementation of the first stage of our methodology. As a result, audio activities from the two suggested datasets were classified, achieving baseline results. The SVM model was trained on the embedding features extracted from the VGGish model, resulting in an accuracy of 85% for the SINS dataset and 82% for the ESC-50 dataset, respectively. Additionally, findings from the Voice Activity Detection model have been included in the same table. The total audio duration was evaluated alongside the total audio where speech was detected for both datasets. It was indicated by our analysis that only 0.2% of speech was detected in the SINS dataset, while 0.6% of speech was detected in the ESC-50 dataset.

The acoustic activity classifier was trained on the baseline dataset, with the understanding that there is a limited number of audio data. For the SINS dataset, 84% precision and 90% recall were achieved, and for the ESC-50 dataset, 84% precision and 83% recall were achieved. Our approach does not aim to achieve the best results, as there are other approaches, for example, Audio Spectrogram Transformer (AST) that achieve state-of-the-art in many acoustic activity classification tasks [192]. The proposed approach for acoustic activity classification is to be utilised as the evaluation tool to assess the performance of the privacy wall. The VAD results are 0.2% of the SINS Baseline contains speech data and 0.6% of the ESC-50 Baseline contains speech. These results show that the datasets contain a negligible amount of speech data.

The results have been set as a baseline for the methodology, and the study is intended to continue to the next stage to understand the impact of speech on the acoustic environment. Furthermore, the methodology for speech removal and the performance analysis of a privacy-preserved acoustic environment will be presented.

6.2.3 Stage 2: The Impact of Speech on Acoustic Activity Recognition

In the second stage of our methodology, speech data was incorporated into our datasets. This involves utilising the Audio Acoustic Classification and Voice Activity Detector tools from the first stage to assess the impact of audio classification in a speech-rich environment. The objective is to remove speech in an acoustic environment; however, first, an understanding must be gained regarding how adding speech affects the data and audio classification model. These discoveries may pose unanticipated challenges, which will be addressed through adjustments to the methodology. In this stage, synthetic datasets will be created based on public datasets and real-world deployment data, followed by applying the previously developed Acoustic Activity Classification and Voice Activity Detection tools to examine the effects and implications of added speech.

For this stage of our methodology, this is the procedure:

- Dataset Synthesis (Described in 3)
- Apply Acoustic Activity Classification
- Apply Voice Activity Detection
- Compare Results with previous findings

6.2.3.1 Performance Metrics

At stage 1 of our methodology, the goal was to set the baseline for our acoustic activity classification and voice activity recognition based on our proposed dataset, SINS and ESC-50. At this stage, after synthesising variants of a synthetic dataset that includes speech in various sound levels, the acoustic activity classification and voice activity recognition models will be deployed to understand the impact of speech in the acoustic environments. The acoustic activity classification model considers the real environment in which our system will be deployed, where labelled data are limited and hard to collect.

The eight synthesised datasets will be run using the same model as the baseline. For Voice Activity Detection, Silero [186] will be used, and for the acoustic activity classification, a 50% overlap window will be applied to the raw audio data to extract spectrograms of shape 96 x 64. Embedding features of 128-D will then be extracted from those spectrograms using the VGGish model [118]. Subsequently, the 10 spectrograms will be aggregated by calculating their mean, and the extracted data will be prepared for evaluation using the SVM model that was trained earlier from the baseline.

6.2.3.2 Results

During this stage of the methodology, synthetic datasets were generated, and speech data was incorporated into the ESC-50 and SINS datasets. The Acoustic Activity Classification and Voice Activity Detection tools were then applied to the datasets, resulting in some findings. Comparing these results with previous findings will help understand how the addition of speech affects the acoustic activity classification model. In this section, the results of the experiments conducted to date will be presented. Afterwards, the next stage of the methodology will be addressed.

As shown in table 6.1, results were produced using the AAD and VAD model. Beginning with the baseline from the SINS dataset, which was presented at stage 1 of this methodology, achieved 84 % precision and 90% recall from AAD and 0.2 % speech detected from the VAD model. By adding Speech to our data, the accuracy of the AAD deteriorated. Specifically, in the SINS dataset with the added Speech at 100% sound level, the AAD dropped to 50% precision and 51% recall whilst VAD raised to 49% of detected audio that contained speech. By reducing the sound level of the speech data before adding them to the background audio data, minor improvements can be found to the acoustic activity classification and a drop in the percentage of speech detected in the synthetic dataset.

A similar effect was the ESC-50 dataset, with a baseline AAD at 84 % precision and 83% recall with voice activity detected at 0.6 %. After the insertion of speech data at 100 % sound level, the AAD dropped to 70% precision and 69% recall, and voice activity was detected at 67.5 % of the audio data. During the alternations of speech sound level, the lower the volume level of speech data, the better results were obtained in acoustic activity classification, where the voice activity detection percentage was degrading.

Overall, these results provide valuable insights into the effects of speech on acoustic activity classification and the importance of the sound level from the speech audio data. Having achieved those results, stage three of our methodology will

Table 6.1: Precision, Recall and VAD performance from our baseline and overlapped speech on various sound levels.

Dataset	**Precision**	**Recall**	**VAD**
SINS (Baseline without Speech)	84%	90%	0.2%
SINS with added Speech (40% sound level)	59%	58%	0.60%
SINS with added Speech (60% sound level)	52%	52%	1%
SINS with added Speech (80% sound level)	48%	56%	9.89%
SINS with added Speech (100% sound level)	50%	51%	49%
ESC-50 (Baseline without Speech)	84%	83%	0.6%
ESC-50 with added Speech (40% sound level)	81%	75%	6.80%
ESC-50 with added Speech (60% sound level)	74%	74%	7.2%
ESC-50 with added Speech (80% sound level)	71%	73%	17.19%
ESC-50 with added Speech (100% sound level)	70%	69%	67.5%

be entered, where a speech removal model will be applied to the data. The aim of the final part is to ensure that any speech is removed from the synthetic data while keeping the background data intact, thus allowing for the improvement of the acoustic activity classification while the privacy of the occupants regarding conversational data is preserved.

6.2.4 Stage 3: Speech Removal

The final stage of our methodology consists of removing speech from the synthetic datasets for the purpose of preserving the privacy of people living in the acoustic environment regarding speech. The aim is also to improve the acoustic activity classification results by removing speech from the data. As shown in figure 6.4, the pipeline involves the removal of speech using a speech removal model. Afterwards, the proposed Acoustic Activity Classification and Voice Activity Detection models will be used to analyse the changes in the data concerning the amount of speech remaining in the audio data and to compare the Acoustic Activity Classification performance with previous results.

For the Acoustic Activity Classification, spectrograms will be extracted from the "Speech-Free" Audio Data as input to the VGGish model, from which feature embedding vectors will be obtained and evaluated using the developed SVM model.

For Voice Activity Detection, the Silero VAD public model will be applied to the "Speech-Free" raw audio data.

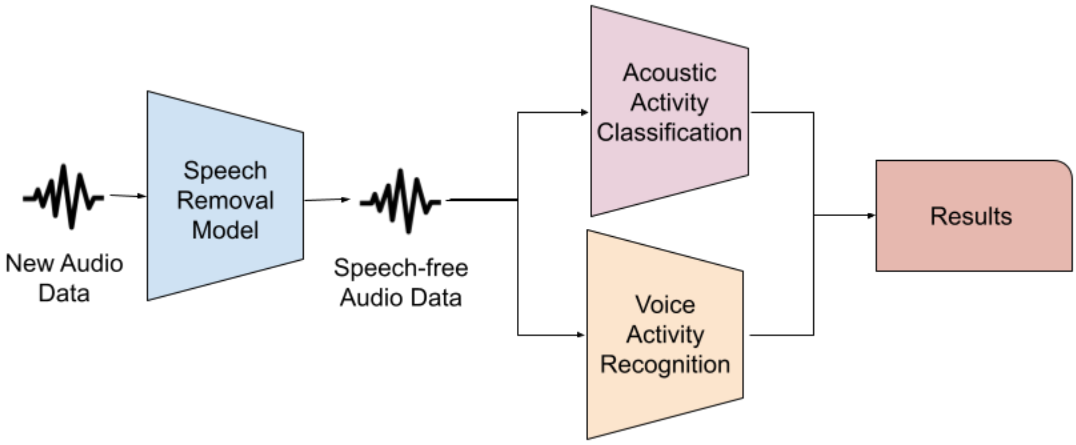


FIGURE 6.4: Deploying Speech Removal Model alongside VGGish Model

6.2.4.1 Experimenting with Public Models

Before moving on to the development of a custom model aimed at removing speech from background audio data, a variety of public models were explored. In the area of Speech Separation [197, 198], solutions for the Cocktail Party Effect [148] and Audio Source Separation [199] were published by some authors who developed their own models. Three models were chosen for testing using the public dataset ESC-50 and the synthetic version of ESC-50, where speech was synthesised at a 100% sound level.

A major benefit is gained from downloading and using other researchers' models, as the outcomes of their research can be obtained and opportunities for improvement can be explored. The conclusion was reached by experimenting with these public models:

- Facebook Denoiser [157]
- SepFormer [150]
- ConvTasNet [198]

These models were downloaded from a dataset library called Hugging Face [200]. Hugging Face is hosting a plethora of public models for researchers to download and use. These models were developed for various purposes. I will briefly explain how these models were created by their authors.

ConvTasNet and SepFormer were trained using the WSJ0-2Mix dataset [151], where their focus was to separate audio sources between 2 or more speakers. ConvTasNet consists of an Encoder, a deep convolutional neural network that does the speech separation followed by a Decoder. Where SepFormer was created after the publication of Attention is All You Need, [149], and they are utilising an attention layer.

The Facebook Denoiser model was focused on cleaning up the noise from the background sound, which is the reverse of what is being aimed to improve the quality of speech. The Valentine and DNS datasets were used to train the Facebook Denoiser [161, 201]. A U-Net structure was utilised in the model [159], which was initially developed for biomedical image segmentation. These experiments were conducted to understand the capabilities of these public models, and thus, only the ESC-50 dataset with added speech at 100% was tried. To achieve the results shown in table 6.2, the same methodology was implemented for acoustic activity classification and voice activity detection, which were explored in stages 1 and 2, after the public models were deployed on ESC-50 data with added speech. Baseline results from stage 1 and ESC-50 with added speech at 100% sound level are also presented as a comparison to the new findings.

Table 6.2: Results from the initial experiments on AAD and VAD before and after removing speech data using various public models for speech removal or speech separation.

Dataset	**Precision**	**Recall**	**VAD**
ESC-50 (Baseline without Speech)	84%	83%	0.6%
ESC-50 with added Speech	70%	69%	67.5%
ESC-50 Removed Speech - Denoiser	67%	66%	6.55%
ESC-50 Removed Speech - SepFormer	51%	51%	36.34%
ESC-50 Removed Speech - ConvTasNet	52%	60%	47.21%

After applying the speech removal models, the results from the AAD have deteriorated. However, the VAD model has shown improvement in terms of detecting the amount of speech. For instance, the Denoiser, by Facebook, has performed the best in removing speech, with only 6.55 % of audio-containing speech, as compared to the original dataset, which had 67.5% of speech. Despite this, the AAD results have worsened by 67% precision and 66% recall, as compared to the ESC-50 with added speech, which had 70% precision and 60% recall, and the baseline, which had 82% precision and 83% recall.

