ROBUST AUDIO CLASSIFICATION: INTEGRATING TV SOUND DETECTION

by

Mohamed Mehdi Bourahla

A dissertation submitted to the faculty of The University of North Carolina at Charlotte in partial fulfillment of the requirements for the degree of Master of Science in Computer Science.

Charlotte

2023

Approved by:

____

Dr. Minwoo Jake Lee

____

Dr. Min Shin

____

Dr. Razvan Bunescu

____

Dr. Jing Yang

Abstract

MOHAMED MEHDI BOURAHLA. Robust Audio Classification: Integrating TV Sound Detection. (Under the direction of DR. MINWOO LEE and DR. MIN SHIN)

The emergence of voice-controlled systems has transformed the way users engage with technology, providing unparalleled ease and accessibility. Nonetheless, these systems encounter difficulties in distinguishing intended commands from unintentional triggers, especially when competing with TV broadcast sounds. This dissertation addresses this issue by creating and testing multiple techniques to differentiate ambient sound environments from TV broadcast audio.

The study presents four unique methods for detecting TV audio: Energy Balance Metric, YAMNet-based TV Audio Detection, Fine-Tuning YAMNet with LSTM Backend, and an Ensemble of YAMNet-LSTM Models. Each method employs audio signal processing and machine learning techniques to provide a comprehensive and diverse perspective on TV sound detection.

Comprehensive evaluations are conducted using various datasets, including EAR-Aging, EAR-Divorce, SINS, and a custom lab dataset. These datasets cover a broad range of real-world situations, including natural ambient recordings and controlled laboratory environments, providing a comprehensive assessment of each technique. The assessment centers on crucial benchmarks such as the ability to detect TV noise, identifying non-TV sounds with precision, and overall performance under different acoustic conditions and hardware setups.

The thesis findings highlight the capabilities and drawbacks of each technique for precisely identifying sounds from TVs, providing valuable insights on their practical use in voice-controlled systems. Particularly, the Ensemble model offers a promising combination of unique features, resulting in improved accuracy and adaptability.

This study contributes innovatively to the ambient sound classification domain, presenting novel solutions to accurately identify TV sounds. The optimization of voice-controlled systems can enhance reliability and improve user experience in everyday environments, opening new avenues for exploration. Future work can extend these methodologies to broader applications, including smart home automation and assistive technologies for people with disabilities.

Acknowledgements

First of all, I would like to thank God the Greatest for giving me the strength and patience to complete this work. Every letter typed on this thesis and every information withheld is by His grace glory to Him.

I extend my deepest gratitude to my supervisors, Dr. Minwoo Jake Lee and Dr. Min Shin, for their invaluable guidance, unwavering support, and insightful feedback throughout the course of this research project. Their expertise and mentorship have been pivotal in shaping the direction and success of this work.

Special thanks are due to Samuel Tate and David Farynyk for their significant contributions and assistance. Their expertise, dedication, and collaborative spirit have been instrumental in advancing this project, and I am deeply appreciative of their efforts and insights.

I am grateful to the faculty members at the College of Computing and Informatics at UNC Charlotte. Their profound knowledge and inspiring lectures have greatly enriched my academic journey and played a crucial role in my intellectual growth.

My sincere appreciation goes to the members of the Charmlab, whose collaboration, support, and camaraderie have been a source of motivation and encouragement. The stimulating discussions and exchange of ideas within this group have significantly enhanced my research experience.

I am also profoundly thankful to Fulbright for granting me the opportunity to pursue my master's degree in the United States. This experience has been transformative, offering a platform for academic and personal growth, and for this, I am immensely grateful.

Finally, I extend my heartfelt gratitude to my parents, my family and cherished friends—Hadjer, Kader, Karim, Lyna, Rosa, Rym and Vijay. Your unwavering support and faith in me have been the bedrock of my journey. Your constant encouragement, coupled with your unconditional love and understanding, have been the driving forces behind my perseverance and achievements.

This journey has been a collaborative effort, and it is with a heart full of gratitude that I acknowledge the contributions of each individual and institution that has been a part of this endeavor.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

MFCCs	Mel Frequency Cepstral Coefficients
GMMs	Gaussian Mixture Models
HMMs	Hidden Markov Models
CNN	Convolutional Neural Network
CRNN	Convolutional Recurrent Neural Network
ASR	Automatic Speech Recognition
SVM	Support Vector Machine
RBF	Radial Basis Function
EAR	Electronically Activated Recorder
SC	Side Conversation
TNR	True Negative Rate
EBM	Energy Balance Metric
FFT	Fast Fourier Transform
ROC	Receiver Operating Characteristic

Chapter 1 Introduction

The prevalence of smart devices in contemporary households has led to an escalating dependence on audio signals for various applications. Consequently, audio classification, the process of sorting audio clips into distinct categories based on their content, has become a crucial area of investigation in the broader scope of machine learning and signal processing [1].

Despite the considerable progress made in the realm of audio classification, significant challenges remain in the field. The vast range of sounds present in real-world settings, coupled with the diverse contexts in which they happen, oftentimes lead to incorrect categorizations, hindering the efficacy of audio-only systems.

To overcome these difficulties, robust audio classification was introduced. Robust audio classification prioritizes accuracy and resilience in the face of real-world audio challenges, unlike conventional audio classification which may struggle with noisy or ambiguous input [2].

Various methods of robust audio classification have been developed over time. Some researchers prioritize identifying subtle audio cues in complex soundscapes [3], while others focus on enhancing noise-cancellation techniques [4].

Within the myriad avenues of robust audio classification, one particularly intriguing aspect involves classifying sounds that emanate from televisions [5]. With televisions being a prevalent media device in many households, their soundscape presents a challenging yet critical field for audio classifiers, as it combines speech, music, and ambient sounds.

Interestingly, a Burger King advertisement hilariously triggered multiple Google Home devices unintentionally [6]. Simply uttering a phrase intended for TV viewers triggered unintended device activations, highlighting the importance of improved sound classification techniques that can distinguish TV sounds from others.

The significance of identifying TV sounds extends beyond preventing accidental triggers. It has the potential to redefine how devices interact in multimedia-rich environments, leading to smarter and intentional interactions. Furthermore, as more and more applications depend on ambient audio analysis, ranging from surveillance systems to data analytics, the ability to precisely distinguish TV sounds becomes increasingly essential.

Problem Statement

Amidst the cacophony of daily life sounds, it can be difficult to distinguish and categorize those emanating from a television. This thesis explores solutions to identifying unintentional audio triggers from TV broadcasts while remaining sensitive to human interactions. By pushing the boundaries of audio classification, this research aims to serve as a benchmark for future innovation in the field.

Chapter 2 Background

The field of audio classification holds great potential but is complex due to its foundational concepts and evolving methodologies. Prior to delving into our research, it's important to establish a comprehensive understanding of the underlying principles, technologies, and challenges that have shaped the field. This chapter delves into audio classification, beginning with an exploration of its fundamental nature and challenges. We then explore how machine and deep learning paradigms have revolutionized audio data processing. Finally, we elucidate the pivotal concept of transfer learning, which often acts as a bridge between generic datasets and specific audio classification tasks. Lastly, we address the detailed acoustic distinctions between human voices and TV noise, establishing the essentiality of precise differentiation in our research. As we proceed through this chapter, readers will acquire the necessary knowledge and context to understand the subsequent discussions and conclusions fully.

2.1 Audio Classification

Audio classification involves analyzing audio recordings and categorizing them into predefined labels or classes [7]. It is a subfield of audio signal processing, which focuses on studying audio signals and their processing methods. Various terms are used to refer to audio classification in literature and different application domains. Terms like ”audio recognition,” ”audio tagging,” ”sound classification,” and ”acoustic event detection” are occasionally utilized alternately or in closely linked domains. Although the fundamental framework pertains to identifying and arranging audio material, subtle discrepancies may exist contingent on the detailed chore, dataset, or implementation in inquiry. For example, ”acoustic event detection” frequently highlights identifying distinct events or incidents in an audio flow, whereas ”audio tagging” may concentrate on assigning one or more tags to a particular audio clip based on its content. Audio classification technology plays a key role in improving the quality and accessibility of modern life. Its adaptability and accuracy underpin a wide range of applications that provide convenience, security, and empowerment.

Audio classification begins with data acquisition, where the quality and quantity of audio samples directly affect the performance of subsequent classification models [5]. These audio samples can be collected from a variety of media, such as microphones, online datasets, or field recordings. Once acquired, the raw audio typically undergoes preprocessing. The main objective of preprocessing is to eliminate noise and other undesired artifacts from the audio signal, enhancing signal quality and subsequently improving speech recognition accuracy. Key preprocessing techniques include noise reduction (to eliminate background and other unwanted noise), voice activity detection (to identify the presence of speech within an audio input), signal normalization (to adjust the audio input to a standard level), and feature extraction (to convert the voice input into a set of features suitable for further processing) [8]. Feature extraction plays a central role in audio classification, pinpointing the distinct characteristics or properties of sounds that serve as discriminative markers. The extensive body of literature in this area highlights a spectrum of techniques, broadly grouped into four categories: cepstral, temporal, image-based, and spectral features. Cepstral features, particularly Mel Frequency Cepstral Coefficients (MFCCs), have gained prominence across diverse audio domains [9, 10]. Temporal features, extracted directly from sound waveforms, include attributes like Zero-Crossing Rate and Energy Entropy [11]. The rise of deep learning has introduced the adoption of image-based features, with spectrograms standing out as effective visual representations of audio signals [12]. Lastly, spectral features encompass transformations of temporal ones, with techniques such as the MPEG-7 feature set proving valuable in various studies [13]. As audio classification continues to evolve, the choice of features remains pivotal, shaping the accuracy and efficiency of the classification models. Before the advent of modern machine learning, audio classification largely relied on these manually generated features, and simple classifiers such as Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs) were used for tasks such as speech recognition and basic sound classification [14]. However, the advent of deep learning initiated a new wave of techniques. Deep neural network-based models, such as the Convolutional Neural Network (CNN), have consistently outperformed traditional classifiers. Various CNN architectures, ranging from PiczackCNN to Resource Adaptive CNN, have been carefully tailored for improved performance in audio classification [15, 16]. Other deep learning architectures such as the Tensor Deep Stacking Network, Convolutional Recurrent Neural Network (CRNN), and Image Recognition Networks (e.g., AlexNet, GoogLeNet) have further expanded the classification landscape. Deep Belief Neural Networks have emerged as a remedy for certain drawbacks of traditional DNNs, providing better efficiency in certain contexts [17]. Despite this diversity, choosing an optimal classifier isn't easy. Researchers must balance performance against computational cost. For example, while deep learning models such as CNNs offer high accuracy, they also require significant computing power. On the other hand, models like SVM or GMM may offer a more balanced tradeoff between accuracy and computational cost. The ideal classifier should therefore not only be accurate, but also computationally efficient and robust to potential noise in the audio data [18]. Building upon this rich foundation, the technique of transfer learning has gained immense popularity in the audio classification domain. Transfer learning has become pivotal in audio classification. It allows models trained on expansive datasets like AudioSet [5] to be fine-tuned for specific tasks, often outperforming models trained from scratch. This approach is not only efficient but also advantageous when the target dataset is limited in size. Prominent pre-trained audio models, such as YAMNet and VGGish exemplify this trend [19, 5]. By leveraging these models, researchers can rapidly prototype and adapt to specialized audio challenges without the need for exhaustive feature extraction or training.

The remarkable progress and achievements of deep learning in contemporary classification tasks are evident, especially in the field of sound and audio classification. Recent research shows that deep learning techniques have greatly improved diverse domains, including environmental sound detection [19], Automatic Speech Recognition (ASR) [20], music/acoustic classification [21], medical diagnostics [22], and more. This advancement in audio classification has greatly benefited individuals with disabilities or limited mobility. The improved accuracy in voice recognition allows for hands-free device control, providing the ability to manage tasks such as adjusting room temperatures, activating lights, or unlocking entrances without physical intervention. Additionally, this feature enhances the functionality of voice-activated digital assistants across various platforms, including smart homes and cars, delivering timely and accurate information to users [23]. In the leisure industry, audio classification's precision is irrefutable. For instance, a groundbreaking smartphone app that distinguishes a baby's cry from background noises, instantly notifying caretakers [24]. Music streaming services utilize this technology to categorize and suggest genres with efficiency, simplifying users' journey in discovering their preferred music without cumbersome searches [25]. From a security standpoint, audio classification has significant benefits. Systems equipped with this advanced technology can detect potentially dangerous sounds such as the shattering of glass or unfamiliar footsteps, providing enhanced protection for homes and businesses [26]. Bioacoustic research uses this tool for monitoring wildlife, which is crucial for conservation and preserving biodiversity [3]. In the commercial sector, audio classification is revitalizing customer service. Automated platforms can now provide more personalized and prompt responses to client inquiries, enhancing the overall user experience while reducing the need for human intervention [27].

Deep learning models in sound recognition systems are susceptible to various challenges that can compromise their performance. A primary concern is environmental noise, which often obscures finer acoustic details and results in the potential loss of significant information [28]. This problem is exacerbated when distinguishing sounds that emanate from devices like televisions. Sounds originating from TV broadcasts, which include a diverse range of audio from dialogues to music tracks, can be particularly misleading for classifiers. Differentiating between actual real-world events and broadcasted sounds remains a significant hurdle, especially in environments where TVs play in the background. Another critical hurdle in constructing a proficient sound recognition system is the procurement of vast and meticulously annotated datasets. Due to concerns over privacy, alongside ethical and legal implications, collecting extensive sound datasets can be challenging. These data limitations critically hinder the efficient training of deep neural networks, as they heavily rely on vast training data to refine their predictive capacities [29]. Traditional methods of audio feature extraction, although foundational to the field, often fall short in identifying superior feature representations, impacting the efficacy of sound recognition systems [30].

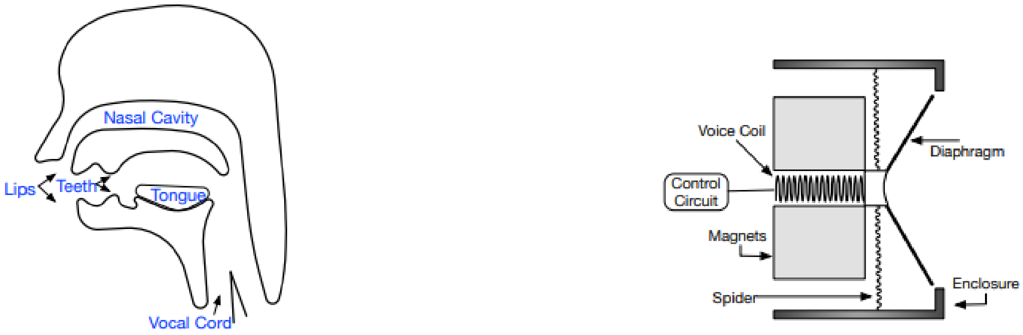
Robustness and adaptability are paramount in crafting a high-performing sound recognition framework. Many existing systems suffer performance degradation under varying conditions, such as reverberations, different types of noise, or channel discrepancies. Moreover, the reliability of these systems is often contingent on expert-driven annotations, which might not always be readily available. Imbalance in data is another substantial impediment, with certain sound categories being underrepresented. This disproportionality can undermine the model's performance, especially in the presence of pervasive environmental noises. Compounding these challenges is the fact that curating a comprehensive sound dataset is both labor-intensive and resource-demanding. This reality underscores the need for innovative solutions and methodologies that can robustly and accurately classify audio, even amidst these prevalent challenges.

2.2 Acoustic Difference Between Human Voices and TV Noise

Understanding the differences between human vocal sounds and electronic noise is essential for any audio classification or processing task. Clear and concise comprehension of these nuances enables accurate audio processing. Let's explore the fundamental distinctions between these two types of sound.

The vocal cords are the intricate mechanisms at the base of the larynx, responsible for human auditory expression. They vibrate to produce the fundamental frequency, which gives each person's voice a distinct pitch and tonal quality. The frequency of these vibrations varies based on individual characteristics. Factors such as age and gender can influence this frequency. As a general observation, the frequency range of human voices mainly lies between 85 and 255 Hz, providing the fundamental resonance and tone associated with spoken or sung words [31].

In contrast, electronic speakers are marvels of modern engineering. The main purpose of speakers is to faithfully reproduce sounds over a wide frequency range, resulting in a comprehensive audio experience for listeners. In today's high-quality electronic speakers, there is considerable focus on amplifying low-frequency sounds, which is known as the "sub-bass over-excitation" phenomenon. This phenomenon is unique to electronic speakers and produces pronounced low-frequency signals, specifically in the sub-bass region (20-80Hz). The signals necessary for creating richer, more intricate sounds in devices vastly contrast with natural human voice frequency patterns. This intentional design aspect heightens the listening experience while also generating a unique acoustic signature that clearly distinguishes electronic sounds from organic vocal sounds [31]. Figure 2.1 illustrates the structures that create a human voice and electronic sound, highlighting the differences between the two mechanisms.