6.2.4.2 Creating a purposed Speech Removal Model

The Speech Removal Model is being developed with the goal of removing all speech data from the environment and improving acoustic activity classification results. Since the public models that were experimented with are not suitable for this task, a model based on specific requirements is being created and trained. The model being developed is also compatible with the VGGish model for performance analysis. In this research, the work that included Acoustic Activity Classification after the speech was removed, alongside Voice Activity Detection, will be presented. Any future work can incorporate the proposed speech removal model with other audio analysis tasks.

Based on the recent results presented in table 6.2, it has been determined that the Facebook Denoiser performed the best, leading to the decision to use the structure of their model. The model was adjusted to fit the spectrogram size of VGGish, 96 x 64. A U-Net structure, specifically a Deep Convolutional Autoencoder with symmetric skip connections [159], was utilised. The U-Net Model was adapted for spectrograms to facilitate speech removal.

For the training data, the synthetic dataset was used, created from the real deployment data. The SINS dataset synthesised LibriVox Speech data. The data were split for training and evaluation at 70/30 for both the SINS dataset and the SINS dataset with speech data. The pre-processing of the dataset was done by extracting spectrograms from our audio datasets using Librosa [196] at four directories:

- Train with Speech
- Eval with Speech
- Train without Speech
- Eval without Speech

The training data are divided into two categories: Train without Speech and Train with Speech. Train without Speech comes from the SINS dataset and Train with Speech originates from SINS with added Speech at 100 % sound level. The other variations of our SINS with added speech synthetic datasets were used for evaluation purposes. The spectrograms were generated from raw audio files using a sliding window of 0.96 seconds, which corresponds to 15360 frames for the 16kHz sampling rate audio file. This sliding window has a step length of 0.48 seconds or 768 frames, which equals a 50% step.

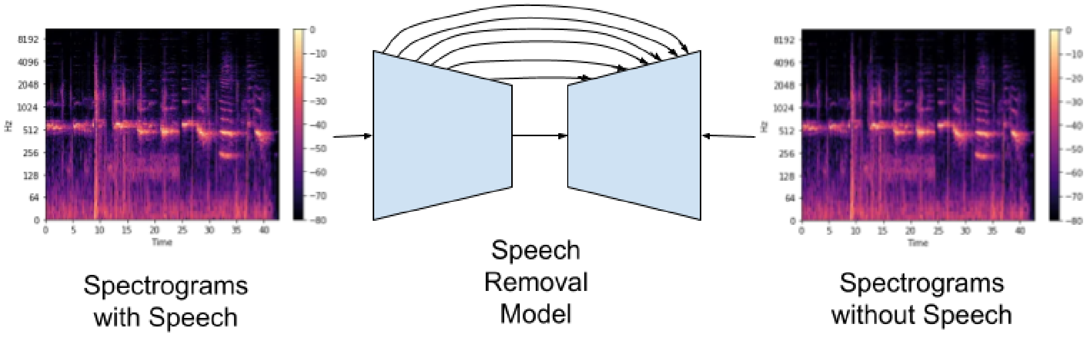


FIGURE 6.5: Training Speech Removal Model

The proposed model uses Train with Speech as input and Train without Speech as the output target. The network accepts spectrograms that use a global scaling factor, which maps them to a distribution of values between -1 and 1.

The model has an encoder and a decoder. The encoder consists of 5 convolutional layers, each with LeakyReLU, maxpooling, and dropout. The decoder is symmetric to the encoder and features skip connections. The final activation layer uses a hyperbolic tangent function (Tanh), resulting in an output distribution between -1 and 1.

The encoder:

1. **Input Layer:** This layer takes an input image (for our purpose a spectrogram) of size (96, 64, 1) where 96 x 64 represents the height and width, and 1 represents the number of channels (grayscale image was used for spectrogram)
2. **Convolution Block 1 (Conv1):**
   (a) Two consecutive 3x3 convolutions with 16 filters.
   (b) LeakyReLU Activation.
3. **Max Pooling 1:** Downsamples by factor of 2 (2x2).
4. **Convolution Block 2 (Conv2):** Same as Conv1, but with 32 filters
5. **Max Pooling 2:** Downsamples by factor of 2 (2x2).
6. **Convolution Block 3 (Conv3):** Similar to Conv2, but with 64 filters.
7. **Max Pooling 3:** Downsamples by factor of 2 (2x2).
8. **Convolution Block 4 (Conv4):**
   (a) Similar to Conv2, but with 128 filters.
   (b) Includes Dropout (0.5) for regularisation.
9. **Max Pooling 4:** Downsamples by factor of 2 (2x2).
10. **Convolution Block 5 (Conv5):**
    (a) Similar to Conv2, but with 256 filters.
    (b) Includes Dropout (0.5) for regularisation.

The decoder:

1. **Upconvolution 6 (Up6):**
   (a) Upsamples by a factor of 2 (2x2).
   (b) 3x3 convolution with 128 filters and LeakyReLU activation.
2. **Concatenation 6:** Combines Up6 features with Conv4 features.
3. **Convolution Block 6 (Conv6):** Similar to Conv2, but with 128 filters.
4. **Upconvolution 7 (Up7):** Similar to Up6, uses Conv6 features and concatenates with Conv3.
5. **Convolution Block 7 (Conv7):** Similar to Conv2, but with 64 filters.
6. **Upconvolution 8 (Up8):** Similar to Up6, uses Conv7 features and concatenates with Conv2.
7. **Convolution Block 8 (Conv8)**: Similar to Conv2, but with 32 filters.
8. **Output Layer:** 1x1 convolution for segmentation mask (96 x 64, 1 channel).

The input and output of our speech removal model are represented using spectrograms, enabling their combination with further audio analysis. In this example, embedding features are extracted using the VGGish model after the speech has been removed from the original audio data, and the analysis continues with the produced spectrograms.

6.2.4.3 Hyper Parameter Optimisation

In addition to creating the speech removal model and using it in combination with audio analysis, also a hyper-parameter optimisation was performed to fine-tune the model's performance.

Here are the parameters that were explored (The following parameters were tuned using Keras Tuner library [202]):

- **size_filter_in**:
  - The number of filters in the first convolutional layer.
  - This controls the feature map depth throughout the network.
  - Options: 16, 32, 48, 64, 80, 96, 112 and 128
- **dropout_rate**:
  - The dropout rate is applied after specific convolutional layers.
  - To prevent overfitting.
  - Options: 0.0, 0.1, 0.2, 0.3, 0.4, 0.5.
- **activation**:
  - The activation function used in convolutional layers
  - Introduces non-linearity for improved learning.
  - Options: ReLU (rectified linear unit), Leaky ReLU, ELU (Exponential Linear Unit), Sigmoid.
- **batch_size**:
  - The number of samples used for gradient updates during training.
  - Affects training speed and memory usage.
  - Options: 32, 64 and 128
- **learning_rate**:

  - The learning rate used by the optimiser.
  - Controls the step size taken during gradient descent.
  - Range: 1e-6 to 1e-2 in logarithmic scale

Here is the configuration that achieved the best validation loss after 60 epochs:

- **size_filter_in**: 112 (increased from the default range)
- **optimiser**: SGD (instead of Adam)
- **dropout_rate**: 0 (no dropout applied)
- **activation**: leaky_relu (confirmed a good choice)
- **batch_size**: 32 (default value)
- **learning_rate**: 2.9456e-05 (significantly lower than the default range)
- **Epochs**: 20 (early stopping based on validation performance)

Based on the results of Hyperparameter Optimisation, the following conclusions were drawn:

1. The model's ability to learn complex features may have been enhanced by increasing the initial number of filters (size_filter_in). For this particular dataset and network architecture, the SGD optimiser may have been more appropriate than Adam.
2. A dropout rate of 0 (i.e. disabling dropout) may indicate that sufficient regularisation was achieved through other factors, such as the activation function.
3. Leaky ReLU remained the optimal activation function for this task.
4. The default batch size (32) proved to be efficient.
5. A significantly lower learning rate (2.9456e-05) was crucial for achieving convergence and avoiding overfitting.

6. The model achieved good performance relatively quickly, as evidenced by early stopping at 20 epochs.

6.2.4.4 Results

Table 6.3 displays the Baseline results for stages 1 and 2 of the SINS and ESC-50 datasets. As noted in stage 2, adding speech in the ESC-50 dataset tends to lower the accuracy of acoustic activity classification. At 100% speech sound level, the accuracy drops from 82% (baseline without speech) to 71%. However, as speech sound level decreases, the impact on accuracy becomes less prominent.

Similarly, adding speech to the SINS dataset also diminishes classification accuracy. The accuracy reduction is most noticeable at 100% speech sound level (49% versus 85% baseline) and gradually improves as the speech sound level decreases. The expectation of an acoustic environment is that the speech sound level should not be very high compared to the background sounds because of the nature of the environment and where the microphones are set up. The person living at home does not always face the smart device that contains the microphone, and they might not be located in very close proximity to the microphone. As a solution, a multi volume level approach was utilised to understand the effectiveness of our system in filtering out conversational content from the audio signal while maintaining high accuracy in acoustic activity classification.

In stage 3, our methodologies produced the following results. Removing speech from the ESC-50 dataset with added speech enhances classification accuracy compared to scenarios where speech was present. For example, eliminating speech entirely (0% voice activity detected) yields an accuracy of 80%, which is very close to the baseline (82% accuracy). As for the SINS dataset, similar to ESC-50, removing speech from the SINS dataset with added speech increases classification accuracy compared to having speech present. In this case, the improvement is even more significant, where speech removal at all sound levels leads to better results than having speech in the audio signal.

Table 6.3: Performance metrics for AAD (Precision and Recall) and VAD on various datasets.

Dataset	**Precision**	**Recall**	**VAD**
ESC-50 (Baseline without Speech)	**84%**	**83%**	**0.6%**
ESC-50 with added Speech (40% sound level)	**81%**	**75%**	**6.80%**
ESC-50 with added Speech (60% sound level)	74%	74%	7.2%
ESC-50 with added Speech (80% sound level)	71%	73%	17.19%
ESC-50 with added Speech (100% sound level)	70%	69%	67.5%
ESC-50 after Speech Removed (40% sound level)	**85%**	**85%**	**0%**
ESC-50 after Speech Removed (60% sound level)	80%	79%	0%
ESC-50 after Speech Removed (80% sound level)	74%	74%	0%
ESC-50 after Speech Removed (100% sound level)	73%	81%	0%
SINS (Baseline without Speech)	**84%**	**90%**	**0.2%**
SINS with added Speech (40% sound level)	**59%**	**58%**	**0.60%**
SINS with added Speech (60% sound level)	52%	52%	1%
SINS with added Speech (80% sound level)	48%	56%	9.89%
SINS with added Speech (100% sound level)	50%	51%	49%
SINS after Speech Removed (40% sound level)	**71%**	**77%**	**0%**
SINS after Speech Removed (60% sound level)	71%	70%	0%
SINS after Speech Removed (80% sound level)	65%	65%	0%
SINS after Speech Removed (100% sound level)	62%	70%	0%

One of the primary advantages of using this particular methodology is that the capability to eliminate any speech content from the audio signal completely is provided. This is particularly important as the privacy of the occupants is prioritised. The Voice Activity Detection model is utilised to understand if any speech is included in the dataset being used. According to the Voice Activity Detection column, the speech removal model developed effectively removes all speech content in both the ESC-50 and SINS datasets and all the synthetic alternative datasets created.

Additionally, this approach has been shown to improve AAD from the data with added speech after the speech was removed, making it a valuable tool for various applications. By removing speech audio content, this methodology allows for a more thorough analysis of the acoustic properties of the signal, providing a more accurate understanding of the audio and the environment in which it was recorded.