(a) Vocal cords and articulators in speech production.

(b) Mechanism of sound production in speakers.

Figure 2.1: Figures illustrating the differences between human and electronic sound production mechanisms [31].

Electronic speakers' sub-bass variations pose unique challenges not present with human voices. As highlighted by Blue et al. [31], there are two primary issues in capturing these variations. One challenge is the variability of sub-bass over-excitation among different electronic speakers due to their enclosure design. Specifically, the resonant frequency of a speaker enclosure is correlated with its physical dimensions. For instance, if a rectangular enclosure is used, it will produce three resonant frequencies, with each corresponding to a pair of parallel walls. In contrast, if the enclosure is cube-shaped, its three resonance components will be identical, causing sub-bass over-excitation to peak at a single frequency. Nevertheless, when the enclosure's dimensions differ from each other, it will have three distinct resonant frequencies, resulting in a more evenly distributed over-excitation in the sub-bass region. The temporal variability of sub-bass in electronic speakers is complex. Depending on the phoneme's frequency components, a speaker can produce varying amounts of sub-bass for different commands. This is due to how the frequencies stimulate the enclosure's resonance. To clarify, some frequencies cause more intense resonance than others. Additionally, organic speaker recordings may include extra sub-bass from background sources. External disturbances, such as accidental impacts or collisions, may elevate sub-bass levels. These abrupt sub-bass peaks can lead to a command's FFT providing an inaccurate depiction of the existing sub-bass levels within the commands. For a visual illustration of these intricate dynamics, please see Figure 2.2. This portrayal underscores the significant divergence in acoustic patterns between human voices and electronic speakers, underscoring the importance of precision in audio classification systems designed to distinguish between these sound sources.

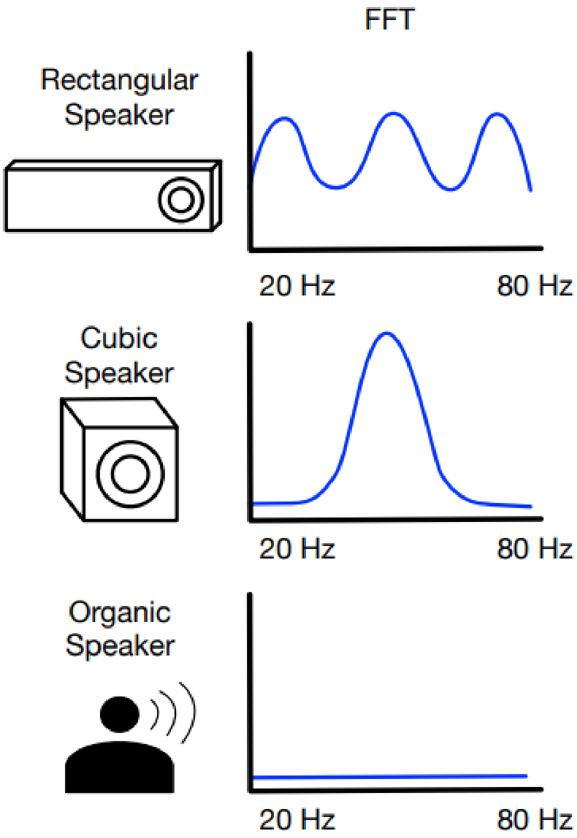


Figure 2.2: The dimensions of an electronic speaker's enclosure determine the different frequencies in the sub-bass region that are over excited [31]

Chapter 3 Related Work

The endeavor to distinguish between human voices and TV noise is a convoluted yet intriguing challenge that has garnered the attention of various researchers across several fields. As we delve into this chapter, we will unearth the comprehensive landscape of previous works that have ventured to tackle this problem, understanding their methodologies and key findings. To begin with, we'll shed light on the wider context wherein TV detection has been contemplated. From there, we'll navigate through studies that have approached the problem without the assistance of machine learning. Here, we will spotlight the seminal work of Blue et al. [31], which crafted an innovative metric grounded in the realm of sub-bass variation. In the latter sections, we will transition to a discussion on research that has harnessed the power of machine learning and deep learning paradigms. Emphasis will be on preprocessing techniques, choice of features, and the utilization of various public datasets, providing readers a holistic understanding of the state-of-the-art.

3.1 TV Detection

In the context of audio processing and classification, ”TV noise” refers to the composite sound output originating from television broadcasts. This encompasses a wide range of audio signals such as dialogues, music, sound effects, and even periods of silence or ambient noise during the transmission. What makes TV noise distinct is its inherently varied and dynamic nature. Unlike consistent sounds, like a steady beep or the hum of an appliance, TV noise can rapidly oscillate between loud explosions in an action scene to the subtle whispers of a dramatic dialogue. Moreover, the sound from televisions is modulated by the electronic speakers designed to reproduce a broad spectrum of frequencies, which often results in unique acoustic signatures such as the sub-bass over-excitation phenomenon [31]. This rich and multifaceted audio landscape makes TV noise a challenging candidate for audio classification tasks, especially when the objective is to distinguish it from other sound sources, like human voices.

The distinction between human voices and TV noise isn't merely an academic challenge but has practical implications across several domains. As we traverse the landscape of prior works, we'll recognize that this intricate problem of TV noise differentiation has been considered in various application areas, each presenting its unique challenges and nuances.

3.1.1 Action Recognition

Audio signals provide a wealth of information about human activities, particularly in domestic environments, for the purpose of Action Recognition. A wide variety of sounds, including conversations and appliance noises, describe the ongoing actions that are occurring. However, it is vital to differentiate TV noise from other sound sources, since televisions can imitate real-world sounds, leading to possible confusion. Precise differentiation between sound sources plays a crucial role in smart home contexts, where instant comprehension of human activities is essential for automation and user experience. While our main objective focuses on distinguishing between TV noise and human voices, Dekkers et al. [32] derived from the SINS database, which comprised continuous recordings of a single individual living in a vacation home over a week. This data was captured using a network of 13 microphone arrays distributed throughout the home. The continuous recordings were segmented into 10-second audio clips. Activities such as "Cooking," "Watching TV," and "Working" were among the predefined classes. The authors extracted features in frames of 40 ms with a 50% overlap, using 40 mel bands covering a frequency range from 50 to 8000 Hz. The authors designed a single classifier model that takes a single channel as input. The classifier was based on a Neural Network architecture comprising two convolutional layers and one dense layer. The input to the network was log mel-band energies. The highest accuracy achieved was for the "Watching TV" activity, with an impressive 99.59%. However, other activities like "Dishwashing" and

”Other” had significantly lower accuracies, with 76.73% and 44.74% respectively. It is worth noting that there is no overlap between activities in SINS dataset, which guarantees unambiguous distinction among recorded events.

3.1.2 Sound Event Recognition

Sound Event Recognition (SER) is the task of automatically identifying and classifying different sounds or events in an auditory environment. The main objective of SER is to recognize various acoustic events, which could range from environmental noises like rain or traffic to human-centric sounds such as laughter or clapping. These systems aim to enhance machine understanding of real-world auditory scenes, providing potential applications in surveillance, environmental monitoring, and smart home systems, among others. A key challenge in SER is dealing with the inherent variability and complexity of real-world soundscapes. This variability comes from the multitude of potential overlapping sound sources, the diverse nature of sounds themselves, and the acoustics of the environment. Distinguishing between sounds like TV noise and other sources becomes especially crucial in such settings, given the rich and multifaceted nature of broadcast content. A significant contribution to SER is by Plakal and Ellis [19], who developed YAMNet. YAMNet is a deep learning model designed for audio event classification, pretrained on the expansive AudioSet-YouTube corpus [5]. The model employs the MobilenetV1 depthwise separable convolution architecture, which is efficient in terms of computational cost. YAMNet is designed to output scores across 521 audio event classes, capturing a wide range of real-world sounds. To process the audio data, the authors employed a series of steps to transform raw audio into a format suitable for their model. All audio was resampled to 16 kHz mono. The authors then computed a spectrogram using magnitudes of the Short-Time Fourier Transform with parameters such as a window size of 25 ms and a window hop of 10 ms. This spectrogram was mapped to 64 mel bins covering the frequency range of 125-7500 Hz to produce a mel spectrogram. To stabilize the mel spectrogram, the authors applied a logarithmic transformation with a small offset to avoid computing the logarithm of zero. The resulting features were framed into overlapping examples of 0.96 seconds, each covering 64 mel bands and 96 frames of 10 ms. These patches were then input into the MobilenetV1 model, producing a set of activations which were subsequently averaged and processed to produce the final 521 per-class output scores for each 960 ms input waveform segment. In terms of performance, the model achieves a mean average precision of 0.306 on the AudioSet evaluation set, demonstrating its ability to classify diverse sound events with reasonable accuracy.

3.1.3 Keyword Spotting

Keyword Spotting (KWS) systems, a key component of voice assistants and smart devices, are designed to accurately detect specific words or phrases in ambient noise. The task becomes particularly challenging in environments filled with multimedia broadcasts, such as live TV broadcasts. For these systems, distinguishing between human-initiated commands and TV noise is critical to ensure responsiveness and reduce false activations. A well-known challenge with KWS is the accidental activation of voice assistants triggered by unplanned external stimuli. A classic example of this problem occurred in 2017, when a Burger King ad intentionally said, ”OK Google, what is the Whopper burger?” to activate nearby Google Assistant devices and relay information about the burger [6]. In another incident, a news anchor accidentally ordered dollhouses on viewers' Amazon Echo devices by uttering a command live on air, highlighting the vulnerability of such systems to unintentional activation [33]. Such incidents underscore the importance of designing KWS systems that are robust against unintentional triggers, especially from broadcasts. The work of Huang et al. [34] and Liang et al. [35] further underscores this need, as they have specifically addressed KWS in environments with potential TV noise interference.

Huang et al. [34] address the significant challenge of KWS performance degradation in the presence of dominant music or TV noise. To address this, they present a novel ”keyword sifter” system that uniquely bridges the preprocessing front-end with the neural network KWS classifier, allowing them to refine noise reduction by leveraging the insights of the back-end machine. The architectural foundation of their system is based on a five-layer CNN, and their input consists of stacked feature vectors. While their system operates primarily on monophonic audio, they also use stereo microphones to extend its capabilities, implementing adaptive noise reduction. In addition, a noise reduction technique based on the short-time Fourier transform proves to be crucial for seamless KWS feature extraction. Empirical evaluations illustrate the superiority of their sifter system, especially in noisy environments with music or TV distractions, and confirm its effectiveness compared to baseline approaches. The essence of their contribution lies in the symbiotic blend of microphone array speech enhancement and deep learning, which together strengthen KWS in adversarial auditory environments.

Lian et al. [35] address the dual challenges of KWS and speaker verification (SV) for voice assistants, emphasizing the importance of reliable voice trigger detection and ensuring authorized activations. Their research culminates in the design of the Multi-task Deep Cross-attention Network , a groundbreaking system that synergistically leverages a KWS subnetwork and a SV subnetwork to serve both functions with increased efficiency. At the heart of MTCANet are the Deep Cross-Attention module, designed to facilitate the mutual benefits of KWS and SV tasks; the Shared Encoder , which extracts common beneficial speech features for both tasks while enhancing noise resilience; and the Soft Attention module, adept at pinpointing critical frequency features. Remarkably, MTCANet sets a new performance benchmark, outperforming established models such as ECAPA-TDNN [36] and Convmixer [37] on metrics such as equal error rate (EER), minimum detection cost function , and accuracy. This unified approach to KWS and SV, supported by its cross-attention mechanism, sets a new precedent in the field of voice assistant technology and achieves remarkable advances even under challenging auditory conditions.

3.1.4 Voice Interfaces Security

Voice interfaces, deeply embedded in Internet of Things (IoT) devices, are both a convenience and a potential security vulnerability. The challenge of ensuring that voice commands are truly coming from users, rather than malicious adversaries, is becoming paramount to maintaining the security and reliability of these systems.

Blue et al. [31] shed light on the security dilemmas surrounding voice interfaces. They vividly illustrate how any sound-emitting entity within the receptive range of a voice device can compromise its operation, potentially triggering unwanted actions such as unsolicited purchases, unauthorized access, or other unforeseen consequences. To address this dilemma, their research aims to distinguish sounds originating from real human speakers from those of electronic counterparts, thereby providing a bulwark against malicious command infiltration. Their research has identified a unique acoustic signature that distinguishes the human vocal apparatus from electronic sound output devices. Called "sub-bass over-excitation," this signature refers to pronounced low-frequency audio elements inherent to electronic loudspeakers but alien to human vocal emissions. The root of this phenomenon lies in the enclosures in which these loudspeakers are housed. Using this distinction, the authors create the "energy balance metric," a measure specifically tailored to capture sub-bass over-excitation. To operationalize this discovery, they develop a detector that categorically bifurcates organic from electronic audio sources based on this over-excitation. In terms of empirical validation, Blue and his team establish the impressive effectiveness of their detector in distinguishing between organic and synthetic sound sources. The accuracy metrics are compelling: in quiet acoustic environments, their detector boasts a true positive rate (TPR) of a flawless 100%, contrasted with a low false positive rate (FPR) of 1.72%. Even in acoustically hostile, noisy areas, the TPR remains robust at 95.7%, while the FPR rises slightly to 5.0%. In addition, the detector demonstrates resilience to the latest audio attack strategies, such as garbled audio injections and codec transcoding attacks. The team's findings prove that sub-bass over-excitation is not a quirk of a particular electronic loudspeaker genre, but is emblematic of its very fabric. This finding cements the validity and universality of their solution for shielding voice interfaces from unauthorized command incursions.

Moving on to Ahmed et al. [38], the authors address the vulnerability posed to voice assistants by potential voice spoofing exploits. Of particular concern are voice replay attacks, in which malicious entities record and replay voice commands, and voice synthesis attacks, in which advanced models are used to mimic victims' voices. Their answer to this dilemma is "Void," a state-of-the-art voice liveness detection mechanism designed to detect voice spoofing schemes. What sets Void apart is its sharp focus on spectral power deviations between real human voices and those regurgitated through speakers. Its distinguishing features include the recognition of spectral power decay and peak patterns, as well as the incorporation of Linear Prediction Cepstrum Coefficients (LPCC) to give it a broader auditory spectrum to work with. Void's evaluation results are commendable. On a dataset teeming with over 255,000 voice samples from a variety of devices and participants, it boasted an EER of just 0.3%. Even on a more eclectic public dataset, the EER was 11.6%. Compared to its deep learning brethren, Void outperforms by being more resource-efficient and faster. Its prowess extends to its versatility, fending off a range of sophisticated voice spoofing stratagems, underscoring its effectiveness in protecting against voice replay attacks.

3.2 Discussion

As we navigated through the vast realm of related work, it becomes imperative to analyze the research holistically. We've synthesized the various methodologies and contributions into a concise table to provide clarity on the different approaches. Then, we will shed light on the gaps that exist in current research and how these interstices motivate the direction of this thesis.

Upon critical reflection on the existing literature, several gaps become apparent. Many studies operate on custom datasets, which introduces challenges when trying to make direct performance comparisons between methodologies. A standardized benchmark dataset would make evaluations more straightforward. Additionally, while some studies show effectiveness in controlled environments, they often neglect to account for the myriad of real-world noises that can impact system efficacy. Moreover, while advanced models demonstrate computational prowess, they frequently require substantial computational resources, rendering them less adaptable for real-time applications. The aim of this master's thesis is not just to build upon the foundation established by earlier works but

Table 3.1: Comparative analysis of various methodologies and their focus areas in the realm of distinguishing human voice from TV noise.

Reference	Focus Area	Method	Notable Findings
Blue et al. [31]	Voice Interfaces Security	Energy balance metric for detecting sub-bass over-excitation	TPR: 100% (quiet environments) & 95.7% (noisy)
Ahmed et al. [38]	Voice Interfaces Security	”Void” (voice liveness detection)	EER: 0.3% on specific dataset
Dekkers et al. [32]	Action Recognition	Neural Network based classifier	”Watching TV” activity accuracy: 99.59%
Plakal and Ellis [19]	Sound Event Recognition	YAMNet (deep learning model)	mAP: 0.306 on AudioSet evaluation set
Huang et al. [34]	Keyword Spotting	”Keyword sifter” system with CNNs	Improved KWS in noisy environments
Liang et al. [35]	Keyword Spotting	MTCANet (Multi-task Deep Cross-attention Network)	EER, minDCF, and accuracy outperforming benchmarks

also to bridge these research voids. The primary objectives include crafting models that are robust to the myriad of real-world noises, as many existing models struggle in such diverse acoustic settings. Furthermore, by utilizing optimization techniques, the goal is to create models that are computationally efficient without sacrificing performance, thus enabling real-time applications.

Chapter 4 Methodology

Building upon the review of related work, this chapter applies the discussed theories to address the challenge of differentiating TV broadcast sounds from daily life audio to prevent inadvertent triggers in voice-controlled systems. The chapter details four methodological approaches developed to tackle the problem of TV sound detection. We have chosen each methodology for its unique potential to contribute to the solution, utilizing various aspects of audio signal processing and machine learning. We will provide detailed information on these methodologies, including their design and the reasoning behind their selection. Furthermore, we will present the datasets - both collected in-lab and from literature - that we will leverage to evaluate the performance of these approaches. These datasets were selected to represent a wide range of real-world situations and assess the durability and versatility of the proposed solutions. We then define the criteria for evaluating the methodologies, including sensitivity in detecting TV noise, discrimination specificity between non-TV sounds, performance in overlapping sound conditions, robustness across different audio hardware, consistency at various distances, and performance in the most difficult scenarios noted. This chapter provides a framework for implementing and testing, preparing the foundation for analyzing and discussing subsequent results.

4.1 Approaches to TV Sound Detection

In order to effectively tackle the challenge of distinguishing TV audio from ambient sounds in everyday environments, this section explores four distinct approaches. These methods offer unique strategies, ranging from leveraging current audio analysis techniques to innovating with advanced machine learning models. Their varied backgrounds bring a wide range of perspectives to the issue, resulting in a thorough comprehension of potential resolutions.

4.1.1 Energy Balance Metric

The Energy Balance Metric, as detailed by Blue et al. [31], offers a novel approach to differentiate audio signals emanating from electronic speakers and human voices. Central to this methodology is the identification of ”sub-bass over-excitation,” a distinct acoustic signature prominent in the output of electronic speakers but not in human vocal emissions. This acoustic phenomenon is particularly evident in the sub-bass region, typically defined as the frequency range between 20 Hz and 80 Hz.

In our study, we adopted this approach to calculate the Energy Balance Metric, which we then used to train a Support Vector Machine (SVM) classifier with a Radial Basis Function (RBF) kernel. The RBF kernel non-linearly maps the input data into a higher-dimensional space, allowing the SVM to classify non-linearly separable data [39].

This approach is rooted in acoustic physics, making it uniquely suited for environments where electronic and human sounds coexist. Its selection is based on the hypothesis that physical characteristics of sound can offer reliable cues for classification, especially in scenarios where digital signal processing techniques alone might fall short.

4.1.2 YAMNet-Based TV Audio Detection

In this approach, we utilize YAMNet [19], a deep convolutional neural network trained on the AudioSet dataset [5], a large-scale collection of 521 human-labeled audio events, to detect TV and radio sound emissions. YAMNet, known for its proficiency in classifying a wide array of audio events, is adapted in our study to focus specifically on TV-related sound classes namely ”Television” and ”Radio”. The process involves extracting predictions scores for these classes from YAMNet’s output and aggregating them to form a unified metric. This aggregation is done by simply summing up the two scores. The rationale behind this approach is to capture the combined presence of both TV and radio elements, which are often intertwined in real-world audio environments. With the aggregated metric, similar to the previous approach, we train an SVM classifier with an RBF kernel to capture a more complex and non-linear relationship within the data.

The choice of YAMNet is driven by its extensive training on a diverse range of audio events, making it a robust foundation for audio classification tasks. This approach leverages YAMNet's pre-trained capabilities to handle the complexity of real-world audio, offering a practical solution that can be quickly deployed without extensive retraining.

4.1.3 Fine-Tuning YAMNet with LSTM Backend

This approach builds upon YAMNet's capabilities by extracting audio embeddings from it and further processing these embeddings with an LSTM (Long Short-Term Memory) model [40]. The LSTM model is designed to recognize patterns in the sequence of embeddings that are indicative of TV sounds, providing a more nuanced understanding than classification based on individual audio frames alone. The method involves extracting YAMNet embeddings for each audio frame, and then feeding these embeddings into the LSTM model. The LSTM model is trained to learn the temporal patterns in the embeddings, enabling it to identify TV sounds based on the sequence of embeddings.

This method combines the strengths of YAMNet's feature extraction with the LSTM's ability to understand temporal dynamics in audio data. It represents an exploration into how deep learning architectures can be synergistically combined to enhance the accuracy and reliability of sound classification, particularly in distinguishing between continuous TV broadcasts and sporadic human conversations.

4.1.4 Ensemble of YAMNet-LSTM Models

In this approach, we develop an ensemble model that integrates various fine-tuned models, each trained on distinct datasets. Employing a Confidence Weighted Ensemble technique, the ensemble strategically combines the strengths of individual models, aiming to create a more robust and accurate system for TV sound detection [41]. This method capitalizes on the confidence level of predictions from each model in the ensemble, ensuring that more reliable predictions are given greater weight in the final decision. The ensemble process involves:

1. Training Individual Models: Each model is trained on a unique dataset, capturing diverse audio characteristics and enabling specialization in various sound scenarios.

2. Aggregating Model Predictions: Predictions are obtained from all the individual models in the ensemble. These predictions are in the form of probability distributions across the two classes (TV sound and non-TV sound).

3. Model Confidence Assessment: The confidence of each model's prediction is determined by the maximum probability value in its output distribution. This value represents the model's certainty in its prediction.

4. Weighting Predictions by Confidence: Each model in the ensemble contributes predictions that are weighted by its confidence level. The confidence level for each classifier's prediction is a critical determinant in this aggregation process. Predictions that exceed a predefined confidence threshold contribute to the final decision, ensuring that the ensemble is guided by the most reliable insights. This approach is critical to improving the accuracy of the ensemble because it selectively amplifies the influence of the most trustworthy models. The weighted prediction for a given model i and sample j is elegantly determined by a function reminiscent of ReLU activation, as described below:

$$p_{\text{weighted}}(i,j) = \begin{cases} p(i,j) & \text{if } p(i,j) > \text{Threshold} \\ 0 & \text{Otherwise} \end{cases} \tag{4.1}$$

   where $p(i,j)$ is the prediction probability vector of model i on the sample j, and the threshold is a hyperparameter that determines the minimum confidence level required for a model's prediction to be considered.

5. Summing Weighted Predictions: The weighted predictions from all models are summed to form the ensemble prediction. This summation effectively combines

the insights from each model, taking into account their respective confidence levels.

$$p_{ensemble} = \sum_{i=1}^{n} p_{weighted}^{(i)} \tag{4.2}$$

where n is the number of models in the ensemble.

6. Normalizing Ensemble Predictions: The summed predictions are normalized by dividing by the number of models to ensure that the ensemble output is a valid probability distribution.

$$p_{ensemble} = \frac{p_{ensemble}}{n} \tag{4.3}$$

7. Determining Final Class Label: The final class label for each sample is determined by selecting the class with the highest probability in the normalized ensemble prediction.

$$label = \arg\max_{c\in\{0,1\}} p_{ensemble}(c) \tag{4.4}$$

where arg max function returns the index of the maximum value, corresponding to the predicted class label.

The Confidence Weighted Ensemble of YAMNet-LSTM models showcases an innovative approach in audio classification [42]. By utilizing a range of models and integrating their predictions based on confidence, this technique aims for high accuracy and reliability in identifying TV sounds. This demonstrates the effectiveness of collaborative model prediction in highly complex classification tasks.

4.2 Datasets

In our study, we utilize four datasets, each unique in its composition and relevance to our research objectives: EAR-Aging, EAR-Divorce, SINS, and a Lab dataset compiled by David Farynyk, a Ph.D. student at UNC Charlotte.

4.2.1 EAR-Aging and EAR-Divorce

These datasets were collected using the Electronically Activated Recorder (EAR), a portable audio recorder designed for ecological behavioral observation. It intermittently records ambient sound bites, capturing the natural auditory environment of the participants' daily activities. This method aligns with ethnographic research principles by maintaining a high degree of naturalism in the raw data [43].

- EAR-Aging Dataset: Gathered over two years from 93 participants aged between 65 and 90. The audio was segmented into 30-second clips, with a final count of 32,984 samples after pruning. Among these, 10,491 samples include TV watching [44].
- EAR-Divorce Dataset: Comprises recordings from 122 individuals over five years, all of whom had experienced divorce. This dataset, post-pruning, contains 49,720 samples, with 14,226 including TV sounds [45].

Both datasets were annotated for various ambient categories and split into training, validation, and testing sets in a 70/15/15 ratio, ensuring no overlap of participants across sets. Table 4.1 summarizes the key characteristics of these datasets.

Table 4.1: Summary of EAR-Aging and EAR-Divorce datasets

Dataset	Total Samples	Samples Including TV	Training/Validation/Testing Split
EAR-Aging	32,984	10,491	70%/15%/15%
EAR-Divorce	49,720	14,226	70%/15%/15%

4.2.2 SINS Dataset

The SINS database [32] consists of continuous audio recordings from a single individual in a vacation home, covering one week. It was collected using a network of 13 microphone arrays distributed over the multiple rooms. We focused on the living room recordings, where "watching TV" and other activities like "phone calls" and "cooking" were identified. After excluding certain activities ("unknown", "don't use" and "other") and corrupted data, the dataset was divided using a 70/15/15 ratio, stratified by activity. Table 4.2 lists the different activities along the mean and standard deviation of the duration of each activity across the 7 selected nodes. As shown in Figure 4.1, the nodes are distributed across the living room, capturing the audio from various locations. Nodes from 1 to 8 (except node 5, which was corrupted) have been selected for our study.

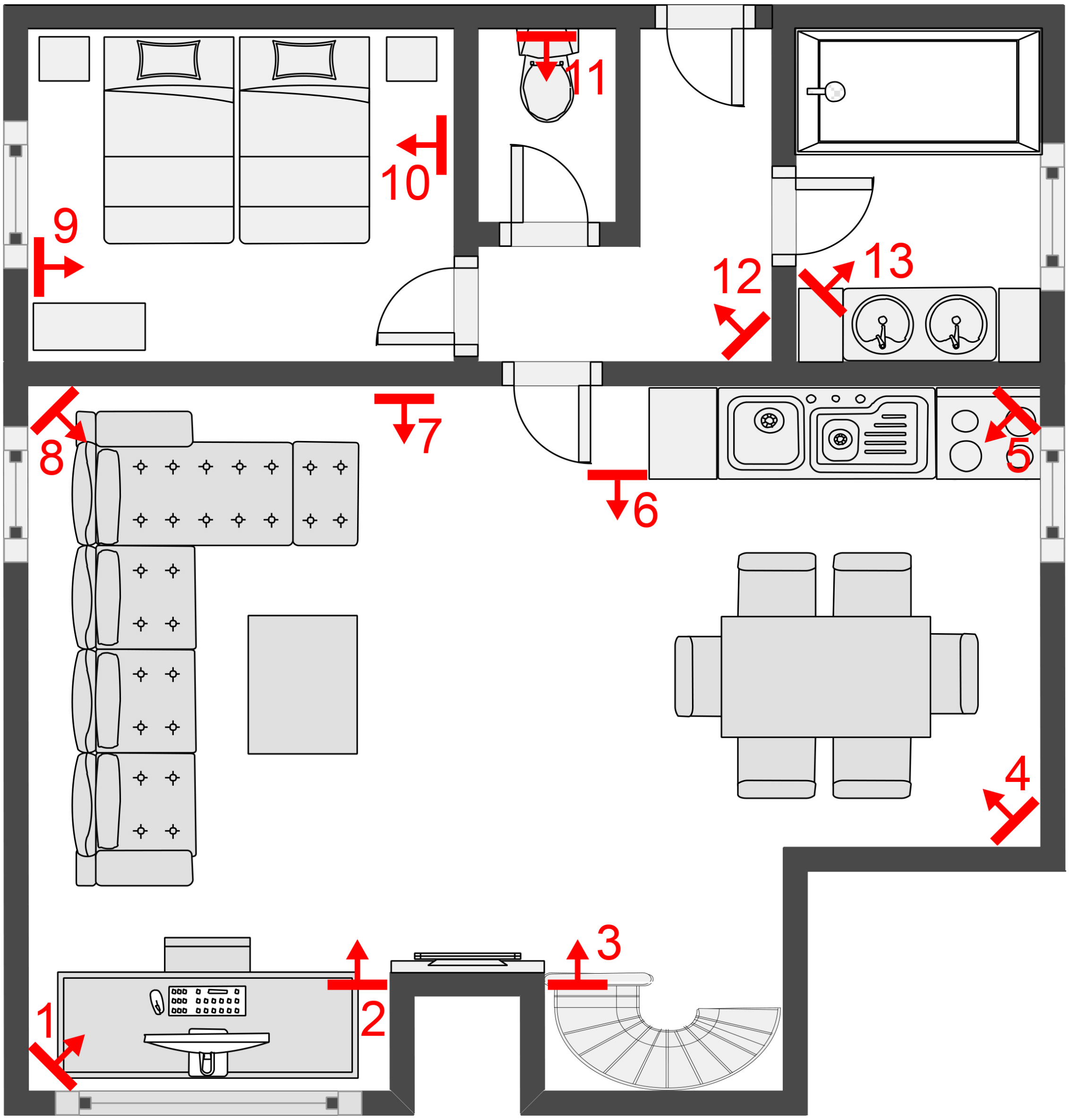


Figure 4.1: 2D floor plan of the SINS dataset [32]

4.2.3 Lab Dataset

The Lab dataset, meticulously assembled by David Farynyk, a Ph.D. student at UNC Charlotte, College of Computing and Informatics, is designed to rigorously evaluate our

Table 4.2: Duration of activities in the SINS dataset (Living Room)

Activity	Mean (min)	Std
Absence	4199.29	53.66
Watching TV	1997.71	25.48
Working	1513.57	21.43
Cooking	312.29	4.33
Calling	178.00	3.46
Eating	146.14	2.80
Other	148.29	7.65
Visit	118.86	1.12
Dishwashing	94.86	2.29
Don't Use	62.86	1.25
Vacuum Cleaning	62.14	2.23
Unknown	3.00	0.00

proposed methodology. It encompasses a range of challenging scenarios, making it a pivotal component of our study. The dataset is categorized into four distinct scenarios:

- TV Sound Only: In this scenario, the focus is solely on capturing TV sounds. The dataset includes recordings using an array of speakers and microphones in different environments, showcasing a variety of content at multiple volume levels.

- TV with Background Conversation: This scenario simulates a common real-world situation where TV audio coexists with background conversation. This setup challenges the model's ability to discern TV audio amidst conversational noise, a key aspect of the study.

- Side Conversation (SC): Here, the focus is on capturing conversations that occur in the same room but not directly in front of the TV. This scenario tests the model's capacity to distinguish between TV sounds and incidental conversational noises in the background.

- Conversation Only: This scenario isolates conversation sounds without any TV audio. It serves as a control scenario, allowing us to assess the model's performance in exclusively conversational environments.

Each of these scenarios was constructed with careful consideration to ensure a balanced and comprehensive dataset:

- Speakers and Microphones: The dataset incorporated a diverse set of audio capture and playback devices to simulate a range of real-world scenarios. This included four different types of speakers: a Small PC, a Small TV, a Midsize PC, and an Amplifier. For recording, four types of microphones were used: TicWatch 3 Pro, Apple Watch, iPhone, and Pixel 4a.

- Environments: Two distinct environments were selected for recording: a Basement and a Carpeted Room. This variance in recording environments is crucial to test the model's adaptability to different acoustic settings.

- Contents: Nine content types were included: Action, News, Baseball, Football, Tennis, Soap Opera, Gameshow, Children's Show, and Commercials. This diversity aimed to challenge the model's ability to accurately identify TV noise across varied audio contexts.

- Volumes and Distances: The recordings were made at three different volume levels (60db, 65db, and 75db) to assess the model's performance in varying loudness conditions. Additionally, for the side conversation data, recordings were done at four distances (5ft, 10ft, 15ft, and 20ft) from the microphone.

Table 4.3: Summary of datasets by categories

	Environment	Content	Microphones	Speakers	Vol./Dist.	Total(min)
SC	Carpet	-	4	-	10ft, 15ft, 20ft, 5ft	260
TV	Basement, Carpet	9	4	4	60db, 65db, 75db	8400
TV & Conv.	-	9	4	1	75db	180
Conv.	Carpet	-	4	-	-	140

4.3 Evaluation Criteria

This section outlines the key evaluation criteria used to assess the performance of the TV sound detection models. Each criterion is designed to test a specific aspect of the model's functionality, ensuring its effectiveness and robustness in real-world scenarios. The Lab dataset, collected specifically for this study, plays a crucial role in evaluating these criteria due to its diverse scenarios and controlled variables.