6.3 Results

Finally, an overview of all results from the experiments conducted during this research was reviewed, and it is pleasing to see that the pipeline performs well in both public datasets and real deployments, as well as in data collections that were conducted. The aim of the pipeline was to completely reduce or remove the speech content from the audio signal gathered from a home environment while improving the acoustic activity classification.

Since the ESC-50 dataset does not contain speech, synthetic datasets were created by incorporating speech audio data into ESC-50 audio data at varying sound levels. An acoustic activity classification model, along with a Voice Activity Detection public model, was utilised to evaluate the effects of speech on these datasets.

Subsequently, some public models aimed at speech separation or speech enhancement were tried, which were attempting to solve a problem similar to the one being addressed. Facebook Denoiser [157], a Speech Enhancement model, ConvTasNet [198], a deep convolutional model built for Speech Separation, and SepFormer [203], which features an architecture based on Transformers with an Attention Layer designed for speech separation, were all tested. Unfortunately, the public models did not successfully remove the speech from the background audio while keeping the background audio intact, necessitating the construction of a proprietary model.

After the speech removal model was created and applied to the synthetic data, it was observed that the acoustic activity classification outcomes were enhanced compared to those of the synthetic datasets. Although the objective was to achieve acoustic classification levels comparable to those of the baseline without speech, that was not accomplished. It is thrilling to announce that the gap between the acoustic activity classification outcomes of the baseline and the synthetic dataset with speech has been successfully narrowed. The performance is significantly better than that of the existing generic models that were tested.

The same structure as that of the SINS Dataset was followed, as the SINS dataset does not contain speech data. Acoustic Activity Classification and Voice Activity Detection were applied to establish the baseline. After speech audio data was introduced to the dataset at various sound levels, a decline in the acoustic activity classification results was noted. By applying the pipeline to these various synthetic datasets originating from the SINS Dataset, it was confirmed that speech was completely removed according to the voice activity detection model, and an improvement in the acoustic activity classification accuracy was achieved to a certain extent.

Table 6.4: AAD (Precision and Recall) and VAD performance metrics on public datasets, real deployments, and alternative synthetic datasets.

Dataset	**Precision**	**Recall**	**VAD**
ESC-50 (Baseline without Speech)	84%	83%	0.6%
ESC-50 with added Speech (40% sound level)	81%	75%	6.80%
ESC-50 with added Speech (60% sound level)	74%	74%	7.2%
ESC-50 with added Speech (80% sound level)	71%	73%	17.19%
ESC-50 with added Speech (100% sound level)	70%	69%	67.5%
ESC-50 after Speech Removed with Denoiser	67%	66%	6.55%
ESC-50 after Speech Removed with SepFormer	51%	51%	36.34%
ESC-50 after Speech Removed with ConvTasNet	52%	60%	47.21%
ESC-50 after Speech Removed (40% sound level)	85%	85%	0%
ESC-50 after Speech Removed (60% sound level)	80%	79%	0%
ESC-50 after Speech Removed (80% sound level)	74%	74%	0%
ESC-50 after Speech Removed (100% sound level)	73%	81%	0%
SINS (Baseline without Speech)	84%	90%	0.2%
SINS with added Speech (40% sound level)	59%	58%	0.60%
SINS with added Speech (60% sound level)	52%	52%	1%
SINS with added Speech (80% sound level)	48%	56%	9.89%
SINS with added Speech (100% sound level)	50%	51%	49%
SINS after Speech Removed (40% sound level)	71%	77%	0%
SINS after Speech Removed (60% sound level)	71%	70%	0%
SINS after Speech Removed (80% sound level)	65%	65%	0%
SINS after Speech Removed (100% sound level)	62%	70%	0%
First Data Collection (No Speech)	80%	86%	0%
Second Data Collection (Including Speech)	78%	77%	6.8%
Removed Speech from Second Data Collection	76%	76%	0%

Lastly, the complete pipeline was applied to the Second Data Collection, where daily activities were recorded by participants using the AudioHive App, a smartphone application developed explicitly for data collection purposes. The results of the acoustic activity classification and voice activity detection (VAD) for both data collections are presented in Table 6.4.

The First Data Collection was conducted under controlled conditions where participants were not allowed to speak. As expected, the VAD module did not detect any speech, ensuring that the dataset consisted purely of activity sounds. In contrast, the Second Data Collection permitted participants to speak during activities, resulting in speech being present in 6.8% of the total dataset.

Applying the proposed pipeline to the Second Data Collection demonstrates its ability to both classify acoustic activities and effectively remove speech content from the data. While the proportion of speech in the dataset was relatively low (6.8%), the pipeline successfully mitigated its influence on classification performance. This indicates the robustness of the approach even with minor speech interference.

The evaluation highlights that precision and recall metrics decline with increasing speech interference, as expected. However, post-processing with the proposed speech removal techniques restores classification performance to levels comparable to datasets without speech interference, particularly when speech levels are moderate (40%-60%). This aligns with the pipeline's goal of providing robust acoustic activity classification while preserving privacy by eliminating conversation content.

6.4 Conclusions

The experimental results demonstrate several key conclusions regarding acoustic activity classification and voice activity detection under conditions of various speech additions to the environmental audio signal.

- The impact of speech
- Identified benefits of speech removal
- Application on real-world data

- Privacy awareness regarding speech

The proposed methodology consisted of three stages:

1. Set the baseline of acoustic activity recognition
2. Recognise the impact of speech on an audio dataset from background sounds.
3. Application methodology of removing the speech content and improving the acoustic activity classification.

As part of the methodology, two key metrics were utilised: acoustic activity classification and voice activity recognition. To begin with, the impact of introducing speech into an acoustic environment was examined. This was crucial for understanding the role of speech in the overall acoustic environment and its effect on the metrics. Next, the overall transformations in the datasets were measured before and after the proposed dataset was applied. This enabled the assessment of the effectiveness of the proposed dataset and the determination of its impact on the overall accuracy of the metrics. By analysing these metrics, valuable insights into the performance of the methodology were gained, allowing for further refinement to achieve more accurate results.

The proposed pipeline for acoustic activity classification is believed to have the potential to be modified and customised for a wide range of acoustic sensing applications in future research. The current classification module can be substituted with other types of classification models or applications, depending on the specific needs of the acoustic home sensing system.

This pipeline can be particularly beneficial in situations where collecting labelled data is challenging or not feasible and where ensuring the privacy of individuals living in the acoustic environment is a top priority. Additionally, this pipeline can help to eliminate speech from audio data, which can be especially useful in cases where speech analysis is not required. Overall, this pipeline has significant potential for advancing the field of acoustic sensing and improving the quality of life for individuals in various settings.

Chapter 7

Conclusion

7.1 Introduction

During this research, an introduction was provided in chapter 1, presenting a basic understanding of addressing a research challenge that incorporates acoustic sensing in a home environment. Moving to the background work in chapter 2, the current work from the research community in acoustic sensing for the elderly in home environments was presented. In chapter 4, the first task of tracking the daily routines of the elderly at home through acoustic sensing was tackled; however, the collection of labelled data was deemed impossible, necessitating an unsupervised methodology. Furthermore, in chapter 5, unique challenges in acoustic environments where elderly people live were identified based on real-world deployment experiences. For instance, privacy regarding speech was recognised as essential not only for the occupant but also for the carers visiting daily to ensure the well-being and safety of the elderly. Finally, in the last contribution, chapter 6 outlined a multi-step methodology aimed at providing appropriate insights for an acoustic environment in a single-occupancy home. This methodology ensures speech-related privacy while improving the accuracy of acoustic activity recognition after the speech is removed. The overall research presents a modern solution to a contemporary environment where audio is utilised as the sole modality.

7.2 Overall Contribution of this Work

The research questions set at the beginning of this research will be reflected upon, and an understanding of how each question was answered.

7.2.1 Q1: Is it feasible to identify changes in daily routine through unsupervised acoustic sensing with performance levels sufficient for detecting significant deviations, such as change of order of activities or missing activities?

Using acoustic sensing, changes in daily routines were identified with relatively high accuracy [10]. As discussed in chapter 4, the development of a novel dimensionality reduction technique supported the implementation of the pattern matching approach for various activity sequences through acoustic sensing. This approach could determine if an activity sequence has changed.

7.2.2 Q2: What are the implications of an acoustic system in a home environment that is designed to protect against misuse of speech data?

In chapter 5, the implications of an acoustic system were discussed in a home environment designed to protect against misuse of speech data through an experimental approach. Some of the main requirements for such a system are:

- The number of people occupying the home is limited.
- The amount of labelled data that was able to be collected in a home environment is very limited.

- The speech audio data should be eliminated because of privacy concerns for residents, carers visiting and family.

- The acoustic system should easily adapt to new environments and homes.

A practical study will be conducted on the impact of these requirements.

7.2.3 Q3: What is the impact on the performance of acoustic activity recognition while providing speech-related privacy in continuous acoustic sensing?

As discussed in chapter 5, masking audio sections containing speech audio negatively affects the audio classification accuracy. Considering a different solution, in chapter 6, a novel methodology was proposed for preserving privacy regarding speech while maintaining acoustic activity classification accuracy.

7.2.4 Q4: Is it feasible to apply speech-related privacy in continuous acoustic sensing while maintaining a drop in acoustic activity recognition accuracy of no more than 10%?

In Chapter 6, the proposed methodology effectively removes speech from the audio signal while maintaining a relatively high accuracy in acoustic activity recognition. Specifically, the methodology was able to keep the drop in acoustic activity recognition accuracy to no more than 10% when tested with both the synthetic dataset and the data collected for this study. However, when applied to a dataset from a real deployment, the proposed methodology experienced a more significant drop of 13% in acoustic activity recognition accuracy. Although the system succeeded in eliminating conversational audio across all datasets, this came at the cost of reduced performance in acoustic activity recognition. The dataset from the real deployment had the most substantial impact on performance, which warrants further investigation to improve these metrics.

7.3 Future Development of This Work

The outcome of this research can be a baseline for other researchers. They can use our outcomes for their studies and develop further work. In this section, the ideas and future plans developed based on our work will be discussed.

The future developments that are possible and important after this research consist of the following:

- **Open Access Data Collections:** In the area of acoustic activity recognition, many researchers publicise their data collections as a public dataset to support the research in the area. When it comes to acoustic sensing in home environment datasets are limited and it is important to pursue a better future research in this area.
- **Speech Obfuscation:** Some scenarios of acoustic activity recognition contain speech, such as a meeting or phone call. In these cases, removing speech completely will negatively affect the recognition of these activities. A possible approach to preserve privacy while allowing detection of such activities can rely on obfuscation of speech so no conversational content will be available in our data while activity audio is clearly preserved.
- **Improve accuracy through hybrid model:**, where a pre-trained speech removal model can sit alongside a model that can be retrained with new data to target a specific environment. An ensemble framework combining pre-trained with newly trained models can have the potential to achieve better accuracy.
- **Shrink our proposed Deep Learning model to a tiny model** to fit in an embedded device. In doing that, the privacy processes will happen on the device, and not upload speech data to the cloud.

The following sections will explain more in-depth what these future methodologies consist of and what they need to be developed.

7.3.1 Open Access Data Collections

At the beginning of this work, one of the main challenges was understanding and utilising available public datasets and deciding what kind of data the researcher should be collecting. A work that is important for future research in acoustic sensing in the home environment is to develop more public datasets available to other researchers. In Chapter **??**, the public datasets and collected data were discussed for this work. Publicising acoustic sensing data from the home environment is important to expand the research in this domain.