- Sensitivity in Detecting TV: Sensitivity measures the model's ability to correctly identify TV sound segments. It is quantified by the TPR, reflecting the model's efficiency in detecting actual TV noise occurrences. The Lab dataset, with its distinct TV sound scenarios, provides a rich testing ground to assess the model's sensitivity across various contents and environments.

$$TPR = \frac{TP}{TP + FN} \tag{4.5}$$

- Specificity in Identifying Non-TV Sounds: Specificity evaluates the model's accuracy in identifying segments without TV sounds. It is indicated by the True Negative Rate (TNR). The inclusion of non-TV scenarios like side conversations in the Lab dataset allows for testing the model's specificity under different acoustic conditions.

$$TNR = \frac{TN}{TN + FP} \tag{4.6}$$

- Performance with Overlapping Sounds: This criterion assesses the model's capability to detect TV sounds amidst other overlapping audio sources. The Lab dataset's TV and conversation overlap scenarios are pivotal for evaluating the model's performance in complex acoustic environments.

- Performance Across Different Microphones or Speakers: This measures the model's consistency across various audio input and output devices. Given the Lab dataset's use of different microphones and speakers, it serves as an ideal platform to test the model's adaptability to diverse hardware characteristics.

- Performance at Different Distances: This criterion gauges the model's detection accuracy at various distances between the microphone and the TV. The Lab dataset, with recordings made at different distances, helps in evaluating the model's reliability under varying proximity conditions.

- Performance in Challenging Scenarios: This evaluates the model's effectiveness in its most challenging scenarios, like side conversations. The Lab dataset includes specific scenarios such as side conversations that test the model's limits and highlight areas for improvement.

Summary

This chapter has laid the groundwork for addressing the challenge of distinguishing TV broadcast audio from ambient audio. We have introduced four different methods, ranging from traditional signal processing to advanced machine learning techniques, each of which offers a unique approach to TV sound detection.

The Energy Balance Metric exploits acoustic properties, YAMNet-based methods use a pre-trained neural network, and the fine-tuning of YAMNet with LSTM demonstrates the synergy of feature extraction and pattern recognition. An ensemble model combining these methods aims at robustness and accuracy.

We have also described various datasets, each of which provides unique challenges and insights for evaluation. The EAR-Aging and EAR-Divorce datasets capture naturalistic environments, SINS adds detailed activity annotations, and the specially designed Lab dataset provides controlled scenarios for thorough testing.

The evaluation criteria presented in this chapter focus on assessing the accuracy, efficiency, and adaptability of the models in different acoustic conditions and with different audio devices. The methods and datasets discussed here provide a strong foundation for the upcoming implementation and analysis phases.

Chapter 5 Implementation

Following the theoretical foundation laid in the previous chapter, this chapter moves into the practical implementation of our proposed methods for TV sound detection. Here, we detail the process of implementing these concepts, covering environment setup, data preparation, model development, and evaluation.

We begin by outlining the computational environment, including the hardware and software configurations used in the study. This is followed by a description of the data preparation steps, highlighting how the datasets were processed for analysis. The core of the chapter focuses on the implementation of each model, from the energy balance metric to the ensemble of YAMNet-LSTM models. We discuss the training strategies, hyperparameter tuning, and methods used to improve model performance and avoid overfitting.

In addition, we discuss the evaluation setup in accordance with the established criteria and provide initial results and insights into the performance of each model. Challenges encountered during implementation and their solutions are also shared, providing a practical perspective on the research process.

In essence, this chapter serves as a bridge between theory and practice, but also as a foundation for the following chapters, where we discuss the results and draw conclusions.

5.1 Environment Setup

The implementation of our study involved a combination of software tools, computational resources, and hardware setups, ensuring a robust and efficient research process. This section details the key components of our environment setup, encompassing software platforms, data storage, computational resources, and hardware for data collection.

5.1.1 Software and Computational Resources

Our study predominantly utilized Python, a versatile programming language with extensive support for data science and machine learning. Python's extensive libraries and frameworks facilitated the development and execution of our machine learning models. Key libraries used include:

- TensorFlow: A powerful open-source library for machine learning and neural network modeling.

- Librosa: A Python library for audio analysis, essential for processing and extracting features from audio data.

- Soundfile: Used for reading and writing audio files, integral to handling our datasets.

- Git: Employed for version control, ensuring efficient management and tracking of code changes throughout the study.

Computational tasks, especially training and evaluating machine learning models, were executed on the University's high-performance computing cluster. The cluster's GPU partition, essential for handling large-scale computations efficiently, included the following specifications:

- Multiple GPU nodes, including Titan V, Titan RTX, V100S, A100, and A40 models, equipped with NVIDIA GPUs.

- High-capacity RAM configurations and advanced Intel Xeon and AMD EPYC CPUs across nodes.

- Fast Infiniband interconnects, facilitating efficient data transfer within the cluster.

These resources were instrumental in accommodating the computational demands of our deep learning models, allowing for efficient training and evaluation processes.

5.1.2 Data Storage and Management

Given the size and sensitivity of the datasets, particularly the EAR datasets containing personal audio recordings, secure and efficient data storage was paramount. The datasets were stored on the University cluster, ensuring both data security and seamless integration with our computational resources. This setup facilitated easy access and manipulation of the data during various stages of the research.

5.1.3 Hardware for Data Collection

For the Lab Dataset, specific hardware was employed to capture a diverse range of audio scenarios. The setup included:

- Microphones: A range of devices such as the TicWatch Pro 3, iPhone, Pixel 4a, and Apple Watch, each offering different recording characteristics.
- Speakers: Various types, including Small PC, Medium PC, Small TV, and an Amplifier, were used to generate a wide spectrum of audio outputs.

5.2 Data Preparation

Data preparation is a critical phase in any machine learning project, and our study was no exception. Each approach we considered required specific data handling techniques, though there were general steps that applied universally. Here, we outline both the common and unique data preparation procedures employed for our datasets.

5.2.1 General Data Preparation

The general process of data preparation included the following steps:

- Data Splitting: All datasets were divided into training, validation, and test sets using a 70/15/15 split ratio. This division ensured a balanced approach, where the training set was used for model development, the validation set for hyperparameter tuning, and the test set for final evaluation.

- Shuffling: Before splitting, the datasets were shuffled to eliminate any inherent ordering that might bias the models.

- Audio Downsampling and Mono Conversion: To streamline computational requirements and maintain consistency across datasets, all audio recordings were downsampled to a 16kHz sampling rate and converted to mono-channel audio.

5.2.2 EAR Datasets

For the EAR-Aging and EAR-Divorce datasets, additional preparation steps were taken:

- Data Pruning: Recordings labeled with ”problem” were removed, as well as those with disagreements between annotators (in the EAR-Divorce dataset) and corrupted files.

- Participant-wise Splitting: To prevent data leakage and avoid the models recognizing specific participants’ voices, we ensured that each participant’s recordings were confined to a single data set (either training, validation, or test).

5.2.3 SINS Dataset

The SINS dataset required a different approach, given its unique structure:

- Node-specific Modeling: As data from each node in the living room represented different spatial audio characteristics, we developed separate models for each node. This approach allowed for capturing the variability in sound across different locations within the room.

- Activity-based Stratification: During splitting, care was taken to stratify the data based on activity labels. This stratification ensured no single activity was overrepresented in any of the sets, thereby reducing the risk of overfitting on specific activities.

- Exclusion of Certain Activities: Activities labeled as ”unknown”, ”don’t use”, and ”other” were removed from the dataset to maintain clarity and focus on relevant audio samples.

5.3 Model Implementation

In this section, we delve into the implementation details of the methodologies we have chosen for TV sound detection. Each approach is described in depth, highlighting the steps involved in the model development and implementation process.

5.3.1 Energy Balance Metric Implementation

The implementation of the Energy Balance Metric (EBM) model, inspired by the work of Blue et al. [31], follows a series of meticulous steps:

1. Sliding Window Segmentation: The audio signals are segmented using a sliding window technique. Each window, with a duration of 0.1 seconds, allows for a time-localized spectral analysis of the audio signal.

2. Fourier Transform and Frequency Cropping: We apply the Fast Fourier Transform (FFT) to each windowed segment. The FFT size is set to 4096, balancing resolution with computational demands. The frequency spectrum obtained from FFT is then cropped to focus on the lower frequencies, specifically between 20 Hz and 250 Hz.

3. Energy Distribution Analysis: The energy in the cropped frequency spectrum is computed. This spectrum is normalized by the total spectral energy to derive the relative energy distribution.

4. Sub-Bass Energy Calculation: We calculate the sub-bass energy by summing the normalized energy within the frequency range up to 80 Hz, emphasizing the lower frequency components.

5. Energy Balance Metric Computation: The EBM for each segment is determined by the ratio of sub-bass energy to the total energy within the evaluated frequency

range. This metric quantifies the balance between sub-bass and higher frequencies.

$$EBM = \frac{E_{Sub-bassRegion}}{E_{TotalRegion}} \tag{5.1}$$

Here, $E_{Sub-bassRegion}$ is the energy in the sub-bass region (20-80 Hz), and $E_{TotalRegion}$ is the total energy in the evaluated region (20-250 Hz).

6. Outlier Removal and Final Calculation: To enhance the robustness of our EBM calculation, we perform an outlier removal process based on the skewness of our data. The final EBM value is then calculated as the median of the remaining EBMs, offering a more stable representation of the energy balance in the audio signal.

Continuing from the computation of the EBM, we proceed to utilize this metric for classification purposes:

7. Classifier Training: In our experimental setup, we initially attempted to determine an optimal threshold for classification by analyzing the Receiver Operating Characteristic (ROC) curve. This approach aimed for simplicity but presented challenges in achieving a satisfactory balance between the TPR and FPR. To explore alternative methods, we conducted further experiments with a SVM using a linear kernel. However, this model did not yield performance metrics as robust as we desired. After extensive evaluation on the validation dataset, we determined that an SVM equipped with a RBF kernel, configured with an auto gamma value, was significantly more effective. This configuration proved crucial in distinguishing between TV broadcast sounds and human speech, leveraging the calculated EBM value for each sample as input features. The adoption of the RBF kernel allowed for a more nuanced and complex interpretation of the EBM data.

The implementation of the EBM model underscores our commitment to using detailed and scientifically sound methods in audio analysis. By focusing on the energy distribution within specific frequency ranges, this approach provides a nuanced view of the acoustic properties of the audio signals, which is critical for distinguishing between TV sound and other ambient noises.

5.3.2 YAMNet-TV Implementation

The YAMNet-TV model implementation effectively leverages the YAMNet architecture, a robust and proven framework in the field of audio event classification. Our implementation focuses on harnessing YAMNet's capabilities to specifically identify and classify TV and radio sound events. Here is a detailed breakdown of the process:

1. Preprocessing: Each audio recording in our dataset is segmented into 30-second chunks. This segmentation approach allows for a comprehensive analysis of each recording, ensuring that the model captures a representative sample of the audio's characteristics. We also down-sample and convert the audio to mono, and normalize the signals to the range of [-1.0, +1.0] for consistency across the dataset.

2. YAMNet Inference on Segments: For each 30-second audio chunk, we utilize the YAMNet model to perform inference. YAMNet operates by analyzing shorter segments of audio, specifically every 0.96 seconds within the larger 30-second chunk. This granular approach enables YAMNet to capture the dynamic nature of audio events throughout the duration of the recording.

3. Extracting Prediction Scores: The YAMNet model outputs a 521-dimensional vector of prediction scores for every 0.96-second segment within the 30-second chunk. Each dimension in this vector corresponds to a specific sound event, including those for ”Television” and ”Radio.”

4. Aggregating Scores: To derive a comprehensive prediction for the entire 30-second audio chunk, we average the prediction scores across all 0.96-second segments within the chunk. This averaging process consolidates the model's predictions into a single, unified score for each sound event across the entire duration of the audio.

5. Combining Television and Radio Scores: From the averaged predictions, we extract

and combine the scores for "Television" and "Radio" sound events. This combined score represents the model's overall assessment of the presence of TV and radio sounds within the 30-second audio chunk.

6. SVM Classification: The combined TV sound prediction score for each 30-second audio chunk is then used as the input feature for the SVM classifier. Similar to the EBM model, we experimented with different SVM kernels and configurations. After extensive evaluation, we determined that an SVM with an RBF kernel and auto gamma value was the most effective in classifying TV sounds.

Through this approach, the YAMNet-TV model capitalizes on the advanced audio classification capabilities of YAMNet while fine-tuning its focus to the specific task of TV and radio sound detection. This implementation showcases the adaptability of pre-trained models in addressing specialized tasks within the domain of audio processing and classification.

5.3.3 YAMNet-LSTM Implementation

The YAMNet-LSTM approach in our study involves fine-tuning the YAMNet model to extract deep audio embeddings, followed by employing an LSTM backend for final classification. This process effectively utilizes YAMNet's ability to capture rich audio features and leverages LSTM's prowess in handling time-series data. The implementation can be broken down into the following key steps:

1. Preprocessing: Audio recordings are preprocessed similarly to the YAMNet-TV implementation, which involves segmentation, downsampling, and normalization.

2. YAMNet Embedding Extraction: Instead of using the class prediction scores, here we extract the deep audio embeddings output by YAMNet. These embeddings capture intricate patterns and characteristics within the audio data.

3. LSTM Model Architecture: The LSTM model comprises two layers of LSTM cells, each with 64 units. The model architecture is designed to process sequences of

embeddings extracted from YAMNet. We use the "glorot uniform" initializer (also known as Xavier initializer) for the LSTM layers and "he normal" for the final dense layer. The LSTM layers are configured to return sequences, allowing the model to learn dependencies across the entire sequence of embeddings.

4. Output Layer and Compilation: The output of the LSTM layers feeds into a dense layer with softmax activation, designed for binary classification (TV sound vs. non-TV sound). The model is compiled using the Adam optimizer with a learning rate of 3e-4 and categorical crossentropy as the loss function. Accuracy is used as the metric for performance evaluation.

5. Early Stopping and Training: To prevent overfitting and optimize generalization, we employ an early stopping callback with a patience of 10 epochs. The callback monitors the validation loss, restoring the best weights upon early stopping. The model is trained for up to 100 epochs, with a batch size of 32. After each epoch, the training data is shuffled to ensure that the model does not learn any ordering in the data.

6. Data Preparation and Model Training: Audio data is first converted into YAMNet embeddings and then fed into the LSTM model for training. The model learns to recognize temporal patterns in these embeddings that are indicative of TV sounds. The training process involves both the training and validation datasets, allowing the model to learn and validate its predictions iteratively.

This approach effectively marries the strengths of YAMNet's advanced feature extraction with LSTM's sequence learning capabilities. By fine-tuning on embeddings and using a sophisticated LSTM architecture, this model aims to achieve high accuracy in identifying TV sounds, especially in complex and temporally varied audio environments. The use of early stopping and training data shuffling further enhances the model's robustness and reliability.

5.3.4 YAMNet-LSTM Ensemble Implementation

For the ensemble model, we integrated multiple YAMNet-LSTM models, each trained on different subsets of the EAR-Aging, EAR-Divorce, and SINS datasets. The objective was to create a robust ensemble that leverages the unique strengths of each individual model. The ensemble was formed using the Confidence Weighted Ensemble technique. In our experiment, we have set the threshold hyperparameter to 0.5, which is the minimum confidence required for a model to contribute to the ensemble's prediction. Here's an overview of the implementation process:

Data Preparation and Model Training

1. Dataset Splitting for EAR-Aging and EAR-Divorce: The EAR-Aging and EAR-Divorce datasets were split into balanced folds to ensure a diverse range of training data for each model. We split each dataset into 4 folds, each containing a balanced proportion of TV and non-TV samples. We also separated the data by participant to prevent data leakage and ensure that each participant's recordings were confined to a single fold. This approach allowed us to train 4 different models on each dataset, each with a unique subset of data.

2. SINS Dataset Preparation: Since the SINS dataset was already divided by nodes, representing different spatial audio recordings, we treated each node as a separate dataset. This approach allowed us to construct individual models for each node, capturing the unique acoustic properties of different locations in the living room.

3. Model Training: Separate YAMNet-LSTM models were trained on each fold of the EAR datasets and each node of the SINS dataset. This resulted in a total of 4 models for EAR-Aging, 4 for EAR-Divorce, and 7 for SINS, each fine-tuned to its specific subset of data.

Ensemble Creation

We built five different ensemble models using the Confidence Weighted Ensemble method. These ensembles were:

- Ensemble of the 4 EAR-Divorce models.
- Ensemble of the 4 EAR-Aging models.
- Ensemble combining EAR-Aging and EAR-Divorce models.
- Ensemble of the 7 SINS models.
- Ensemble combining all EAR models and SINS models, in total 15 models.

This Confidence Weighted Ensemble method is designed to optimize the ensemble's performance by intelligently combining the strengths of individual models. By weighing predictions based on confidence, the ensemble can make more accurate and reliable decisions, particularly in complex and varied audio environments.

5.4 Evaluation Setup

The evaluation phase of our study was meticulously designed to assess the performance of the proposed methodologies under various real-world conditions. To achieve this, we utilized a multi-faceted evaluation setup that incorporated diverse datasets and focused on key performance metrics.

5.4.1 Testing Data

A critical aspect of our evaluation was the use of unseen testing data to ensure the validity and reliability of our results. This testing data was reserved from each of the EAR-Aging, EAR-Divorce, and SINS datasets, ensuring that the models were assessed on audio samples they had not encountered during the training phase. This approach helped in evaluating the models' generalization capabilities.

5.4.2 Lab Dataset as the Primary Evaluation Source

The Lab dataset, collected by David Farynyk, played a pivotal role in our evaluation process. Given its comprehensive collection parameters, it was ideal for a thorough assessment of the models. The Lab dataset encompasses multiple scenarios:

- TV Sound Only and Non-TV scenarios provided a clear differentiation between TV and non-TV sounds, allowing us to measure the models' sensitivity and specificity accurately.
- Overlapping Sounds (TV and Conversation) scenarios presented challenging conditions where TV sounds coexist with or are masked by other sound sources, testing the models' capabilities in complex audio environments.
- Side Conversation scenarios allowed us to evaluate the models' ability to distinguish between TV sounds and side conversations, which are often mistaken for TV sounds.
- The variety of speakers and microphones used in the dataset enabled us to assess the models' robustness across different audio capturing and playback devices.
- The inclusion of different volumes and distances in the dataset allowed us to evaluate the models' performance under various acoustic conditions.

5.4.3 Performance Metrics

To effectively evaluate the methodologies, we focused on two primary metrics: sensitivity and specificity. These metrics are particularly relevant for our study as they provide insights into the models' abilities to correctly identify TV sounds (sensitivity) and accurately discern non-TV sounds (specificity).

- Sensitivity: This metric measures the proportion of actual TV sound instances that the model correctly identifies. A high sensitivity indicates the model's effectiveness in detecting TV sounds.

- Specificity: This metric assesses the model's ability to correctly identify instances where TV sound is not present. High specificity implies the model's accuracy in avoiding false positives in non-TV sound conditions.

This comprehensive evaluation setup was essential for a thorough and nuanced understanding of the strengths and limitations of each proposed methodology. It provided us with valuable insights into how these methodologies might perform in real-world applications, where the variability of audio environments and scenarios is vast.

Summary

This chapter provides a detailed roadmap for implementing our methods for TV sound detection. It outlines the complexities of setting up a robust computational environment using advanced resources and diverse hardware for data collection. The data preparation phase has been tailored to the specific needs of each methodology to ensure accurate representation of real-world scenarios. The implementation of each model, from the Energy Balance Metric to the ensemble of YAMNet-LSTM models, is highlighted, emphasizing the technical intricacies and strategic choices that underpin their functionality. The chapter also highlights the training strategies, hyperparameter tuning, and methods used to improve model performance and avoid overfitting.

Moving on to the evaluation setup, a comprehensive framework was established to test the effectiveness of our methodologies under diverse and challenging conditions. The use of unseen test data and the strategic use of the lab dataset underscored our commitment to ensuring the validity and reliability of our findings. The chapter combines theoretical insights with the resulting results and discussions, demonstrating the rigorous and methodical approach taken in this research.

Chapter 6 Evaluation

Building on the implementation phase, we systematically evaluate the performance of each model using a variety of datasets and scenarios. Our evaluation strategy is designed not only to validate the effectiveness of the models in controlled test environments, but also to test their robustness and adaptability in real-world situations, as represented by the diverse scenarios in the Lab dataset.

The models under evaluation include the Energy Balance Model, YAMNet-TV, YAMNet LSTM, and the YAMNet Ensemble. Each of these models, tailored to address the challenges of distinguishing TV sound from ambient sound, will be tested using specific datasets and criteria. The EAR-Aging, EAR-Divorce, and SINS datasets provide the initial testing ground, where the models' capabilities are measured against unseen data. Next, the Lab dataset, with its intricate and varied audio scenarios, provides a comprehensive platform for understanding the models' performance in more complex and realistic settings.

The primary metrics used for evaluation are sensitivity and specificity, chosen for their relevance in accurately quantifying the models' ability to detect TV sounds and distinguish them from non-TV audio.

The purpose of this chapter is to provide a detailed and transparent account of our evaluation process, and to provide insight into the strengths and limitations of each model.

6.1 Evaluation Methodology

The methodology for evaluating our models is structured to provide a comprehensive understanding of their performance under various conditions. This section outlines the steps and considerations that formed the basis of our evaluation strategy, ensuring that the models are tested in a manner that reflects both their theoretical capabilities and practical applicability.

6.1.1 Model Selection for Evaluation

For the evaluation, the following models were selected based on their distinct approaches to TV sound detection:

- EBM Model: This model uses the energy balance in audio signals to differentiate between TV sound and non-TV sound.
- YAMNet-TV Model: Leveraging the YAMNet architecture, this model focuses on classifying audio as TV sound based on the presence of television and radio classes in the audio.
- YAMNet-LSTM Model: This approach combines YAMNet's feature extraction with LSTM's sequential pattern recognition, offering a nuanced perspective on audio classification.
- YAMNet-LSTM Ensemble Models: These models are ensembles of multiple YAMNet-LSTM models, trained on different datasets or nodes, providing a collective decision based on various individual predictions.

6.1.2 Testing on EAR-Aging, EAR-Divorce, and SINS Datasets

The initial phase of our evaluation involved testing the models on the reserved test sets of the EAR-Aging, EAR-Divorce, and SINS datasets. This phase aimed to assess the models' performance in controlled conditions, with each dataset offering unique challenges:

- EAR-Aging and EAR-Divorce: These datasets provided real-life scenarios with diverse ambient sounds, including TV audio. Evaluating the models on these datasets allowed us to test their ability to identify TV sounds amidst everyday noises and conversations.

- SINS Dataset: With its focus on spatial audio recordings in a living environment, the SINS dataset tested the models' capacity to handle different acoustic conditions and activities, including TV watching.

6.1.3 Comprehensive Testing on the Lab Dataset

The Lab dataset, designed to simulate a wide range of real-world scenarios, served as the primary source for an exhaustive evaluation:

1. Scenario-Based Testing: The models were tested across various scenarios, including TV sound only, TV with background conversation, side conversation, and conversation only. This allowed us to evaluate the models' performance in distinguishing TV sounds in the presence and absence of overlapping audio.

2. Hardware Variability: By utilizing different combinations of speakers and microphones, we assessed the models' robustness and consistency across varying hardware setups.

3. Acoustic Conditions: The impact of different volumes and distances on the models' performance was evaluated, reflecting their adaptability to different sound levels and spatial configurations.

6.1.4 Metrics for Evaluation

Sensitivity and Specificity remained the primary metrics for this evaluation phase. These metrics provided a balanced perspective on each model's accuracy in correctly identifying TV sounds (Sensitivity) and avoiding false identifications in non-TV scenarios (Specificity).

6.2 Evaluation Results

6.2.1 EBM Model

The EBM model was evaluated on the EAR-Aging, EAR-Divorce, SINS, and Lab datasets. The results of this evaluation are presented below, followed by a detailed analysis of the model's performance.

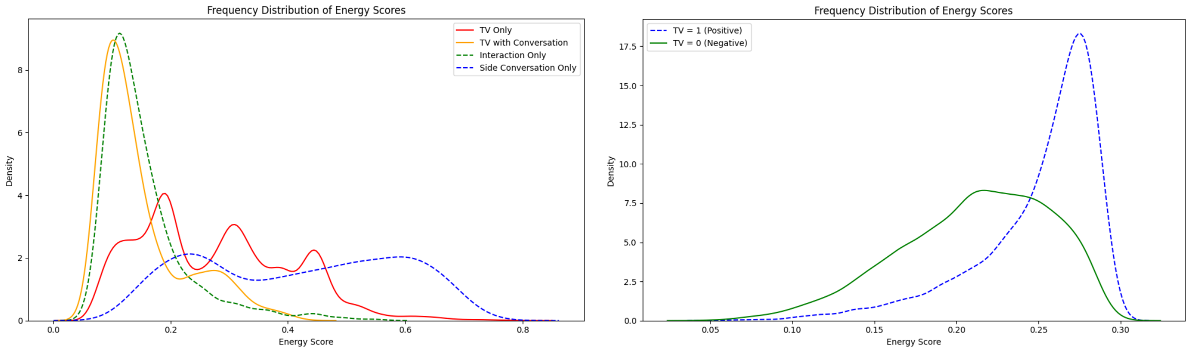
Analysis of the Energy Metric Distributions

The distribution graphs of energy scores reveal key insights into the performance of the EBM approach under various conditions:

- Scenario Analysis: Figure 6.1 compares the distribution of TV and non-TV classes in EAR-Aging dataset, indicating a clear distinction between the energy scores. The peak for the TV class is higher, suggesting a higher energy balance metric when the TV is present. We also support this analysis using the Lab dataset with different scenarios. The distribution presents overlapping between scenarios. This overlap, particularly between "TV with Conversation" and "Interaction Only," suggests that the EBM may struggle to distinguish between TV sounds and conversations in the presence of overlapping audio sources.

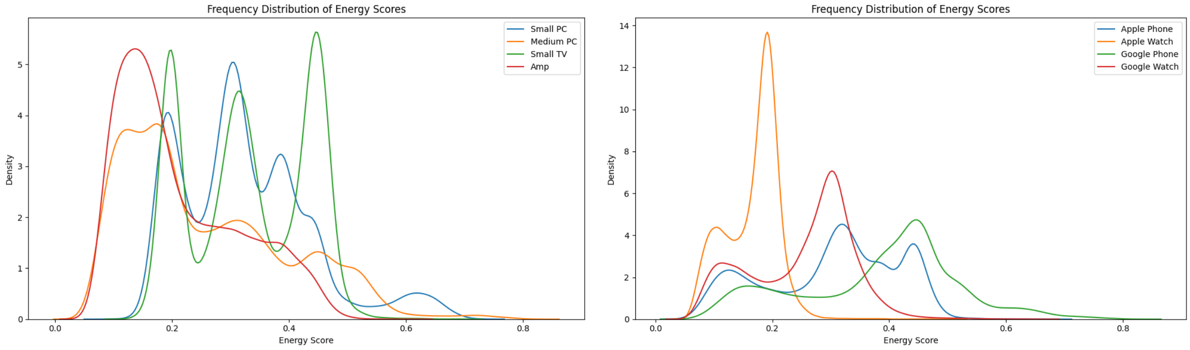
- Hardware Variability: Figure 6.2 shows the frequency distribution of energy scores across different recording devices and specially the speakers. The distinct peaks for each device indicate variability in energy score distribution, suggesting that the EBM is sensitive to the hardware used for audio capture.

- Volume Analysis: Figure 6.3, the distribution of energy scores at different volume levels shows that the volume can significantly impact the energy scores. Higher volumes seem to yield higher energy scores, which could potentially aid in distinguishing TV sounds at varying loudness levels.



(a) Across different scenarios (Lab dataset) (b) TV vs. Non-TV classes (Aging dataset))

Figure 6.1: Composite figure showing the frequency distribution of energy scores across different scenarios and for TV vs. Non-TV classes.



(a) Across different speakers (b) Across different microphones

Figure 6.2: Composite figure showing the frequency distribution of energy scores across different speakers and microphones (Lab dataset).

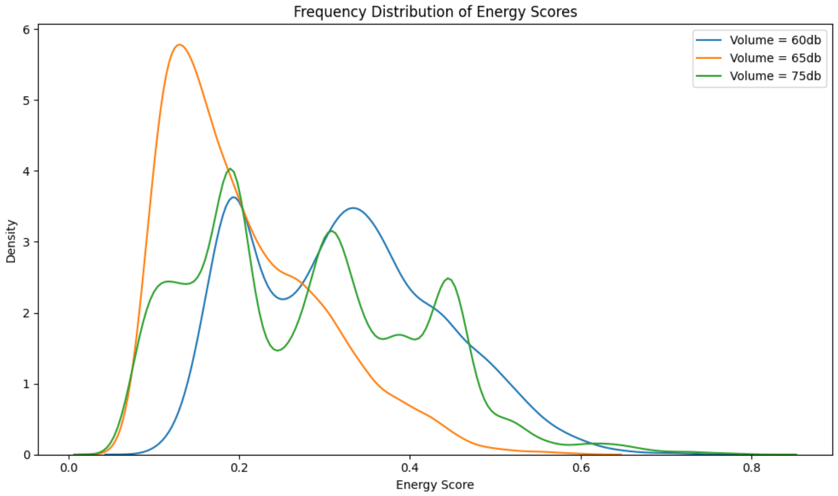


Figure 6.3: Frequency distribution of energy scores at different volume levels (Lab dataset).

Results Evaluation

EAR-Aging Dataset On the EAR-Aging dataset, the EBM exhibited moderate recall and specificity, with the following results:

- Recall: 65.88%
- Specificity: 71.68%
- Accuracy: 69.91%

These outcomes suggest that the metric can identify TV sound with a reasonable degree of accuracy, though there is room for improvement, particularly in minimizing false positives.

EAR-Divorce Dataset For the EAR-Divorce dataset, the EBM showed excellent recall but lower specificity:

- Recall: 95.00%
- Specificity: 54.11%
- Accuracy: 61.90%

This indicates a high sensitivity to TV sounds but at the cost of a greater number of false positives, which is undesirable in a practical setting.

SINS Dataset The performance on the SINS dataset varied, with an average recall higher than specificity:

- Average Recall: 80.30% (±11.18%)
- Average Specificity: 40.27% (±13.68%)
- Average Accuracy: 60.28% (±02.14%)

The variability across nodes is indicative of the metric's sensitivity to the audio capture environment.

Lab Dataset The Lab dataset results underscored the challenges in detecting TV sounds amidst overlapping audio:

- Recall: 91.65%
- Specificity: 32.43%
- Accuracy: 63.30%

These findings highlight the difficulty in distinguishing TV audio within complex acoustic environments, especially when background conversations are present.

Detailed Analysis on the Lab Dataset

A comprehensive analysis of the EBM on the Lab dataset reveals its ability to distinguish between TV sound and conversation:

Scenarios Analysis The distribution of energy scores across different scenarios in the Lab dataset showed that the metric could not effectively differentiate between "TV with Conversation" and "Conversation Only" scenarios. This limitation was also observed in the EAR-Aging and EAR-Divorce datasets, where a significant overlap between TV sounds and discussions was noted, leading to the EBM's poorer performance.

Isolated Speaker Analysis When isolating the analysis to one speaker and comparing "TV Only" with "Conversation Only" scenarios, the results improved significantly:

- Recall: 80%
- Specificity: 69%
- Accuracy: 74%

Particularly notable is the performance for the "Small TV" speaker, with the following improved metrics:

- Recall: 97.68%
- Specificity: 80.04%
- Accuracy: 88.86%

The results across different speakers indicate that the EBM can perform well in scenarios with less acoustic complexity.

Volume Analysis The EBM's performance also varied across different volume levels:

- At 60db: High recall and specificity, leading to an accuracy of approximately 83.29%.
- At 65db: Moderate recall and specificity, with an accuracy of 64.03%.
- At 75db: Substantial recall and specificity, resulting in an accuracy of 76.33%.

These results suggest that volume levels can influence the energy scores, potentially aiding in sound discrimination in more uniform environments.

Microphone Analysis The analysis based on the microphone dimension yielded varied results:

- Apple Phone: High specificity of approximately 89.79% and a recall of 72.85%, achieving an accuracy of 81.32%.

- Apple Watch: Moderate specificity and recall, with an overall accuracy of 70.41%.
- Google Phone: Very high specificity at approximately 93.03% and a recall of 78.89%, leading to an accuracy of 85.96%.
- Google Watch: Good specificity and recall, resulting in an accuracy of 78.07%.

These results demonstrate that the choice of microphone can substantially affect the EBM's performance, suggesting the need for model tuning specific to the recording device.

Threshold Sensitivity An important observation is the significant variance in optimal threshold levels between datasets; the threshold for EAR-Aging was around 0.243, whereas for SINS it was 0.625. This indicates the EBM's high sensitivity to the audio recording setup, which can hinder its generalizability across datasets.

6.2.2 YAMNet-TV Model

The YAMNet-TV model, designed to specifically detect TV and radio sound events, was subjected to a comprehensive evaluation process across multiple datasets. This subsection presents the results of the evaluation and provides a detailed analysis of the model's performance in various scenarios.

Results Evaluation

The YAMNet-TV model's performance across different datasets highlighted its capability to identify TV sound with varying degrees of success:

EAR-Aging Dataset On the EAR-Aging dataset, the model demonstrated moderate recall and higher specificity, yielding an overall accuracy of approximately 69.29%. These results indicate the model's balanced capability in identifying TV sounds and distinguishing them from other audio events.