Collecting data, especially labelled data, is a laborious endeavour, and many researchers will benefit by utilising public datasets in addition to collecting their data. DCASE challenge [143] was a major contributor in bringing researchers together to solve problems related to audio sensing by providing tasks and datasets for researchers worldwide to try and experiment with.

DCASE is a prime example of how research could be improved and developed because of the availability of public datasets in a specific domain. Producing a public dataset is not easy for a researcher, but these endeavours will support the future of research.

7.3.2 Speech Obfuscation

In Chapter 6, it was discussed that removing speech from certain environments whilst keeping the background sounds intact. This acts like a privacy firewall for the home occupant, where speech content is removed completely, and further analysis is performed on the background sounds. On some occasions, the presence of speech might be important for some activities, like having a meeting or during a phone call. The proposed system from 6, as shown in figure 7.1, is unsuitable for these scenarios because the speech content is completely removed from the audio signal.

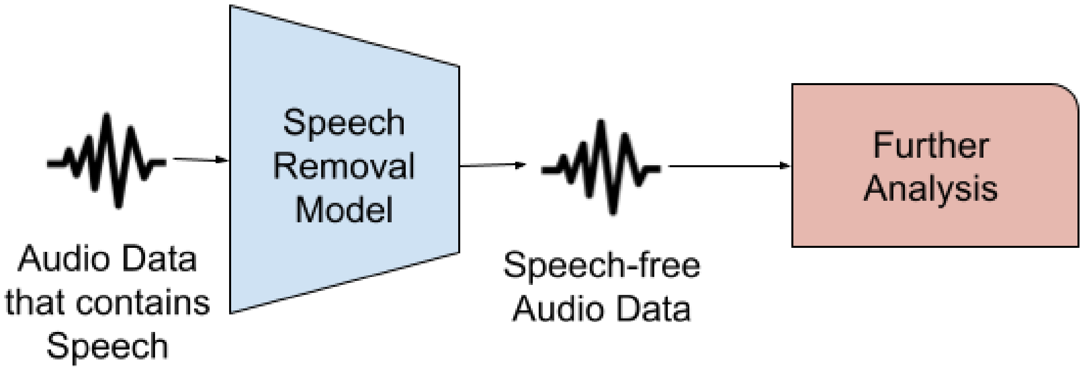


FIGURE 7.1: Proposed Speech Removal Methodology

The methodology considered suitable for these scenarios, which represents an improvement over the method outlined in chapter 6, incorporates speech separation, speech obfuscation, and finally, speech reintroduction to the audio signal. In figure 7.2, a model similar to speech removal is utilised, but the speech-only audio data is not discarded. It is proposed that these data be recycled using a speech obfuscation model, where the content of the speech data is discarded and replaced with a default text that has no real meaning to the occupant.

The home occupant could be recorded while having a private conversation and the proposed methodology will separate the background sounds from the conversational audio data. Then, the speech obfuscation model will use the speaker's sound and a text-to-speech methodology to convert a fabricated text, for example, "I am going to the park", so the vital conversational information from the home occupant will not be used for any further analysis.

Afterwards, the obfuscated speech content will be overlapped with the background-only audio data and used for further analysis. This methodology could be useful in scenarios when additional analysis is needed to understand activities like phone calls. Still, the home occupants do not want to share their conversational audio information. The system also protects visitors, caregivers, and family members when they visit the home occupant, and their conversations can stay private while receiving the benefits of having a passive acoustic system.

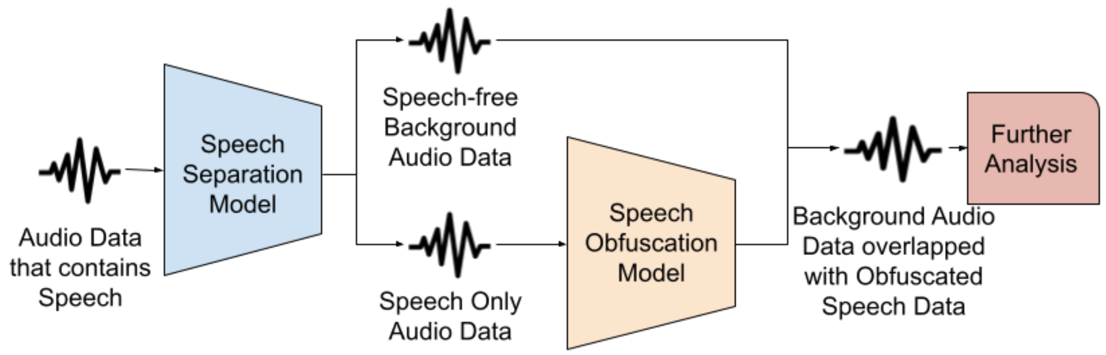

Audio Data that contains Speech
Speech Separation Model
Speech-free Background Audio Data
Speech Only Audio Data
Speech Obfuscation Model
Background Audio Data overlapped with Obfuscated Speech Data
Further Analysis

FIGURE 7.2: Speech Obfuscation Methodology

7.3.3 Hybrid Model

The other part of our proposed methodology in chapter 6 that could be improved is the training part of the speech removal model. The proposed speech removal model is trained on data synthesised from a real environment and speech data collected in a controlled environment. For the Hybrid Model, a framework is proposed that can be deployed with no knowledge of the new environment, and our model will be trained based on the data from that environment.

As shown in 6.2, generic speech removal models are not performing well, but they remove some parts of speech. As shown in figure 7.3, the proposed deployment of two models: the speech removal model as the one suggested in chapter 6 alongside a generic speech removal model, such as the Facebook Denoiser [157].

These two models will run in parallel initially; both will try to remove speech content. Afterwards, the two results will be compared based on how much speech could be detected. Finally, the background audio signal with the less speech detected will move to further analysis, like acoustic activity recognition. If the generic speech removal model performs better than the custom speech removal model, which is expected at the beginning of the experiment in a new environment, the data from the generic model will be used to retrain the custom model online. This process will be repeated with all new audio data that contain speech until the custom speech removal model performs better than the generic speech removal model.

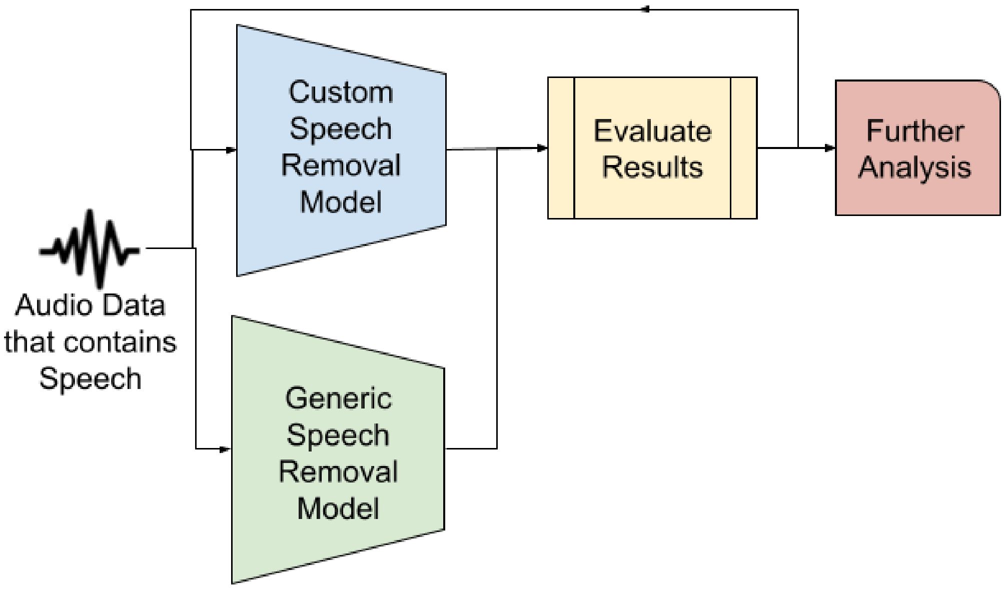

Custom Speech Removal Model
Evaluate Results
Further Analysis
Audio Data that contains Speech
Generic Speech Removal Model

Figure 7.3: Hybrid Model for Speech Removal Methodology

In the work presented in this thesis, data collection should take place for a certain period, and the recorded data needs to be manually labelled. The hybrid model approach could be deployed without the need to manually label the audio data to build the custom speech removal model. Speech-aware acoustic sensing could be performed to analyse the audio signal further.

The main goal of the hybrid model methodology is to ease the deployment's initial setup. The system will be able to be trained on new data from the current environment unsupervised, with no need to do manual labelling and data collection at the beginning of our acoustic environment setup. Every new home environment equipped with an acoustic sensing system will start using the system immediately.

7.3.4 Tiny Model

For the last future extension of this research regarding privacy, this research is focused on removing any conversational data from the audio data, focusing on private conversations, for a privacy-aware acoustic sensing environment. Privacy is a huge area that does not only involve speech data [204]. Uploading the data to a server could be a privacy concern that could be minimised. Having a system that does all the pre-processing on-device could be beneficial. There are multiple ways to tackle this challenge:

- Pre-process data on-device and upload on server speech-free audio data.
- Pre-process data on-device, including further analysis.

Depending on further analysis, the speech removal pipeline and any model used for further analysis should have a small memory footprint so they can be saved and run on the device. For this future development, the focus is on the privacy-preserving data pre-processing part, with our focal point on removing speech data and uploading the speech-free data on the server for further analysis. This methodology will help the system's acceptability from the home occupant and their family. By removing the speech data from the background audio signal before it leaves the always-on recording device, it is possible to significantly enhance the privacy and acceptance of the system. Potential approaches to achieve this can involve following a "teacher-student" model [205] or applying model pruning [206] on the trained model in order to produce models with smaller memory and processing footprint.

7.3.5 Future of Smart Care Homes with Audio as their Main Modality

In chapter 5, the challenges of having an acoustic sensing system in domestic environments were discussed, where the focus is to help them without the burden of managing and charging wearable devices. The reality is that a passive system can be beneficial, but further requirements arise because of the nature of the always-on audio recording devices installed in people's private homes. There is an understanding that these problems could be solved, as mentioned in chapter 6, by developing a suitable pipeline that can satisfy the needs of the home occupants, family members and carers whilst being able to use audio as a modality for passive sensing.

Smart care homes can greatly benefit from utilising a passive system that supports the elderly living at home. Continuous health monitoring through acoustic sensing and a voice-enabled AI assistant that provides tailored companionship to the elderly could relieve stress from the elderly's family and carers. Identifying cognitive deterioration, as mentioned in chapter 4, through acoustic sensing is revolutionising caring capabilities. Where previously, the carer needed to be aware of any changes in the elderly's daily routine, using an automated system can indicate how the elderly are doing, and the carer could support them accordingly.

One possible area for further work is using remote monitoring and alerting systems for caregivers or family members. Such systems could detect anomalies or falls and indicate the urgency of the alert, allowing swift assistance to be provided. Additionally, voice analysis tools could be used to detect emotional changes in elderly individuals' speech data, which could be useful for virtual visits through a smart care home system. Overall, smart care systems that rely on audio sensing have the potential to provide preventive and compassionate solutions for elderly individuals who want to live at home. These systems can improve their quality of life, assist caregivers, and create a secure, autonomous, and respectful environment prioritising privacy.

7.4 Final Reflections

The study has now been completed. It explored the possibility of using acoustic sensing to track changes in a home environment. Based on existing knowledge and the work of other researchers, it was concluded that a novel dimensionality reduction algorithm could be helpful. It was also proven that activity changes could be detected using pattern matching with dynamic time warping.