EAR-Divorce Dataset The results on the EAR-Divorce dataset showed a relatively high recall but lower specificity, which resulted in a moderate accuracy of 57.24%. This suggests the model's tendency to correctly identify TV sounds, albeit with a higher rate of false positives.

SINS Dataset On the SINS dataset, the model's average recall and specificity were close, with an overall average accuracy of 71.51% (±9.34%). The close performance metrics imply a consistent detection capability across different recording nodes within the dataset.

Lab Dataset The evaluation on the Lab dataset revealed a good balance between recall and specificity, with an accuracy of 71.74%. This indicates that the model is capable of performing well even in complex acoustic scenarios present in the Lab dataset.

Detailed Analysis on the Lab Dataset

A deeper examination of the YAMNet-TV model's performance on the Lab dataset provided insights into its effectiveness under varied conditions:

TV and Conversation Scenario In scenarios where TV audio was present along with background conversation, the model's recall and specificity were particularly high, leading to an impressive accuracy of 84.92%. This demonstrates the model's robustness in distinguishing TV sounds amidst overlapping conversational noise.

Isolated Speaker Analysis When analyzed with isolated speakers, the model's performance varied, with some speakers yielding higher recall and specificity than others. For instance, with Speaker 1 (s1), the model achieved a high recall of 87.24% and specificity of 96.52%, leading to an accuracy of 91.88%. These results suggest that the model may be more attuned to certain speaker characteristics that align with its detection algorithm.

Volume Analysis The model's recall was noticeably different across varying volume levels. At 60db, the recall was relatively low at 25.52%, but the specificity was very high at 97.68%, resulting in an accuracy of 61.60%. Conversely, at 75db, both recall and specificity were high, resulting in an impressive accuracy of 87.82%. This indicates the model's sensitivity to volume changes, affecting its detection accuracy.

Microphone Analysis Similar to the volume analysis, the model's performance was influenced by the choice of microphone. With Microphone 1 (m1), the model had a recall of 76.57% and a specificity of 96.98%, resulting in an accuracy of 86.77%. These results suggest that the model may be optimized for specific recording devices, impacting its generalizability.

Threshold Sensitivity

The evaluation also revealed the YAMNet-TV model's sensitivity to threshold settings. Optimal thresholds varied across datasets, suggesting that the model might require fine-tuning to adapt to the acoustic properties of different data sources. The adaptability of the model's threshold setting is crucial for its application in diverse environments.

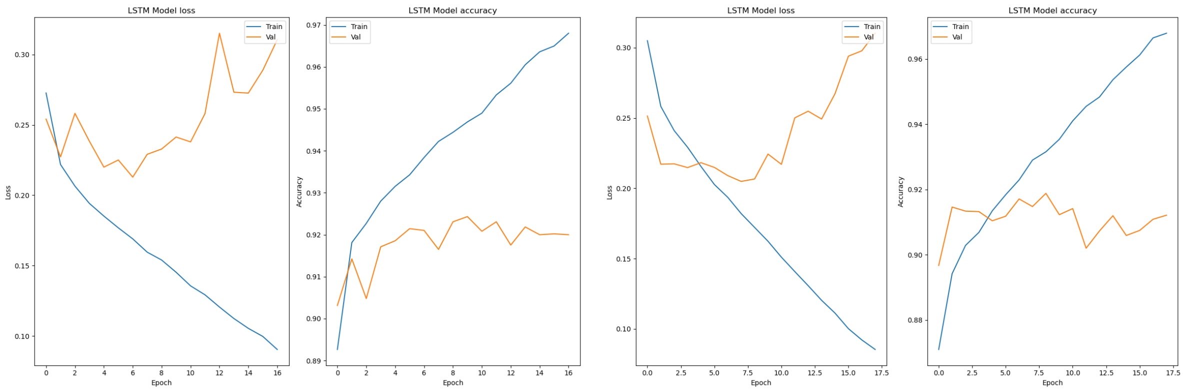
6.2.3 YAMNet-LSTM Model

The YAMNet-LSTM model combines the deep audio embeddings extracted from YAMNet with the sequential pattern recognition capabilities of LSTM networks. This section details the model's implementation and the results of its training evaluation, including an analysis of the training curves and early stopping mechanism.

Training Evaluation and Early Stopping

During the training of the YAMNet-LSTM models, we utilized an early stopping callback to monitor validation loss and prevent overfitting. This approach allows the training to halt when no improvement is observed, reverting to the best model weights achieved during the training process. The training curves for both EAR-Aging and EAR-Divorce models are indicative of the model's learning progression over epochs.

For the EAR-Divorce model, training ceased after 17 epochs, while for the EAR-Aging model, it stopped after 16 epochs. These points represent the moments when the models reached their optimal performance on the validation sets, beyond which no significant improvement in loss reduction was detected.



(a) EAR-Aging model

(b) EAR-Divorce model

Figure 6.4: Training curves for the YAMNet-LSTM models on EAR-Divorce and EAR-Aging datasets.

Model Evaluation Results

The evaluation of the YAMNet-LSTM model across different datasets was aimed at assessing its capacity to generalize and perform consistently in diverse audio environments. Here we detail the performance metrics achieved on each dataset after the model training concluded.

EAR-Divorce Dataset Performance

This model was trained on the EAR-Divorce dataset and evaluated on the reserved test set, EAR-Aging and SINS datasets. The results of this evaluation are presented in Table 6.1.

The YAMNet-LSTM model, when evaluated on the EAR-Divorce dataset's test set, displayed a high level of accuracy. These metrics reflect the model's robust ability to detect TV sounds while maintaining a low rate of false positives. On the EAR-Aging dataset, the model demonstrated similar performance with high accuracy and specificity, indicating its consistency across the two datasets. The evaluation on the SINS dataset yielded an impressive specificity. The high specificity indicates the model's strong perfor-

Table 6.1: Performance of YAMNet-LSTM Model Trained on EAR-Divorce Dataset Across Different Test Sets

Dataset	Accuracy	Sensitivity	Specificity
EAR-Divorce Test Set	89.97%	71.48%	95.92%
EAR-Aging	88.65%	71.05%	96.38%
SINS	88.41% ± 7.14%	77.49% ± 14.08%	99.34% ± 0.56%
Lab - SC Only	75.19%	-	75.19%
Lab - TV with Conv	66.11%	66.11%	-
Lab - TV Only	77.02%	77.02%	-
Lab - Interaction Only	99.65%	-	99.65%

mance in distinguishing between different types of audio events across various recording nodes and activities. The YAMNet-LSTM model's performance was further assessed on the Lab dataset, which offered a variety of challenging scenarios designed to test the model's robustness. The insights from the Lab dataset are particularly revealing. In scenarios where only side conversations were present, the model achieved a modest accuracy. This outcome demonstrates the model's difficulty in differentiating TV sounds from ambient conversational noises. Conversely, in the TV with Conversation and TV Only scenarios, the model exhibited reasonable accuracy and recall, suggesting its capability to detect TV sounds in the presence of or absence of background conversations. In contrast, for the interaction-only scenario, the model displayed excellent specificity, proving its ability to reject false positives when TV sound was absent. We propose in the following a deeper analysis of the model's performance in the Lab dataset.

Impact of Volume on Detection Accuracy Figure 6.5 displays the model's accuracy in TV Only scenarios across different volume levels. The model's accuracy improves with increasing volume, suggesting that the model is more effective at higher volume levels where TV sound characteristics are presumably more pronounced.

Impact of Distance on Detection Accuracy for Side Conversations As shown in Figure 6.6, the model's accuracy decreases as the distance between the side conversation and the microphone increases. This highlights the model's specificity to the proximity of a side conversation, with closer distances yielding better detection of side conversations.

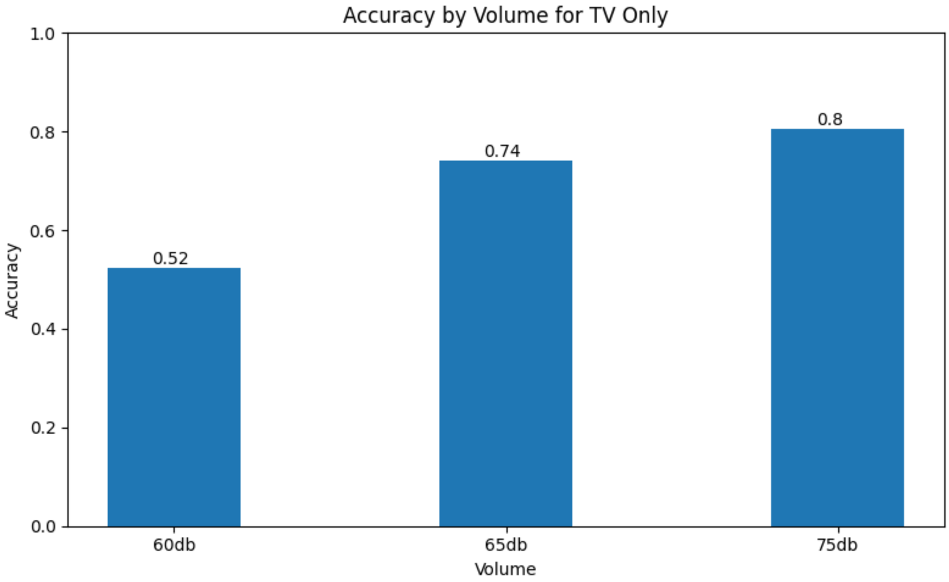


Figure 6.5: Accuracy by volume for TV Only scenarios on the Lab dataset.

It can be further inferred that the model designates 5ft and 10ft as Interactions.

Impact of Microphone Choice on Detection Accuracy Figure 6.7 demonstrates the model's accuracy across different microphones for Side Conversation, TV with Conversation, and TV Only scenarios, respectively. The results indicate that microphone choice has a significant impact on the model's accuracy.

Impact of Speaker Choice on Detection Accuracy Figure 6.8 presents the model's accuracy in TV Only scenario based on the speaker used to play the TV sound. The Small TV speaker resulted in the highest accuracy, which may be due to its sound profile being more effectively captured and processed by the model. The model's accuracy was lower for the Amplifier due to its high bass and low treble, which may have been more challenging for the model to detect.

EAR-Aging Dataset Performance

This model, trained on the EAR-Aging dataset, underwent evaluation on several test sets, including its own reserved test set, the EAR-Divorce dataset, the SINS dataset, and different scenarios within the Lab dataset. The performance metrics are summarized in

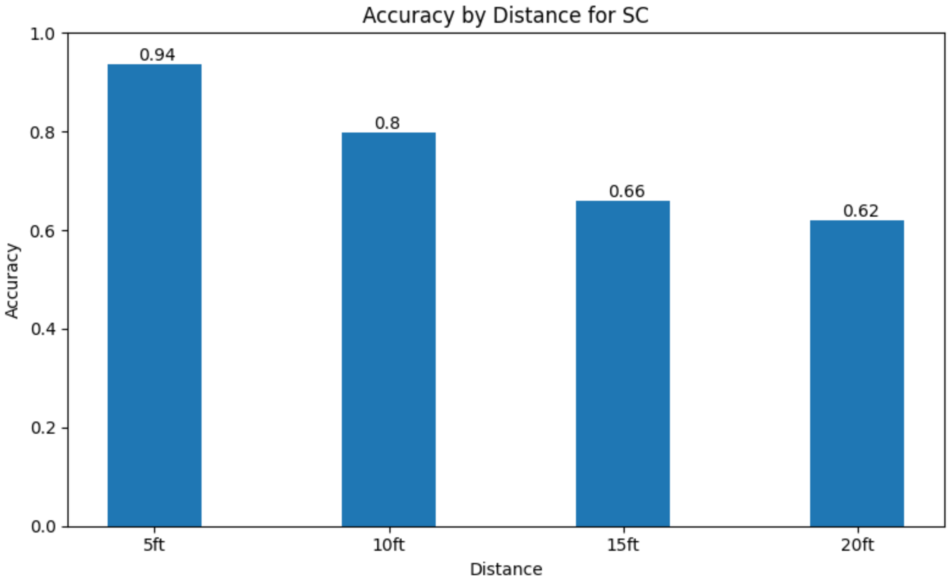


Figure 6.6: Accuracy by distance for Side Conversation scenarios on the Lab dataset.

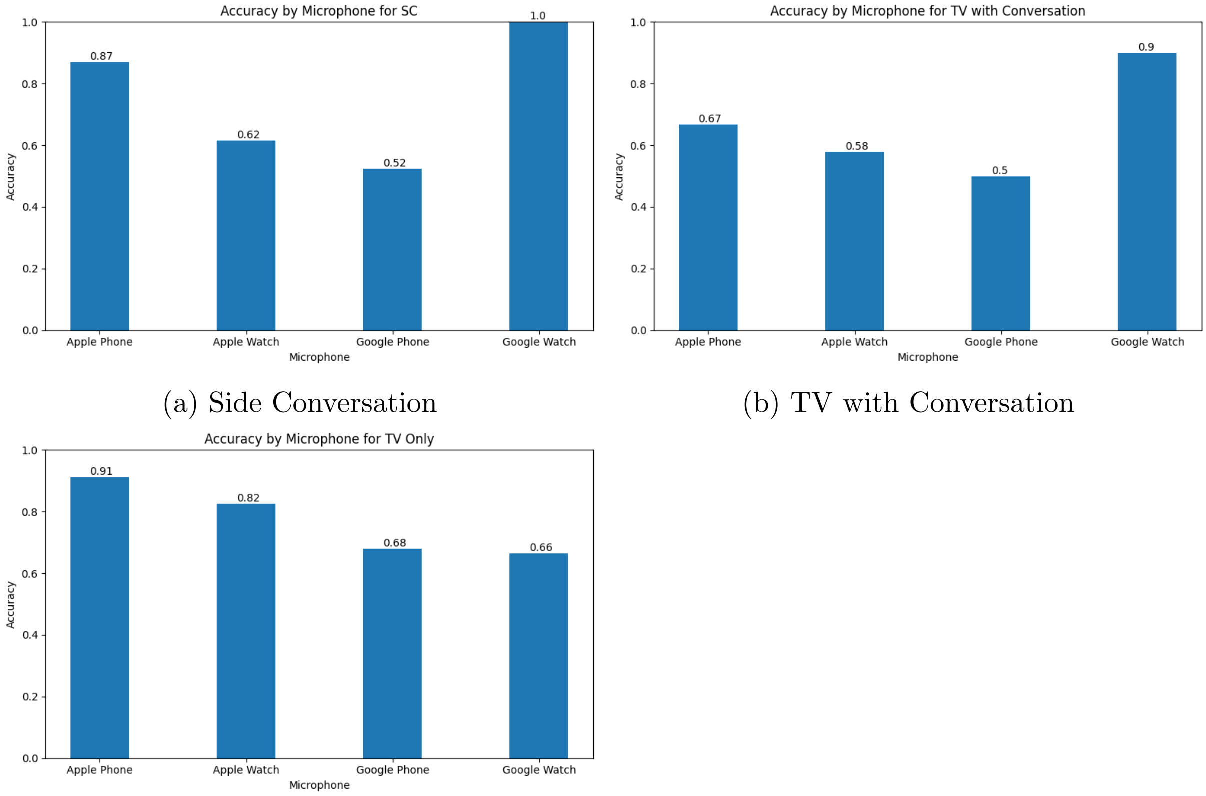
Table 6.2, which outlines the model's accuracy, sensitivity, and specificity across these varied datasets.

Table 6.2: Performance of YAMNet-LSTM Model Trained on EAR-Aging Dataset Across Different Test Sets

Dataset	**Accuracy**	**Sensitivity**	**Specificity**
EAR-Aging Test Set	91%	81.44%	95.20%
EAR-Divorce	89.06%	78.07%	92.59%
SINS	96.69%	96.95%	96.42%
Lab - SC Only	52.5%	-	52.5%
Lab - TV with Conv	79.44%	79.44%	-
Lab - TV Only	89.61%	89.61%	-
Lab - Interaction Only	98.94%	-	98.94%

The model's performance on the EAR-Aging dataset's test set demonstrates high effectiveness, with both sensitivity and specificity above 80%. When evaluated on the EAR-Divorce dataset, the model maintains a similar level of performance, indicating its robustness and generalizability across datasets with different characteristics, similar to EAR-Divorce performance.

On the SINS dataset, the model achieves outstanding accuracy and specificity, highlighting its capability to distinguish between various sound events accurately. This result is particularly notable given the dataset's diversity of recording conditions and the pres-



(a) Side Conversation

(b) TV with Conversation

(c) TV Only

Figure 6.7: Accuracy by microphone for different scenarios on the Lab dataset.

ence of multiple activities.

The Lab dataset's challenging scenarios offered a nuanced view of the model's performance. In scenarios featuring only side conversations, the model struggled, as evidenced by low specificity. This suggests difficulties in distinguishing ambient conversational noises from TV sounds.

In contrast, in scenarios where the TV sound was present with or without conversation (TV with Conversation and TV Only scenarios), the model showed high accuracy and sensitivity, demonstrating its ability to detect TV sounds effectively in the presence of competing audio sources.

The Interaction Only scenario, where no TV sound was present, revealed the model's high specificity, reaffirming its strength in accurately rejecting non-TV sound events.

Impact of Distance The accuracy of this model in identifying side conversations demonstrates a marked decline as the distance between the sound source and the mi-

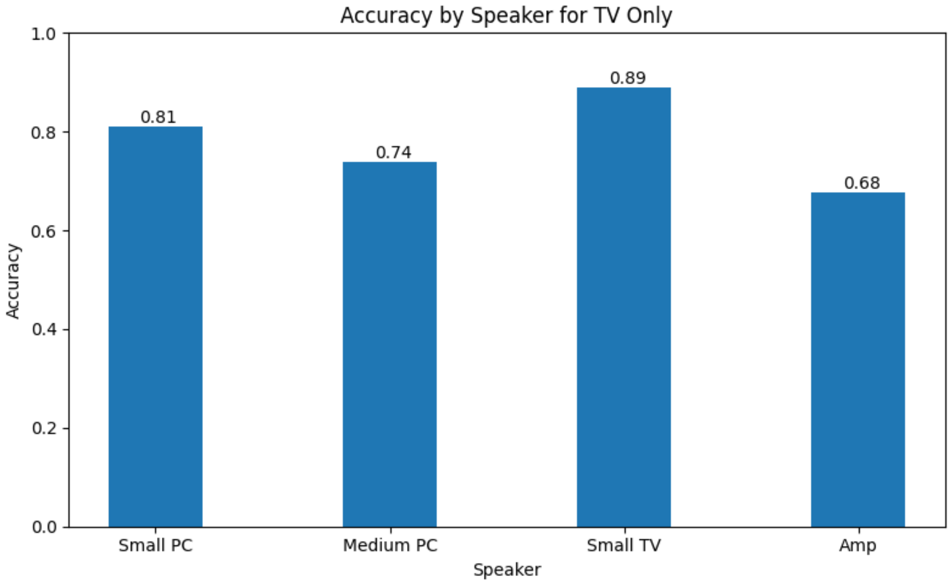


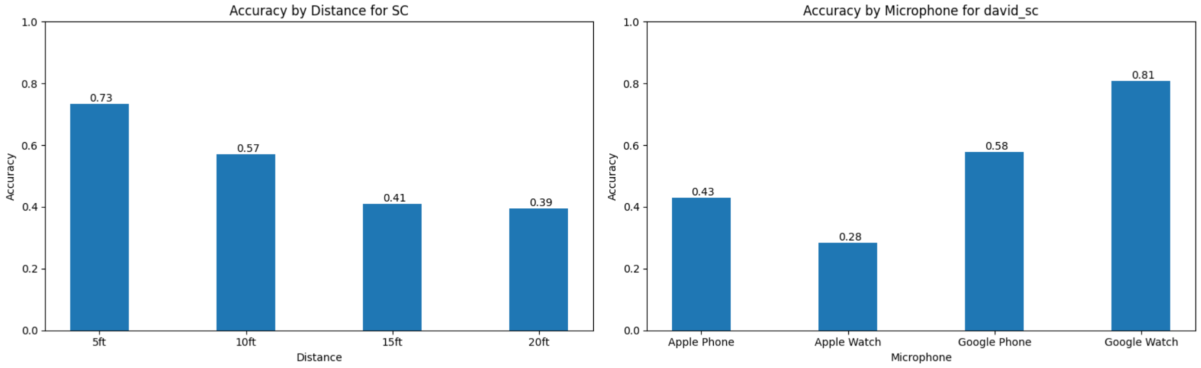
Figure 6.8: Accuracy by speaker for TV Only scenarios on the Lab dataset.

crophone increases. As depicted in Figure 6.9a, at a closer range of 5 feet, the model achieves its peak accuracy. However, as the distance extends to 20 feet, there is a noticeable drop in performance, underscoring the model's sensitivity to distance.

Impact of Microphone Selection Similarly, the choice of microphone significantly affects the model's accuracy. In Figure 6.9b, we observe that different microphones yield varying levels of accuracy in side conversation detection. The superior performance of the Google Watch microphone indicates its potential for higher quality audio capture in this context, while the Google Phone's lower accuracy points to a possible mismatch between the model's training data and the microphone's characteristics.

SINS Dataset Performance Evaluation

The YAMNet-LSTM model's capabilities were further assessed using the SINS dataset, which features diverse environmental sounds captured in a domestic setting. The following analysis highlights the model's performance on this dataset and its ability to generalize to other datasets, including EAR-Divorce, EAR-Aging, and various scenarios within the Lab dataset.



(a) Model accuracy across different distances. (b) Model accuracy with various microphones.

Figure 6.9: Analysis of model performance in side conversation scenarios based on distance and microphone choice.

SINS Dataset The model showcased exceptional performance on its native SINS dataset, as indicated by the nearly perfect accuracy and specificity. This high level of performance suggests that the model can effectively handle environmental sounds and nuances in domestic audio recordings. Table 6.3 summarizes the model's performance on the SINS dataset and other test sets.

Generalization to EAR Datasets When tested on the EAR-Divorce and EAR-Aging datasets, the model experienced a drop in sensitivity, indicating a challenge in detecting TV sounds within datasets it was not trained on. However, the specificity remained high, showcasing the model's ability to correctly reject non-TV sounds.

Performance on the Lab Dataset The Lab dataset, with its varied audio scenarios, posed a significant challenge to the model. The model performed remarkably well in side conversation and interaction-only scenarios, suggesting it has learned to distinguish well between environmental sounds and specific human interactions. In contrast, the TV with Conversation and TV Only scenarios saw a decrease in accuracy, emphasizing the model's difficulty in detecting TV sounds amidst other audio signals.

The model's impressive performance on the SINS dataset and interaction-only scenarios within the Lab dataset suggests it has significant potential for applications in smart home environments and other areas where environmental sound recognition is crucial.

Table 6.3: Performance of YAMNet-LSTM Model Trained on SINS Dataset Across Different Test Sets

Dataset	**Accuracy**	**Sensitivity**	**Specificity**
SINS Test Set	99.81% ± 0.16%	99.72% ± 0.33%	99.90% ± 0.15%
EAR-Divorce	79.25% ± 0.72%	44.32% ± 8.58%	90.48% ± 2.73%
EAR-Aging	82.46% ± 1.72%	53.18% ± 7.01%	95.33% ± 0.92%
Lab - SC Only	99.86% ± 0.22%	-	99.86% ± 0.22%
Lab - TV with Conv	35.63% ± 14.98%	35.63% ± 14.98%	-
Lab - TV Only	46.42% ± 15.61%	46.42% ± 15.61%	-
Lab - Interaction Only	99.50% ± 1.10%	-	99.50% ± 1.10%

However, the reduced sensitivity in scenarios involving TV sounds calls for further model refinement.

6.2.4 YAMNet-LSTM Ensemble Models

To implement this approach, we integrated several YAMNet-LSTM models that were trained on various subsets from the EAR-Aging, EAR-Divorce, and SINS datasets. We ensured each model had a diverse range of training data by dividing the EAR-Aging and EAR-Divorce datasets into balanced folds. We further divided the datasets into multiple folds, with each one containing a balanced proportion of TV and non-TV samples. The SINS dataset was segmented into nodes, each representing a distinct spatial audio recording. For each fold of the EAR datasets and SINS dataset node, separate YAMNet-LSTM models were trained. This resulted in a total of four models for EAR-Aging, four for EAR-Divorce, and seven for SINS. Afterwards, we applied the Confidence Weighted Ensemble method to create five distinct ensemble models:

- Divorce: This ensemble model was trained on the EAR-Divorce dataset and comprised four YAMNet-LSTM models.

- Aging: This ensemble model was trained on the EAR-Aging dataset and comprised four YAMNet-LSTM models.

- SINS: This ensemble model was trained on the SINS dataset and comprised seven YAMNet-LSTM models.

- Divorce-Aging: This ensemble model was trained on the EAR-Divorce and EAR-Aging datasets and comprised eight YAMNet-LSTM models.

- All Ensembles: This ensemble model was trained on the EAR-Divorce, EAR-Aging, and SINS datasets and comprised 15 YAMNet-LSTM models.

Table 6.4 summarizes the performance of the ensemble models on the test sets from the EAR-Divorce, EAR-Aging, and SINS datasets, as well as the Lab dataset.

Table 6.4: Performance of YAMNet-LSTM Ensemble Models Across Different Test Sets

Ensembles	**Divorce**	**Aging**	**SINS**	**Divorce-Aging**	**All Ensembles**
EAR-Divorce Test Set					
Accuracy	84.39%	86.57%	79.65%	86.78%	87.77%
Sensitivity	82.02%	83.77%	45.85%	83.47%	78.12%
Specificity	85.16%	87.47%	90.52%	87.84%	90.88%
EAR-Aging Test Set					
Accuracy	79.90%	90.20%	83.35%	89.58%	89.99%
Sensitivity	86.03%	84.44%	55.36%	84.44%	80.36%
Specificity	77.21%	92.74%	95.65%	91.84%	94.22%
SINS Test Set					
Accuracy	93.89%	98.35%	99.83%	97.36%	98.98%
Sensitivity	88.78%	100.00%	99.66%	98.02%	98.64%
Specificity	99.01%	96.70%	100.00%	96.70%	99.32%
Lab Dataset - Side Conversation Only					
Accuracy	52.12%	15.00%	100.00%	22.31%	62.50%
Specificity	52.12%	15.00%	100.00%	22.31%	62.50%
Lab Dataset - TV with Conversation					
Accuracy	85.83%	98.89%	35.00%	97.78%	90.00%
Sensitivity	85.83%	98.89%	35.00%	97.78%	90.00%
Lab Dataset - TV Only					
Accuracy	93.05%	95.71%	50.52%	96.18%	90.75%
Sensitivity	93.05%	95.71%	50.52%	96.18%	90.75%
Lab Dataset - Interaction Only					
Accuracy	88.73%	88.38%	100.00%	89.44%	99.30%
Specificity	88.73%	88.38%	100.00%	89.44%	99.30%

General Performance Across Datasets: The "All Ensembles" model generally shows strong performance across all datasets, indicating the effectiveness of integrating diverse models to improve generalization.

Lab Dataset - Specific Scenarios: In scenarios with only side conversations, the "SINS Ensemble" achieves perfect accuracy and specificity, which might suggest that models trained on environmental sounds can excel in scenarios that mimic their training conditions. "TV with Conversation" and "TV Only" scenarios show high accuracy with the "Aging Ensemble" and "Divorce-Aging" ensemble, highlighting their potential in recognizing TV sounds amidst conversations.

Sensitivity and Specificity Trade-Off: The "SINS Ensemble" shows a trend of high specificity across the board, indicating a strong ability to correctly reject non-TV sounds. The "Aging Ensemble" demonstrates high sensitivity, especially on the SINS dataset, suggesting it is better at detecting the presence of TV sounds compared to other sounds.

Robustness in Complex Environments: In complex environments such as the Lab dataset, which simulates real-world conditions, there is a notable decrease in performance for certain ensembles, particularly in "TV with Conversation" scenario. This may reflect the challenges in detecting specific audio events within noisy or overlapping sound environments.

Individual vs. Ensemble Model Performance

When comparing the individual YAMNet-LSTM models trained on specific datasets to their respective ensemble approaches, several patterns emerge that highlight the strengths and weaknesses of ensemble methods. Figure 6.10 summarizes performance of both individual models and their respective ensembles on the EAR-Divorce, EAR-Aging, and SINS datasets. We have also included the results of the fine-tuned YAMNet-LSTM model on the combined EAR-Divorce and EAR-Aging datasets, as well as on the EAR-Divorce and EAR-Aging datasets combined with the SINS dataset (All model) in this figure. These two models offer a more complete comparison of individual and ensemble models.

EAR-Divorce Individual vs. Ensemble: The EAR-Divorce ensemble model shows a decrease in accuracy, sensitivity, and specificity on its own test set compared to the individual model (84.39% vs. 89.97% accuracy). However, on the EAR-Aging and SINS test sets, the ensemble model's performance is closer to the individual model's performance, indicating that the ensemble method may provide a more balanced performance across different datasets.

EAR-Aging Individual vs. Ensemble: On the EAR-Divorce test set, the ensemble model has a lower accuracy compared to the individual model. However, on the SINS test set, the ensemble model significantly outperforms the individual model in accuracy (98.35% vs. 96.69%).

SINS Individual vs. Ensemble: The SINS ensemble model displays outstanding performance on its own dataset, marginally improving upon the individual model's already high metrics.

Lab dataset scenarios: The ensemble models generally show improved performance compared to individual models, with the "All Ensembles" approach often leading to the best results. This is particularly evident in the "Side Conversation Only" and "Interaction Only" scenarios, where accuracies reach up to 62.50% and 99.30%, respectively.

6.3 Discussion

The comprehensive evaluation of the various models developed in this study provides valuable insights into their strengths, limitations, and suitability for different scenarios. This discussion synthesizes the findings from the evaluations, offering a critical analysis of each model's performance and highlighting key observations. Table 6.5 summarizes the consolidated performance of the different models across datasets.

6.3.1 EBM Model

The EBM model's performance was found to be moderately effective in simpler scenarios but struggled in more complex environments. The model demonstrated decent recall and specificity in the EAR-Aging dataset, indicating its potential for identifying TV sound amidst everyday noises. However, its performance in the Lab dataset, particularly in differentiating between ”TV with Conversation” and ”Conversation Only” scenarios, was less impressive. This suggests that while the EBM model can be useful in less noisy environments, its reliability decreases in more acoustically complex settings. Moreover, the approach encounters significant challenges when faced with diversified environmental setups. This was particularly evident in the Lab dataset, where the model's performance improved when the speaker setup was fixed, suggesting a sensitivity to variations in acoustic conditions. The EAR-Aging and EAR-Divorce datasets, featuring multiple participants in varied setups, further exemplify the limitations of this approach in adapting to different environmental contexts. Consequently, while the EBM model shows promise in controlled or homogeneous settings, its applicability in dynamic, real-world situations involving diverse acoustic environments remains limited.

6.3.2 YAMNet-TV Model

The YAMNet-TV model showed a balanced capability in identifying TV sounds across different datasets. Its moderate performance on the EAR-Aging and EAR-Divorce datasets was notable, especially considering the diversity of ambient sounds in these datasets. However, its decreased accuracy in the Lab dataset's ”TV with Conversation” scenario indicates a challenge in processing overlapping audio sources, a common occurrence in real-world environments.

6.3.3 YAMNet-LSTM Model

The YAMNet-LSTM model's combination of deep audio embeddings and sequential pattern recognition proved effective, particularly on the SINS dataset where it demonstrated outstanding specificity. This suggests the model's strong potential in applications requiring environmental sound recognition. However, its performance varied across the Lab dataset's scenarios, indicating that while it can handle some complex acoustic situations well, it may require further tuning for others.

6.3.4 YAMNet-LSTM Ensemble Models

The ensemble models generally exhibited improved performance across various datasets, compared to individual models. This improvement was particularly evident in scenarios involving side conversations and interactions only, demonstrating the ensembles' enhanced capability in distinguishing specific audio events in complex acoustic environments. However, the performance in "TV with Conversation" scenarios was still challenging, reflecting the inherent difficulty in isolating TV sounds from overlapping audio.