In the second part of the study, the focus was shifted to privacy regarding speech, and a set of requirements was deemed necessary for the continuation of the research. Popular machine learning models were utilised; however, modifications were made to adapt to the specific requirements of a real-life environment, where labelled data are limited. A framework was established to guide the final part of the study. The speech removal methodology was designed to consider all requirements from the previous section regarding real-life deployment, particularly prioritising the privacy of those living in such environments.

As smart devices have become commonplace in households, new ethical considerations regarding data collection in home environments are being faced. Researchers are recognised as responsible for exploring ways to safeguard individual privacy in the context of advanced smart technology. It is emphasised that no individual should have to compromise their privacy for the sake of advanced smart home technology or other smart devices.

The field of acoustic sensing in real-world environments has a potential for further development and application. I hope that more researchers and research groups will be interested in this area and that the findings presented in this thesis will prove helpful to them. Additionally, anyone working in an application domain related to this field may also benefit from the insights and knowledge gained from this research.

Chapter 8

Appendix

8.1 Case Study

As part of the research, some involvement in the ADAPTIVE project [207] gave a better understanding of the practical work in deploying audio sensing systems. Within this project, audio was experimented with as a modality for detecting footsteps by a company. Before this, the company employed a multi-sensory approach to support the elderly with smart home devices. Although this project was not directly related to the thesis, valuable insights into the workings of a smart care home were gained, particularly regarding data collection, privacy concerns, and deployment. Challenges and considerations arising from the deployment of sensing technology in a real-world setting were witnessed first-hand, and this experience has informed the research path.

AI-based Dementia Assistive & Passive Technology for non-Invasive Elderly care (ADAPTIVE) [207] was a collaborative project between MiiCare [6], Univesity of Kent, East Kent Hospitals University NHS Foundation Trust and Bristol City Council. Falling is a common issue in the elderly [208, 209], and if the gait of the elderly is not stable, that might contribute to a fall. The ADAPTIVE project aims to use acoustic sensing to detect alternations in the gait of occupants in real-life deployments. The smart home devices that contain microphones were installed in single-occupant households and care homes for data collection and deployment.

8.1.1 Real Life Deployment

Some stages need to be considered before the ADAPTIVE project can take place, as shown in figure 8.1. After the sensor installation, a 4 week period of data collection was done for each house deployment. Because of the nature of the deployment, no cameras were installed, nor was human observation, which means that manual data labelling was needed after the data were collected. The collaborators employed external contractors for this purpose, and the contractors were advised to label when three or more footsteps were taking place. That was a long and expensive process. Afterwards, the Machine Learning (ML) Model was trained based on the collected data that were labelled, and then the ML model was deployed to the single-occupant homes for prediction.

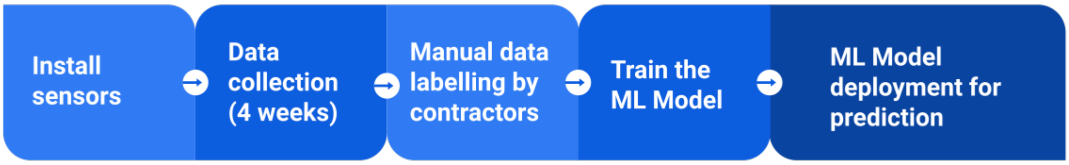

Install sensors
Data collection (4 weeks)
Manual data labelling by contractors
Train the ML Model
ML Model deployment for prediction

FIGURE 8.1: Deployment stages from the case study

A part of the research project was to deploy the smart home device, miiCube, shown in figure 8.2, in one hundred single-occupant households. These homes are primarily inhabited by elderly individuals who receive frequent visits from their caregivers or family members. It is worth noting that some of these elderly individuals may be suffering from mild to moderate dementia. The purpose of this study was to assess the ability to detect footstep and alternation in their footstep pattern using audio as a modality.

The device deployed in these homes was set to record sound in windows of 10 seconds when the Signal to Noise Ratio was above 0 dB. In other terms, the device was recording 10 seconds of audio when the device was detecting that there

FIGURE 8.2: miiCube device by miiCare

was a sound in the room where the device was installed. After the recording, the audio was uploaded to the company's servers, where audio analysis was happening. Because of the nature of these deployments, human observation or the installation of cameras to set the ground truth was not possible. The company had to receive support from external contractors, and they relied on manual labelling through human hearing. Specifically, a cohort of external contractors was employed to manually hear and annotate audio datasets with relevant walking events, and timestamps of each footstep.

Manual annotation was labour-intensive and expensive because it took a long time. From our observation data, collecting labelled data in this kind of real deployment is very hard or maybe impossible without the help of camera sensors or human observers who manually note down activities and times that happen. However, such approaches can be very intrusive and expensive, making them unfeasible for large scale, and long-term collection of labelled data.

While the data collection was taking place, carers of the elderly who were visiting the homes of the elderly expressed their privacy concern. The system recorded people having private conversations during the study because the audio was continuously recorded. This raised concerns about the privacy and security of the individuals visiting or living in the house. Therefore, appropriate measures need to be taken to address this issue and ensure the safety and confidentiality of the elderly people, their family members and carers.

Based on our experience, this project is viewed as a case study, providing an understanding of the benefits and hurdles associated with a smart care home. From what was observed, the collection of labelled data in a real-world environment is not considered an easy task. If cameras or human observers are not present, manual labelling needs to be conducted, which is regarded as time-consuming and expensive. In addition to ensuring the privacy of the occupants, a system needs to be created that is capable of removing any speech content from the background, allowing for the performance variation to be showcased after the speech is removed in regular acoustic sensing classification.

Bibliography

[1] Dhanya Pramod. Assistive technology for elderly people: State of the art review and future research agenda. *Science & Technology Libraries*, 42(1): 85–118, 2023.

[2] NHS. Virtual wards, 2023. URL https://www.england.nhs.uk/virtual-wards/.

[3] Sumit Majumder, Emad Aghayi, Moein Noferesti, Hamidreza Memarzadeh-Tehran, Tapas Mondal, Zhibo Pang, and M Jamal Deen. Smart homes for elderly healthcare—recent advances and research challengesxvx. *Sensors*, 17 (11):2496, 2017.

[4] Sandeep Kumar Vashist, E Marion Schneider, and John HT Luong. Commercial smartphone-based devices and smart applications for personalized healthcare monitoring and management. *Diagnostics*, 4(3):104–128, 2014.

[5] Muhammad Ashfaq, Jiang Yun, and Shubin Yu. My smart speaker is cool! perceived coolness, perceived values, and users' attitude toward smart speakers. *International Journal of Human–Computer Interaction*, 37(6):560–573, 2021.

[6] Kelvin Summoogum, John Wall, Ophir Levy, Shan Mantri, Andy Smith, Debayan Das, Huy Phan, Ian McLoughlin, Terry Whittaker, and Dan Parsons. Miicare. https://www.miicare.co.uk/, 2023.

[7] Rebecca Adaimi, Howard Yong, and Edison Thomaz. Ok google, what am i doing? acoustic activity recognition bounded by conversational assistant

interactions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(1):1–24, 2021.

[8] Louis G Pol. Rapid growth in the elderly population of the world. *Brain and Spine Surgery in the Elderly*, pages 3–15, 2017.

[9] Helen C Kales, Laura N Gitlin, and Constantine G Lyketsos. Assessment and management of behavioral and psychological symptoms of dementia. *Bmj*, 350, 2015.

[10] Pavlos Nicolaou and Christos Efstratiou. Tracking daily routines of elderly users through acoustic sensing: An unsupervised learning approach. In *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, pages 391–396. IEEE, 2022.

[11] Mohsen Amiribesheli and Hamid Bouchachia. A tailored smart home for dementia care. *Journal of Ambient Intelligence and Humanized Computing*, 9:1755–1782, 2018.

[12] Sebastiaan TM Peek, Eveline JM Wouters, Joost Van Hoof, Katrien G Luijkx, Hennie R Boeije, and Hubertus JM Vrijhoef. Factors influencing acceptance of technology for aging in place: a systematic review. *International journal of medical informatics*, 83(4):235–248, 2014.

[13] Charlie Wilson, Tom Hargreaves, and Richard Hauxwell-Baldwin. Smart homes and their users: a systematic analysis and key challenges. *Personal and Ubiquitous Computing*, 19:463–476, 2015.

[14] Marie Chan, Daniel Estève, Christophe Escriba, and Eric Campo. A review of smart homes—present state and future challenges. *Computer methods and programs in biomedicine*, 91(1):55–81, 2008.

[15] Diane J Cook. How smart is your home? *Science*, 335(6076):1579–1581, 2012.

[16] A. Monteriù, M. Prist, E. Frontoni, S. Longhi, F. Pietroni, S. Casaccia, and M. Revel, G. A smart sensing architecture for domestic monitoring: Methodological approach and experimental validation. *Sensors*, 18:2310, 2018. doi: 10.3390/s18072310.

[17] Si H. Fang H., Tang P. Feature selections using minimal redundancy maximal relevance algorithm for human activity recognition in smart home environments. *Journal of Healthcare Engineering*, 2020:1–13, 2020. doi: 10.1155/2020/8876782.

[18] Fermina Rojo-Pérez, Gloria Fernández-Mayoralas, Maria-João Forjaz, María-Eugenia Prieto-Flores, and Pablo Martínez-Martín. Residential environment and health conditions among older-adults in community-dwelling in spain: What influences quality of life? *Environmental gerontology in Europe and Latin America: Policies and perspectives on environment and aging*, pages 149–174, 2016.

[19] John J. Jr. Callahan. *Introduction: Aging in Place*, chapter 1-3. Baywood Publishing Company, Amityville, New York, 1993.

[20] Gavin J Andrews, Malcolm Cutchin, Kevin McCracken, David R Phillips, and Janine Wiles. Geographical gerontology: The constitution of a discipline. *Social Science & Medicine*, 65(1):151–168, 2007.

[21] Xavier Badia, Montserrat Roset, Michael Herdman, and Paul Kind. A comparison of united kingdom and spanish general population time trade-off values for eq-5d health states. *Medical Decision Making*, 21(1):7–16, 2001.

[22] Helen Bartlett and Nancy Peel. *Healthy Ageing in the Community*, pages 98–109. Routledge, Taylor & Francis Group, London, 2005.

[23] Better Lives Joint Programming Initiative More Years. More years better lives. https://jp-demographic.eu/?set_language=e, 2023.

[24] Lu H. Zheng X., Tan H. Research on the intention of use of remote monitoring technology for the elderly based on technology acceptance model.

Advances in Social Science, Education and Humanities Research, 2021. doi: 10.2991/assehr.k.211011.009.

[25] Howland J., Hackman H. H., Taylor A. C., O'Hara K. L., Liu J. K., and Brusch J. L. Older adult fall prevention practices among primary care providers at accountable care organizations: a pilot study. *Plos One*, 13:e0205279, 2018. doi: 10.1371/journal.pone.0205279.

[26] Samantha J Heintzelman and Laura A King. Routines and meaning in life. *Personality and social psychology bulletin*, 45(5):688–699, 2019.

[27] Bayard E Lyons, Daniel Austin, Adriana Seelye, Johanna Petersen, Jonathan Yeargers, Thomas Riley, Nicole Sharma, Nora Mattek, Katherine Wild, Hiroko Dodge, et al. Pervasive computing technologies to continuously assess alzheimer's disease progression and intervention efficacy. *Frontiers in aging neuroscience*, 7:102, 2015.