6.3.5 Comparative Insights

When comparing individual models with their corresponding ensemble versions, several trends emerged:

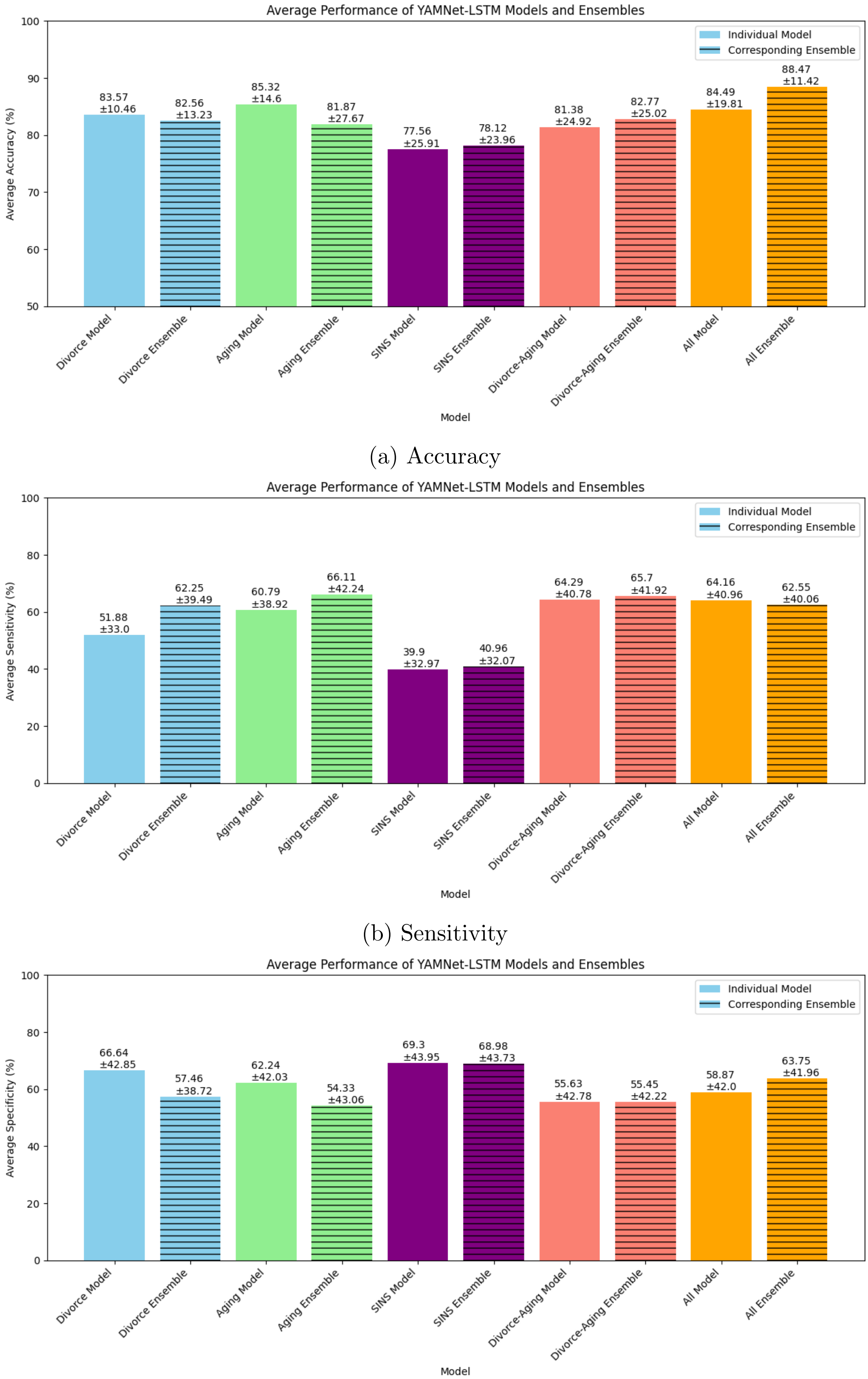
The ensemble models often provided a more balanced performance across different datasets. In the Lab dataset, ensemble models generally outperformed individual models, especially in scenarios with clear-cut audio events (e.g., side conversations only). The "All Ensembles" model frequently led to the best results, indicating the effectiveness of integrating diverse models for improved generalization.

6.3.6 General Observations

Complexity of Real-world Scenarios: All models faced challenges in scenarios with overlapping audio sources, a common situation in real-world environments. This underscores the need for further research and development in handling such complex acoustic scenarios. Sensitivity to Environmental Factors: The models' performance was influenced by factors such as volume, distance, and the choice of microphone or speaker. This variability highlights the importance of considering these factors in practical applications. Potential Applications: While the models showed varying degrees of effectiveness, their potential in applications like smart home environments, where environmental sound recognition is crucial, is evident. However, their limitations in more challenging scenarios suggest the need for continued development and refinement.

Table 6.5: Consolidated Performance of Different Models Across Datasets

Model/Approach	**Dataset**	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**
EBM Model	EAR-Aging	69.91	65.88	71.68
	EAR-Divorce	61.90	95.00	54.11
	SINS	60.28 ± 11.18	80.30 ± 11.18	40.27 ± 13.68
	Lab Dataset	63.30	91.65	32.43
YAMNet-TV Model	EAR-Aging	69.29	61.16	72.86
	EAR-Divorce	57.24	72.27	52.40
	SINS	71.51 ± 9.34	0.7208 ± 15.78	70.93 ± 6.41
	Lab Dataset	71.74	70.66	94.26
YAMNet-LSTM (Divorce)	EAR-Divorce Test	89.97	71.48	95.92
	EAR-Aging	88.65	71.05	96.38
	SINS	88.41 ± 7.14	77.49 ± 14.08	99.34 ± 0.56
	Lab - SC Only	75.19	-	75.19
	Lab - TV w/ Conv	66.11	66.11	-
	Lab - TV Only	77.02	77.02	-
	Lab - Interaction Only	99.65	-	99.65
YAMNet-LSTM (Aging)	EAR-Aging Test	91.00	81.44	95.20
	EAR-Divorce	89.06	78.07	92.59
	SINS	96.69	96.95	96.42
	Lab - SC Only	52.50	-	52.50
	Lab - TV w/ Conv	79.44	79.44	-
	Lab - TV Only	89.61	89.61	-
	Lab - Interaction Only	98.94	-	98.94
Ensemble (All Models)	EAR-Aging	89.99	80.36	94.22
	EAR-Divorce	87.77	78.12	90.88
	SINS	98.98	98.64	99.32
	Lab - SC Only	62.50	-	62.50
	Lab - TV w/ Conv	90.00	90.00	-
	Lab - TV Only	90.75	90.75	-



(a) Accuracy

(b) Sensitivity

(c) Specificity

Figure 6.10: Comparison of individual YAMNet-LSTM models and their respective ensembles across different datasets.

Chapter 7 Conclusion and Future Works

7.1 Contributions

This dissertation has explored the challenging problem of TV sound detection in diverse acoustic environments. The primary contributions of this research can be summarized as follows:

- **Development of Novel Models:** We assessed and tested existing models for TV sound detection and analyzed their performance. We then developed a new model that utilizes ensemble techniques to enhance its accuracy and resilience.

- **Extensive Evaluation Across Datasets:** Our models were rigorously tested on the EAR-Aging, EAR-Divorce, SINS, and a specially curated Lab dataset. This comprehensive evaluation strategy not only demonstrated the models' effectiveness but also their limitations in real-world settings.

- **Insights into Model Performance:** The research provided valuable insights into factors affecting model performance, such as hardware variability, acoustic conditions, and environmental complexity. These insights can guide future developments in the field.

- **Advancement in Sound Detection Techniques:** The research has contributed to the broader field of sound detection by demonstrating the potential and challenges of using machine learning models for audio classification in complex and noisy environments.

7.2 Future Works

This research opens several avenues for future work:

- **Enhancing Model Generalizability:** Future work could focus on improving the models' ability to generalize across more diverse datasets and real-world scenarios, potentially using more sophisticated ensemble techniques or deep learning architectures.
- **Addressing Overlapping Audio Sources:** There is a need for further research to develop models that can more effectively distinguish between overlapping audio sources, such as TV sound and conversation.
- **Expanding Dataset Diversity:** Collecting and incorporating more varied datasets, including those with different languages and cultural contexts, can help in enhancing the robustness of the models.
- **Real-Time Processing and Application:** Implementing these models in real-time applications, such as smart home devices or surveillance systems, and assessing their performance in such settings would be a valuable extension of this research.

7.3 General Conclusion

This dissertation has made significant strides in addressing the complex problem of TV sound detection in varied environments. Through the development and evaluation of several models, we have demonstrated the potential of machine learning approaches in audio classification tasks. However, the research also highlights the challenges that arise due to the complexity of real-world audio environments. While the models developed show promise, there is room for improvement, particularly in terms of generalizability and performance in overlapping sound scenarios. The insights gained from this research contribute to the broader understanding of sound detection and offer a foundation for future advancements in this exciting and rapidly evolving field.

Bibliography

[1] Theodoros Giannakopoulos and Aggelos Pikrakis. *Introduction to Audio Analysis: A MATLAB Approach*. 1st. USA: Academic Press, Inc., 2014. ISBN: 0080993885.

[2] Lie Lu, Hao Jiang, et al. “A robust audio classification and segmentation method”. In: *Proceedings of the ninth ACM international conference on Multimedia*. 2001, pp. 203–211.

[3] Dan Stowell, Michael D Wood, et al. “Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge”. In: *Methods in Ecology and Evolution* 10.3 (2019), pp. 368–380.

[4] Y. Ephraim and D. Malah. “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32.6 (1984), pp. 1109–1121. DOI: 10.1109/TASSP.1984.1164453.

[5] Jort F. Gemmeke, Daniel P. W. Ellis, et al. “Audio Set: An ontology and human-labeled dataset for audio events”. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017, pp. 776–780. DOI: 10.1109/ICASSP.2017.7952261.

[6] Sapna Maheshwari. *Burger King ‘O.K. Google’ Ad Doesn’t Seem O.K. With Google*. 2017. URL: https://www.nytimes.com/2017/04/12/business/burger-king-tv-ad-google-home.html.

[7] Baljinder Kaur and Jaskirat Singh. “Audio Classification: Environmental sounds classification”. In: (2021). ffhal-03501143f.

[8] Anantshesh Katti and M. Sumana. “Pipeline for Pre-processing of Audio Data”. In: *IOT with Smart Systems*. Ed. by Jyoti Choudrie, Parikshit Mahalle, et al. Singapore: Springer Nature Singapore, 2023, pp. 191–198. ISBN: 978-981-19-3575-6.

[9] Beth Logan et al. “Mel frequency cepstral coefficients for music modeling.” In: *Ismir*. Vol. 270. 1. Plymouth, MA. 2000, p. 11.

[10] Hemanta Kumar Palo, Mahesh Chandra, et al. “Recognition of human speech emotion using variants of mel-frequency cepstral coefficients”. In: *Advances in Systems, Control and Automation: ETAEERE-2016* (2018), pp. 491–498.

[11] Wenjun Yang and Sridhar Krishnan. “Combining temporal features by local binary pattern for acoustic scene classification”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.6 (2017), pp. 1315–1321.

[12] Md Rayhan Ahmed, Towhidul Islam Robin, et al. “Automatic Environmental Sound Recognition (AESR) Using Convolutional Neural Network.” In: *International Journal of Modern Education & Computer Science* 12.5 (2020).

[13] Stavros Ntalampiras, Ilyas Potamitis, et al. “Automatic recognition of urban soundscenes”. In: *New directions in intelligent interactive multimedia* (2008), pp. 147–153.

[14] L.R. Rabiner and B.H. Juang. *Fundamentals of Speech Recognition*. Pearson Education signal processing series. Pearson Education, 1993. ISBN: 9788129701381. URL: https://books.google.com/books?id=hoVLAAAACAAJ.

[15] Karol J Piczak. “Environmental sound classification with convolutional neural networks”. In: *2015 IEEE 25th international workshop on machine learning for signal processing (MLSP)*. IEEE. 2015, pp. 1–6.

[16] Zheng Fang, Bo Yin, et al. “Fast environmental sound classification based on resource adaptive convolutional neural network”. In: *Scientific Reports* 12.1 (2022), p. 6599.

[17] Oguzhan Gencoglu, Tuomas Virtanen, et al. “Recognition of acoustic events using deep neural networks”. In: *2014 22nd European signal processing conference (EUSIPCO)*. IEEE. 2014, pp. 506–510.

[18] Anam Bansal and Naresh Kumar Garg. “Environmental Sound Classification: A descriptive review of the literature”. In: *Intelligent Systems with Applications* 16 (2022), p. 200115. ISSN: 2667-3053. DOI: `https://doi.org/10.1016/j.iswa.2022.200115`. URL: `https://www.sciencedirect.com/science/article/pii/S2667305322000539`.

[19] Manoj Plakal and Dan Ellis. *YAMNet.* TensorFlow implementation of YAMNet for audio event classification. Jan. 2020. URL: `https://github.com/tensorflow/models/tree/master/research/audioset/yamnet`.

[20] Anmol Gulati, James Qin, et al. “Conformer: Convolution-augmented transformer for speech recognition”. In: *arXiv preprint arXiv:2005.08100* (2020).

[21] Keunwoo Choi, György Fazekas, et al. “Convolutional recurrent neural networks for music classification”. In: *2017 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2017, pp. 2392–2396.

[22] Jesús Bernardino Alonso Hernández, Marıa Luisa Barragán Pulido, et al. “Speech evaluation of patients with alzheimer’s disease using an automatic interviewer”. In: *Expert Systems with Applications* 192 (2022), p. 116386.

[23] Kevin Berner and Alana N Alves. “A scoping review of literature using speech recognition technologies by individuals with disabilities in multiple contexts”. In: *Disability and Rehabilitation: Assistive Technology* 18.7 (2023), pp. 1139–1145.

[24] Joanna J Parga, Sharon Lewin, et al. “Defining and distinguishing infant behavioral states using acoustic cry analysis: is colic painful?” In: *Pediatric research* 87.3 (2020), pp. 576–580.

[25] Sungkyun Chang, Donmoon Lee, et al. “Neural audio fingerprint for high-specific audio retrieval based on contrastive learning”. In: *ICASSP 2021-2021 IEEE Inter-*

national Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. 2021, pp. 3025–3029.

[26] Yujiang Pu and Xiaoyu Wu. "Audio-guided attention network for weakly supervised violence detection". In: *2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE).* IEEE. 2022, pp. 219–223.

[27] Alexandros Papangelis, Mahdi Namazifar, et al. "Plato dialogue system: A flexible conversational ai research platform". In: *arXiv preprint arXiv:2001.06463* (2020).

[28] Julien Meyer, Laure Dentel, et al. "Speech recognition in natural background noise". In: *PloS one* 8.11 (2013), e79279.

[29] Ian Goodfellow, Yoshua Bengio, et al. *Deep Learning.* `http://www.deeplearningbook.org`. MIT Press, 2016.

[30] Francesc Alıas, Joan Claudi Socoró, et al. "A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds". In: *Applied Sciences* 6.5 (2016), p. 143.

[31] Logan Blue, Luis Vargas, et al. "Hello, is it me you're looking for? differentiating between human and electronic speakers for voice interface security". In: *Proceedings of the 11th ACM Conference on Security & Privacy in Wireless and Mobile Networks.* 2018, pp. 123–133.

[32] Gert Dekkers, Steven Lauwereins, et al. "The SINS database for detection of daily activities in a home environment using an acoustic sensor network". In: *Detection and Classification of Acoustic Scenes and Events 2017* (2017), pp. 1–5.

[33] Shaun Nichols. *TV anchor says live on-air 'Alexa, order me a dollhouse' - Guess what happens next.* 2017. URL: `https://www.theregister.com/2017/01/07/tv_anchor_says_alexa_buy_me_a_dollhouse_and_she_does/`.

[34] Yiteng Huang, Thad Hughes, et al. "Supervised noise reduction for multichannel keyword spotting". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE. 2018, pp. 5474–5478.

[35] Xingwei Liang, Zehua Zhang, et al. “Multi-task deep cross-attention networks for far-field speaker verification and keyword spotting”. In: *EURASIP Journal on Audio, Speech, and Music Processing* 2023.1 (2023), p. 28.

[36] Brecht Desplanques, Jenthe Thienpondt, et al. “Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification”. In: *arXiv preprint arXiv:2005.07143* (2020).

[37] Dianwen Ng, Yunqi Chen, et al. “Convmixer: Feature interactive convolution with curriculum learning for small footprint and noisy far-field keyword spotting”. In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE. 2022, pp. 3603–3607.

[38] Muhammad Ejaz Ahmed, Il-Youp Kwak, et al. “Void: A fast and light voice liveness detection system”. In: *29th USENIX Security Symposium (USENIX Security 20).* 2020, pp. 2685–2702.

[39] Chih-Wei Hsu, Chih-Chung Chang, et al. *A practical guide to support vector classification.* 2003.

[40] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[41] Dedy Rahman Wijaya, Farah Afianti, et al. “Ensemble machine learning approach for electronic nose signal processing”. In: *Sensing and Bio-Sensing Research* 36 (2022), p. 100495.

[42] Duc Thuan Do, Tien Thanh Nguyen, et al. “Confidence in Prediction: An Approach for Dynamic Weighted Ensemble”. In: *Intelligent Information and Database Systems: 12th Asian Conference, ACIIDS 2020, Phuket, Thailand, March 23–26, 2020, Proceedings, Part I 12.* Springer. 2020, pp. 358–370.

[43] Matthias R Mehl. “The electronically activated recorder (EAR) a method for the naturalistic observation of daily social behavior”. In: *Current directions in psychological science* 26.2 (2017), pp. 184–190.

[44] Angelina J Polsinelli, Suzanne A Moseley, et al. “Natural, everyday language use provides a window into the integrity of older adults’ executive functioning”. In: *The Journals of Gerontology: Series B* 75.9 (2020), e215–e220.

[45] Karey L O’Hara, Austin Grinberg, et al. “PREPRINT: Contact and Psychological Adjustment following Divorce/Separation”. In: (2019).