[28] Zoe Arvanitakis, Raj C Shah, and David A Bennett. Diagnosis and management of dementia. *Jama*, 322(16):1589–1599, 2019.

[29] Mariola Bidzan, Leszek Bidzan, and Ilona Bidzan-Bluma. Neuropsychiatric symptoms and faster progression of cognitive impairments as predictors of risk of conversion of mild cognitive impairment to dementia. *Archives of Medical Science*, 13(5):1168–1177, 2017.

[30] Valeria Manera, Emmanuelle Chapoulie, Jérémy Bourgeois, Rachid Guerchouche, Renaud David, Jan Ondrej, George Drettakis, and Philippe Robert. A feasibility study with image-based rendered virtual reality in patients with mild cognitive impairment and dementia. *PloS one*, 11(3):e0151487, 2016.

[31] Sara Saeedi, Adel Moussa, and Naser El-Sheimy. Context-aware personal navigation using embedded sensor fusion in smartphones. *Sensors*, 14(4): 5742–5767, 2014.

[32] Sumit Majumder and M Jamal Deen. Smartphone sensors for health monitoring and diagnosis. *Sensors*, 19(9):2164, 2019.

[33] Francisco M Garcia-Moreno, Maria Bermudez-Edo, José Luis Garrido, Estefanía Rodríguez-García, José Manuel Pérez-Mármol, and María José Rodríguez-Fórtiz. A microservices e-health system for ecological frailty assessment using wearables. *Sensors*, 20(12):3427, 2020.

[34] Elham Nejadsadeghi, Shahab Papi, Maria Cheraghi, Samaneh Norouzi, Fatemeh Hosseini, and Ghodratollah Shakeri Nejad. Factor affecting the activities of daily living among aging people during the covid-19 pandemic–a structural equation modelling. *Menopause Review/Przegląd Menopauzalny*, 21(2):111–116, 2022.

[35] Tahir Masud and Robert O Morris. Epidemiology of falls. *Age and ageing*, 30:3–7, 2001.

[36] Panagiotis Tsinganos and Athanassios Skodras. A smartphone-based fall detection system for the elderly. In *Proceedings of the 10th International Symposium on Image and Signal Processing and Analysis*, pages 53–58. IEEE, 2017.

[37] George Vavoulas, Charikleia Chatzaki, Thodoris Malliotakis, Matthew Pediaditis, and Manolis Tsiknakis. The mobiact dataset: Recognition of activities of daily living using smartphones. In *International conference on information and communication technologies for ageing well and e-health*, volume 2, pages 143–151. SciTePress, 2016.

[38] Suranga Seneviratne, Yining Hu, Tham Nguyen, Guohao Lan, Sara Khalifa, Kanchana Thilakarathna, Mahbub Hassan, and Aruna Seneviratne. A survey of wearable devices and challenges. *IEEE Communications Surveys & Tutorials*, 19(4):2573–2620, 2017.

[39] Milad Asgari Mehrabadi, Iman Azimi, Fatemeh Sarhaddi, Anna Axelin, Hannakaisa Niela-Vilén, Saana Myllyntausta, Sari Stenholm, Nikil Dutt, Pasi Liljeberg, Amir M Rahmani, et al. Sleep tracking of a commercially

available smart ring and smartwatch against medical-grade actigraphy in everyday settings: instrument validation study. *JMIR mHealth and uHealth*, 8 (11):e20465, 2020.

[40] Zhuo Wang, Vignesh Ramamoorthy, Udi Gal, and Allon Guez. Possible life saver: A review on human fall detection technology. *Robotics*, 9(3):55, 2020.

[41] Farhad Shahmohammadi, Anahita Hosseini, Christine E. King, and Majid Sarrafzadeh. Smartwatch Based Activity Recognition Using Active Learning. *Proceedings - 2017 IEEE 2nd International Conference on Connected Health: Applications, Systems and Engineering Technologies, CHASE 2017*, pages 321–329, 2017. doi: 10.1109/CHASE.2017.115.

[42] Sourav Bhattacharya and Nicholas D. Lane. From smart to deep: Robust activity recognition on smartwatches using deep learning. *2016 IEEE International Conference on Pervasive Computing and Communication Workshops, PerCom Workshops 2016*, 2016. doi: 10.1109/PERCOMW.2016.7457169.

[43] Gary M. Weiss, Jessica L. Timko, Catherine M. Gallagher, Kenichi Yoneda, and Andrew J. Schreiber. Smartwatch-based activity recognition: A machine learning approach. *3rd IEEE EMBS International Conference on Biomedical and Health Informatics, BHI 2016*, pages 426–429, 2016. doi: 10.1109/BHI.2016.7455925.

[44] Falin Wu, Hengyang Zhao, Yan Zhao, and Haibo Zhong. Development of a wearable-sensor-based fall detection system. *International journal of telemedicine and applications*, 2015:2–2, 2015.

[45] Pervasive Systems Team. esense, 2018. esense.io.

[46] F. Kawsar, C. Min, A. Mathur, and A. Montanari. Earables for personal-scale behavior analytics. *IEEE Pervasive Computing*, 17(3):83–89, 2018.

[47] Stefan Liebich, Johannes Fabry, Peter Jax, and Peter Vary. Signal processing challenges for active noise cancellation headphones. In *Speech Communication; 13th ITG-Symposium*, pages 1–5. VDE, 2018.

[48] Piero Rivera Benois, Patrick Nowak, Etienne Gerat, Muhammad Salman, and Udo Zölzer. Improving the performance of an active noise cancelling headphones prototype. In *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, volume 259, pages 889–900. Institute of Noise Control Engineering, 2019.

[49] Jenni Radun, Iida-Kaisa Tervahartiala, Ville Kontinen, Jukka Keränen, and Valtteri Hongisto. Do active noise-cancelling headphones' influence performance, stress, or experience in office context? *Building and Environment*, 266:112102, 2024.

[50] Romit Roy Choudhury. Earable computing: A new area to think about. In *Proceedings of the 22nd International Workshop on Mobile Computing Systems and Applications*, pages 147–153, 2021.

[51] Lillian Graham. A looming epidemic: Personal music player use in young adults and concern for future noise-induced hearing loss. *International Journal of High School Research*, 5(5), 2023.

[52] Fahim Kawsar, Chulhong Min, Akhil Mathur, and Alessandro Montanari. Earables for personal-scale behavior analytics. *IEEE Pervasive Computing*, 17(3):83–89, 2018.

[53] A. P. Vancea and I. Orha. Smart home automation and monitoring system. *Carpathian Journal of Electronic and Computer Engineering*, 11(1):40–43, 3918. doi: doi:10.2478/cjece-2018-0007.

[54] Google. Nest thermostat. https://store.google.com/us/product/nest_thermostat?hl=en-US, 2023.

[55] Parisa Rashidi and Alex Mihailidis. A survey on ambient-assisted living tools for older adults. *IEEE journal of biomedical and health informatics*, 17 (3):579–590, 2012.

[56] Wing WY Ng, Shichao Xu, Ting Wang, Shuai Zhang, and Chris Nugent. Radial basis function neural network with localized stochastic-sensitive autoencoder for home-based activity recognition. *Sensors*, 20(5):1479, 2020.

[57] Sinan Chen, Sachio Saiki, and Masahide Nakamura. Toward affordable and practical home context recognition:—framework and implementation with image-based cognitive api—. *International Journal of Networked and Distributed Computing*, 8(1):16–24, 2019.

[58] Fco Javier Ordónez, Paula De Toledo, and Araceli Sanchis. Activity recognition using hybrid generative/discriminative models on home environments using binary sensors. *Sensors*, 13(5):5460–5477, 2013.

[59] Shuai Zhang, Wing WY Ng, Jianjun Zhang, Chris D Nugent, Naomi Irvine, and Ting Wang. Evaluation of radial basis function neural network minimizing l-gem for sensor-based activity recognition. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–11, 2023.

[60] TLM Van Kasteren, Gwenn Englebienne, and Ben JA Kröse. Activity recognition using semi-markov models on real world smart home datasets. *Journal of ambient intelligence and smart environments*, 2(3):311–325, 2010.

[61] World Health Organisation. Ageing: Global population - world health organization (who). https://www.who.int/news-room/questions-and-answers/item/population-ageing, 2010.

[62] Taekjin Han, Wonho Kang, and Gyunghyun Choi. Ir-uwb sensor based fall detection method using cnn algorithm. *Sensors*, 20(20):5948, 2020.

[63] Jie Yin, Qing Zhang, and Mohan Karunanithi. Unsupervised daily routine and activity discovery in smart homes. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5497–5500. IEEE, 2015.

[64] Shwetambara Kekade, Chung-Ho Hseieh, Md. Mohaimenul Islam, Suleman Atique, Abdulwahed Mohammed Khalfan, Yu-Chuan Li, and Shabbir Syed

Abdul. The usefulness and actual use of wearable devices among the elderly population. *Computer Methods and Programs in Biomedicine*, 153:137–159, 2018. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2017.10.008.

[65] A. Behzadi, A. S. Shamloo, K. Mouratis, G. Hindricks, A. Arya, and A. Bollmann. Feasibility and reliability of smartwatch to obtain 3-lead electrocardiogram recordings. *Sensors*, 20:5074, 2020. doi: 10.3390/s20185074.

[66] Eurostat. Database: Your key to european statistics, 2019.

[67] Ranganathan Chandrasekaran, Vipanchi Katthula, and Evangelos Moustakas. Too old for technology? use of wearable healthcare devices by older adults and their willingness to share health data with providers. *Health Informatics Journal*, 27(4):14604582211058073, 2021.

[68] Apple. First iphone. https://www.apple.com/newsroom/2007/01/09Apple-Reinvents-the-Phone-with-iPhone/, Apr 2007.

[69] Stephanie Hui-Wen Chuah, Philipp A Rauschnabel, Nina Krey, Bang Nguyen, Thurasamy Ramayah, and Shwetak Lade. Wearable technologies: The role of usefulness and visibility in smartwatch adoption. *Computers in Human Behavior*, 65:276–284, 2016.

[70] Ioana Iancu and Bogdan Iancu. Designing mobile technology for elderly. a theoretical overview. *Technological Forecasting and Social Change*, 155: 119977, 2020.

[71] Victor P Cornet and Richard J Holden. Systematic review of smartphone-based passive sensing for health and wellbeing. *Journal of biomedical informatics*, 77:120–132, 2018.

[72] Michalis Vrigkas, Christophoros Nikou, and Ioannis A Kakadiaris. A review of human activity recognition methods. *Frontiers in Robotics and AI*, 2:28, 2015.

[73] Hamid Medjahed, Dan Istrate, Jerome Boudy, Jean-Louis Baldinger, and Bernadette Dorizzi. A pervasive multi-sensor data fusion for smart home

healthcare monitoring. In *2011 IEEE international conference on fuzzy systems (FUZZ-IEEE 2011)*, pages 1466–1473. IEEE, 2011.

[74] Po-Huan Chou, Yu-Liang Hsu, Wan-Lung Lee, Yu-Chen Kuo, Chih-Chien Chang, Yuan-Sheng Cheng, Hsing-Cheng Chang, Shyan-Lung Lin, Shih-Chin Yang, and Hsin-Hung Lee. Development of a smart home system based on multi-sensor data fusion technology. In *2017 international conference on applied system innovation (ICASI)*, pages 690–693. IEEE, 2017.

[75] Gibson Chimamiwa, Marjan Alirezaie, Federico Pecora, and Amy Loutfi. Multi-sensor dataset of human activities in a smart home environment. *Data in Brief*, 34:106632, 2021.

[76] Pedro Chahuara, Anthony Fleury, François Portet, and Michel Vacher. Online human activity recognition from audio and home automation sensors: Comparison of sequential and non-sequential models in realistic smart homes 1. *Journal of ambient intelligence and smart environments*, 8(4):399–422, 2016.

[77] Chloë Brown, Jagmohan Chauhan, Andreas Grammenos, Jing Han, Apinan Hasthanasombat, Dimitris Spathis, Tong Xia, Pietro Cicuta, and Cecilia Mascolo. Exploring automatic diagnosis of covid-19 from crowdsourced respiratory sound data. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3474–3484, 2020.

[78] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.

[79] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Tut database for acoustic scene classification and sound event detection. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pages 1128–1132. IEEE, 2016.

[80] Tuomas Virtanen, Mark D Plumbley, and Dan Ellis. *Computational analysis of sound scenes acolornd events*. Springer, 2018.

[81] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Assessment of human and machine performance in acoustic scene classification: Dcase 2016 case study. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 319–323. IEEE, 2017.

[82] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio Spectrogram Transformer. In *Proc. Interspeech 2021*, pages 571–575, 2021. doi: 10.21437/Interspeech.2021-698.

[83] Biyun Ding, Tao Zhang, Chao Wang, Ganjun Liu, Jinhua Liang, Ruimin Hu, Yulin Wu, and Difei Guo. Acoustic scene classification: a comprehensive survey. *Expert Systems with Applications*, page 121902, 2023.

[84] Dan Stowell, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange, and Mark D Plumbley. Detection and classification of acoustic scenes and events. *IEEE Transactions on Multimedia*, 17(10):1733–1746, 2015.

[85] Andrey Temko, Robert Malkin, Christian Zieger, Dušan Macho, Climent Nadeu, and Maurizio Omologo. Clear evaluation of acoustic event detection and classification systems. In *International Evaluation Workshop on Classification of Events, Activities and Relationships*, pages 311–322. Springer, 2006.

[86] Huy Phan, Marco Maaß, Radoslaw Mazur, and Alfred Mertins. Random regression forests for acoustic event detection and classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(1):20–31, 2014.

[87] Jens Schröder, Stefan Goetze, and Jörn Anemüller. Spectro-temporal gabor filterbank features for acoustic event detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(12):2198–2208, 2015.

[88] Kang Li, Yan Song, Li-Rong Dai, Ian McLoughlin, Xin Fang, and Lin Liu. Ast-sed: An effective sound event detection method based on audio spectrogram transformer. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10096853.

[89] Huy Phan, Philipp Koch, Fabrice Katzberg, Marco Maass, Radoslaw Mazur, Ian McLoughlin, and Alfred Mertins. What makes audio event detection harder than classification? In *2017 25th European signal processing conference (EUSIPCO)*, pages 2739–2743. IEEE, 2017.

[90] Annamaria Mesaros, Toni Heittola, Antti Eronen, and Tuomas Virtanen. Acoustic event detection in real life recordings. In *2010 18th European signal processing conference*, pages 1267–1271. IEEE, 2010.

[91] Xiaodan Zhuang, Xi Zhou, Mark A Hasegawa-Johnson, and Thomas S Huang. Real-world acoustic event detection. *Pattern recognition letters*, 31(12):1543–1551, 2010.

[92] Xi Zhou, Xiaodan Zhuang, Ming Liu, Hao Tang, Mark Hasegawa-Johnson, and Thomas Huang. Hmm-based acoustic event detection with adaboost feature selection. In *International Evaluation Workshop on Rich Transcription*, pages 345–353. Springer, 2007.

[93] World Health Organisation. Falls. https://www.who.int/news-room/fact-sheets/detail/falls, 2021.

[94] Michael Cheffena. Fall detection using smartphone audio features. *IEEE Journal of Biomedical and Health Informatics*, 20(4):1073–1080, 2016. ISSN 21682194. doi: 10.1109/JBHI.2015.2425932.

[95] John R Koza, Forrest H Bennett, David Andre, and Martin A Keane. Automated design of both the topology and sizing of analog electrical circuits using genetic programming. *Artificial intelligence in design'96*, pages 151–170, 1996.

[96] IBM. https://www.ibm.com/topics/machine-learning, 2024.

[97] Harrison Kinsley and Daniel Kukieła. *Neural Networks from scratch in Python*. Harrison Kinsley, 2020.

[98] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021.

[99] David A Forsyth and Jean Ponce. *Computer vision: a modern approach*. prentice hall professional technical reference, 2002.

[100] Suman K Saksamudre, PP Shrishrimal, and RR Deshmukh. A review on different approaches for speech recognition system. *International Journal of Computer Applications*, 115(22), 2015.

[101] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature Medicine*, pages 1–11, 2023.

[102] Romain Serizel, Victor Bisot, Slim Essid, and Gaël Richard. Acoustic features for environmental sound analysis. *Computational analysis of sound scenes and events*, pages 71–101, 2018.

[103] Gert Dekkers, Steven Lauwereins, Bart Thoen, Mulu Weldegebreal Adhana, Henk Brouckxon, Bertold Van den Bergh, Toon Van Waterschoot, Bart Vanrumste, Marian Verhelst, and Peter Karsmakers. The sins database for detection of daily activities in a home environment using an acoustic sensor network. *Detection and Classification of Acoustic Scenes and Events 2017*, pages 1–5, 2017.

[104] World Health Organisation. Ageing and health. https://www.who.int/news-room/fact-sheets/detail/ageing-and-health, 2022.

[105] World Health Organisation. Disability. https://www.who.int/en/news-room/fact-sheets/detail/disability-and-health, 2023.

[106] Zied Mnasri, Stefano Rovetta, and Francesco Masulli. Anomalous sound event detection: A survey of machine learning based methods and applications. *Multimedia Tools and Applications*, pages 1–50, 2022.

[107] P Dhanalakshmi, Sengottayan Palanivel, and Vennila Ramalingam. Classification of audio signals using aann and gmm. *Applied soft computing*, 11 (1):716–723, 2011.

[108] Elham Babaee, Nor Badrul Anuar, Ainuddin Wahid Abdul Wahab, Shahaboddin Shamshirband, and Anthony T Chronopoulos. An overview of audio event detection methods from feature extraction to classification. *applied artificial intelligence*, 31(9-10):661–714, 2017.

[109] Frederic Font, Gerard Roma, and Xavier Serra. Sound sharing and retrieval. *Computational analysis of sound scenes and events*, pages 279–301, 2018.

[110] Md. Afzal Hossan, Sheeraz Memon, and Mark A Gregory. A novel approach for mfcc feature extraction. In *2010 4th International Conference on Signal Processing and Communication Systems*, pages 1–5, 2010. doi: 10.1109/ICSPCS.2010.5709752.

[111] Minh Pham, Yehenew Mengistu, Ha Do, and Weihua Sheng. Delivering home healthcare through a cloud-based smart home environment (coshe). *Future Generation Computer Systems*, 81:129–140, 2018.

[112] Marc Peter Deisenroth, A Aldo Faisal, and Cheng Soon Ong. *Mathematics for machine learning*. Cambridge University Press, 2020.

[113] Jake VanderPlas. *Python data science handbook: Essential tools for working with data*. " O'Reilly Media, Inc.", 2016.

[114] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[115] Carl Kingsford and Steven L Salzberg. What are decision trees? *Nature biotechnology*, 26(9):1011–1013, 2008.

[116] Leo Breiman. Bagging predictors. *Machine learning*, 24:123–140, 1996.

[117] Jacob Goldberger, Geoffrey E Hinton, Sam Roweis, and Russ R Salakhutdinov. Neighbourhood components analysis. *Advances in neural information processing systems*, 17, 2004.

[118] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135. IEEE, 2017.

[119] Per-Erik Danielsson. Euclidean distance mapping. *Computer Graphics and image processing*, 14(3):227–248, 1980.

[120] Derek JS Robinson. *An introduction to abstract algebra.* Walter de Gruyter, 2003.

[121] Paul E Black. *Dictionary of algorithms and data structures.* Paul E. Black, 1998.

[122] Anthony C Thompson. *Minkowski geometry.* Cambridge University Press, 1996.

[123] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

[124] Qiuqiang Kong, Changsong Yu, Yong Xu, Turab Iqbal, Wenwu Wang, and Mark D Plumbley. Weakly labelled audioset tagging with attention neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(11):1791–1802, 2019.

[125] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2021.

[126] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Acoustic scene classification: an overview of dcase 2017 challenge entries. In *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 411–415. IEEE, 2018.

[127] Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018, 2015.

[128] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

[129] KULeuven ADVISE research group. Sins. https://github.com/KULeuvenADVISE/SINS_database, 2017.

[130] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[131] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing*, 22(10): 1533–1545, 2014.

[132] Federico Cruciani, Anastasios Vafeiadis, Chris Nugent, Ian Cleland, Paul McCullagh, Konstantinos Votis, Dimitrios Giakoumis, Dimitrios Tzovaras, Liming Chen, and Raouf Hamzaoui. Feature learning for human activity recognition using convolutional neural networks: A case study for inertial measurement unit and audio data. *CCF Transactions on Pervasive Computing and Interaction*, 2(1):18–32, 2020.

[133] Jakob Abeßer. A review of deep learning based methods for acoustic scene classification. *Applied Sciences*, 10(6), 2020.

[134] Hao Duc Do, Son Thai Tran, and Duc Thanh Chau. Speech separation in the frequency domain with autoencoder. *J. Commun.*, 15(11):841–848, 2020.

[135] Li Deng, Ossama Abdel-Hamid, and Dong Yu. A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6669–6673. IEEE, 2013.

[136] Tara N Sainath, Brian Kingsbury, George Saon, Hagen Soltau, Abdel-rahman Mohamed, George Dahl, and Bhuvana Ramabhadran. Deep convolutional neural networks for large-scale speech tasks. *Neural networks*, 64: 39–48, 2015.

[137] Charmi Jobanputra, Jatna Bavishi, and Nishant Doshi. Human activity recognition: A survey. *Procedia Computer Science*, 155:698–703, 2019.

[138] Zhihua Wang, Zhaochu Yang, and Tao Dong. A review of wearable technologies for elderly care that can accurately track indoor position, recognize physical activities and monitor vital signs in real time. *Sensors*, 17(2):341, 2017.

[139] Ahatsham Hayat, Fernando Morgado-Dias, Bikram Pratim Bhuyan, and Ravi Tomar. Human activity recognition for elderly people using machine and deep learning approaches. *Information*, 13(6):275, 2022.

[140] Meriem Zerkouk and Belkacem Chikhaoui. Spatio-temporal abnormal behavior prediction in elderly persons using deep learning models. *Sensors*, 20 (8):2359, 2020.

[141] Karen Rose, Scott Eldridge, and Lyman Chapin. The internet of things: An overview. *The internet society (ISOC)*, 80:1–50, 2015.

[142] Soe Ye Yint Tun, Samaneh Madanian, and Farhaan Mirza. Internet of things (iot) applications for elderly care: a reflective review. *Aging clinical and experimental research*, 33:855–867, 2021.

[143] IEEE AASP. Dcase challenge. https://dcase.community/, 2013.

[144] Thomas Lidy and Alexander Schindler. Cqt-based convolutional neural networks for audio scene classification. In *DCASE*, pages 60–64, 2016.

[145] Christian Schörkhuber and Anssi Klapuri. Constant-q transform toolbox for music processing. In *7th sound and music computing conference, Barcelona, Spain*, pages 3–64, 2010.

[146] Zhao Lu. Sound event detection and localization based on cnn and lstm. *Detection Classification Acoust. Scenes Events Challenge, Tech. Rep*, 2019.

[147] Sharath Adavanne, Archontis Politis, and Tuomas Virtanen. A multi-room reverberant dataset for sound event localization and detection. *arXiv preprint arXiv:1905.08546*, 2019.

[148] Barry Arons. A review of the cocktail party effect. *Journal of the American Voice I/O society*, 12(7):35–50, 1992.

[149] Pengcheng Li, Yan Song, Ian Vince McLoughlin, Wu Guo, and Li-Rong Dai. An attention pooling based representation learning method for speech emotion recognition. *Interspeech 2018*, 2018.

[150] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. Attention is all you need in speech separation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 21–25. IEEE, 2021.

[151] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 31–35. IEEE, 2016.

[152] Yi Luo, Zhuo Chen, and Takuya Yoshioka. Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 46–50. IEEE, 2020.

[153] Giancarlo Kerg, Bhargav Kanuparthi, Anirudh Goyal ALIAS PARTH GOYAL, Kyle Goyette, Yoshua Bengio, and Guillaume Lajoie. Untangling tradeoffs between recurrence and self-attention in artificial neural networks. *Advances in Neural Information Processing Systems*, 33: 19443–19454, 2020.

[154] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. SpeechBrain: A general-purpose speech toolkit, 2021. arXiv:2106.04624.

[155] Sudharsan Ravichandiran. *Getting Started with Google BERT: Build and train state-of-the-art natural language processing models using BERT*. Packt Publishing Ltd, 2021.

[156] Jaeyoung Kim, Mostafa El-Khamy, and Jungwon Lee. T-gsa: Transformer with gaussian-weighted self-attention for speech enhancement. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6649–6653. IEEE, 2020.

[157] Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi. Real time speech enhancement in the waveform domain. In *Interspeech*, 2020.

[158] Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis Bach. Music source separation in the waveform domain. *arXiv preprint arXiv:1911.13254*, 2019.

[159] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[160] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941. PMLR, 2017.

[161] Cassia Valentini-Botinhao et al. Noisy speech database for training speech enhancement algorithms and tts models, 2017.

[162] CK Reddy, E Beyrami, H Dubey, V Gopal, R Cheng, R Cutler, S Matusevych, R Aichner, A Aazami, S Braun, et al. The interspeech 2020 deep noise suppression challenge: Datasets, subjective speech quality and testing framework. arxiv 2020. *arXiv preprint arXiv:2001.08662*, 2020.

[163] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.

[164] Eduardo Fonseca, Manoj Plakal, Frederic Font, Daniel PW Ellis, and Xavier Serra. Audio tagging with noisy labels and minimal supervision. *arXiv preprint arXiv:1906.02975*, 2019.

[165] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[166] Gert Dekkers, Lode Vuegen, Toon van Waterschoot, Bart Vanrumste, and Peter Karsmakers. Dcase 2018 challenge-task 5: Monitoring of domestic activities based on multi-channel acoustics. *arXiv preprint arXiv:1807.11246*, 2018.

[167] Fatih Erden, Senem Velipasalar, Ali Ziya Alkar, and A Enis Cetin. Sensors in assisted living: A survey of signal and image processing methods. *IEEE Signal Processing Magazine*, 33(2):36–44, 2016.

[168] Pavlos Nicolaou and Christos Efstratiou. Audiohive. https://ubicomp-kent.org/projects/audiohive/, 2023.

[169] Robin IM Dunbar. Do online social media cut through the constraints that limit the size of offline social networks? *Royal Society Open Science*, 3(1): 150292, 2016.

[170] Xingjie Liu, Qi He, Yuanyuan Tian, Wang-Chien Lee, John McPherson, and Jiawei Han. Event-based social networks: linking the online and offline social worlds. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1032–1040, 2012.

[171] James Robert. Pydub. https://github.com/jiaaro/pydub, 2024.

[172] Amazon. Alexa. https://developer.amazon.com/echo, 2023.

[173] World Health Organisation. Dementia. https://www.who.int/news-room/fact-sheets/detail/dementia, 2021.

[174] World Health Organisation. Risk reduction of cognitive decline and dementia. https://www.who.int/publications/i/item/risk-reduction-of-cognitive-decline-and-dementia, 2021.

[175] Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007.

[176] Gineke A Ten Holt, Marcel JT Reinders, and Emile A Hendriks. Multi-dimensional dynamic time warping for gesture recognition. In *Thirteenth annual conference of the Advanced School for Computing and Imaging*, volume 300, page 1, 2007.

[177] Mohammad Shokoohi-Yekta, Bing Hu, Hongxia Jin, Jun Wang, and Eamonn Keogh. Generalizing dtw to the multi-dimensional case requires an adaptive approach. *Data mining and knowledge discovery*, 31(1):1–31, 2017.

[178] Mohammad Shokoohi-Yekta, Jun Wang, and Eamonn Keogh. On the non-trivial generalization of dynamic time warping to the multi-dimensional case. In *Proceedings of the 2015 SIAM international conference on data mining*, pages 289–297. SIAM, 2015.

[179] Meinard Müller. Information retrieval for music and motion. *Information Retrieval for Music and Motion*, 2007.

[180] Joseph B Kruskall. The symmetric time warping algorithm: From continuous to discrete. *Time warps, string edits and macromolecules*, 1983.

[181] Louise Corti, Veerle Van den Eynden, Libby Bishop, and Matthew Woollard. *Managing and sharing research data: A guide to good practice*. Sage, 2019.

[182] Lindasalwa Muda, Mumtaj Begam, and Irraivan Elamvazuthi. Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques. *arXiv preprint arXiv:1003.4083*, 2010.

[183] Emre Cakır, Giambattista Parascandolo, Toni Heittola, Heikki Huttunen, and Tuomas Virtanen. Convolutional recurrent neural networks for polyphonic sound event detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(6):1291–1303, 2017.

[184] Yuan Gong, Yu-An Chung, and James Glass. Psla: Improving audio tagging with pretraining, sampling, labeling, and aggregation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3292–3306, 2021.

[185] Shuteng Niu, Yongxin Liu, Jian Wang, and Houbing Song. A decade survey of transfer learning (2010–2020). *IEEE Transactions on Artificial Intelligence*, 1(2):151–166, 2020.

[186] Silero Team. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. https://github.com/snakers4/silero-vad, 2021.

[187] Branislav Sredojev, Dragan Samardzija, and Dragan Posarac. Webrtc technology overview and signaling solution design and implementation. In *2015*

38th international convention on information and communication technology, electronics and microelectronics (MIPRO), pages 1006–1009. IEEE, 2015.

[188] Jan "Yenda" Trmal. Open speech and language resources. https://www.openslr.org/resources.php, 2024.

[189] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

[190] Google and Google Brain. Tensorflow. https://www.tensorflow.org/, 2024.

[191] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

[192] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021.

[193] Jing Han, Tong Xia, Dimitris Spathis, Erika Bondareva, Chloë Brown, Jagmohan Chauhan, Ting Dang, Andreas Grammenos, Apinan Hasthanasombat, Andres Floto, et al. Sounds of covid-19: exploring realistic performance of audio-based digital testing. *NPJ digital medicine*, 5(1):16, 2022.

[194] Nayan Di, Muhammad Zahid Sharif, Zongwen Hu, Renjie Xue, and Baizhong Yu. Applicability of vggish embedding in bee colony monitoring: comparison with mfcc in colony sound classification. *PeerJ*, 11:e14696, 2023.

[195] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[196] Brian McFee, Matt McVicar, Daniel Faronbi, Iran Roman, Matan Gover, Stefan Balke, Scott Seyfarth, Ayoub Malek, Colin Raffel, Vincent Lostanlen, Benjamin van Niekirk, Dana Lee, Frank Cwitkowitz, Frank Zalkow, Oriol Nieto, Dan Ellis, Jack Mason, Kyungyun Lee, Bea Steers, Emily Halvachs, Carl Thomé, Fabian Robert-Stöter, Rachel Bittner, Ziyao Wei, Adam Weiss, Eric Battenberg, Keunwoo Choi, Ryuichi Yamamoto, CJ Carr, Alex Metsai, Stefan Sullivan, Pius Friesch, Asmitha Krishnakumar, Shunsuke Hidaka, Steve Kowalik, Fabian Keller, Dan Mazur, Alexandre Chabot-Leclerc, Curtis Hawthorne, Chandrashekhar Ramaprasad, Myungchul Keum, Juanita Gomez, Will Monroe, Viktor Andreevitch Morozov, Kian Eliasi, nullmightybofo, Paul Biberstein, N. Dorukhan Sergin, Romain Hennequin, Rimvydas Naktinis, beantowel, Taewoon Kim, Jon Petter Åsen, Joon Lim, Alex Malins, Darío Hereñú, Stef van der Struijk, Lorenz Nickel, Jackie Wu, Zhen Wang, Tim Gates, Matt Vollrath, Andy Sarroff, Xiao-Ming, Alastair Porter, Seth Kranzler, Voodoohop, Mattia Di Gangi, Helmi Jinoz, Connor Guerrero, Abduttayyeb Mazhar, toddrme2178, Zvi Baratz, Anton Kostin, Xinlu Zhuang, Cash TingHin Lo, Pavel Campr, Eric Semeniuc, Monsij Biswal, Shayenne Moura, Paul Brossier, Hojin Lee, and Waldir Pimenta. librosa/librosa: 0.10.1, August 2023. URL https://doi.org/10.5281/zenodo.8252662.

[197] DeLiang Wang and Jitong Chen. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1702–1726, 2018.

[198] Yi Luo and Nima Mesgarani. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 27(8):1256–1266, 2019.

[199] Shoji Makino. *Audio source separation*, volume 433. Springer, 2018.

[200] Clément Delangue. Hugging face. https://huggingface.co/datasets, 2024.

[201] Chandan KA Reddy, Vishak Gopal, Ross Cutler, Ebrahim Beyrami, Roger Cheng, Harishchandra Dubey, Sergiy Matusevych, Robert Aichner, Ashkan

Aazami, Sebastian Braun, et al. The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results. In *INTERSPEECH*, 2020.

[202] Francois Chollet et al. Keras, 2015. URL https://github.com/fchollet/keras.

[203] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. Attention is all you need in speech separation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 21–25. IEEE, 2021.

[204] Iulia Ion, Niharika Sachdeva, Ponnurangam Kumaraguru, and Srdjan Čapkun. Home is safer than the cloud! privacy concerns for consumer cloud storage. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*, pages 1–20, 2011.

[205] Zhong Meng, Jinyu Li, Yong Zhao, and Yifan Gong. Conditional teacher-student learning. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6445–6449. IEEE, 2019.

[206] Michael Zhu and Suyog Gupta. To prune, or not to prune: exploring the efficacy of pruning for model compression. *arXiv preprint arXiv:1710.01878*, 2017.

[207] UK Research and Innovation. Adaptive. https://gtr.ukri.org/projects?ref=68239, 2022.

[208] Efraim Aizen. Falls in patients with dementia. *Harefuah*, 154(5):323–6, 2015.

[209] Eresha Fernando, Michelle Fraser, Jane Hendriksen, Corey H Kim, and Susan W Muir-Hunter. Risk factors associated with falls in older adults with dementia: a systematic review. *Physiotherapy Canada*, 69(2):161–170, 2017.