2021-04-08

A Hybrid Method for Auralizing Vibroacoustic Systems and Evaluating Audio Fidelity/Sound Quality Using Machine Learning

Andrew Jared Miller
Brigham Young University

A Hybrid Method for Auralizing Vibroacoustic Systems and Evaluating Audio Fidelity/Sound Quality Using Machine Learning

Andrew Jared Miller

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Scott D. Sommerfeldt, Chair
Jonathan D. Blotter
Tracianne B. Neilsen

Department of Physics and Astronomy

Brigham Young University

ABSTRACT

A Hybrid Method for Auralizing Vibroacoustic Systems and Evaluating Audio Fidelity/Sound Quality Using Machine Learning

Andrew Jared Miller
Department of Physics and Astronomy, BYU
Master of Science

Two separate methods are presented to aid in the creation and evaluation of acoustic simulations. The first is a hybrid method that allows separate low and high-frequency acoustic responses to be combined into a single broadband response suitable for auralization. The process consists of four steps: 1) creating separate low-frequency and high-frequency responses of the system of interest, 2) interpolating between the two responses to get a single broadband magnitude response, 3) adding amplitude modulation to the high-frequency portion of the response, and 4) calculating approximate phase information. An experimental setup is used to validate the hybrid method. Listening tests are conducted to assess the realism of simulated auralizations compared to measurements. The listening tests confirm that the method is able to produce realistic auralizations, subject to a few limitations. The second method presented is a machine learning approach for predicting human perceptions of audio fidelity and sound quality. Several algorithms are compared and various audio features considered in developing the machine learning models. The developed models accurately predict human perceptions of audio fidelity and sound quality in three distinct applications: assessing the fidelity of compressed audio, evaluating the fidelity of simulated audio, and comparing the sound quality of loudspeakers. The high accuracies achieved confirm that machine learning models could potentially supplant listening tests, significantly decreasing the time required to assess audio quality or fidelity.

Keywords: auralization, broadband, acoustic simulation, machine learning, audio fidelity, sound quality, listening tests.

Acknowledgments

I would like to offer my special thanks to Dr. Sommerfeldt for all of his insights and guidance. Your willingness and encouragement as I explored the possibility of a machine learning approach for sound quality evaluation are particularly appreciated. I could not have asked for a better advisor.

Thank you, Dr. Blotter, for your feedback throughout my research. You always posed thought provoking questions that substantially improved the quality of my work.

I would also like to show gratitude to Caterpillar Inc. for funding the project, and David for providing direction while allowing autonomy as we investigated different approaches.

Finally, thank you to my wife, Adi. Your loving support helps me persevere, reach higher, and achieve things I could not on my own.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

The overall goal of this research is to develop methods that will help create realistic acoustic simulations of coupled fluid-structure systems. Specifically, the objective is to simulate the acoustic response inside the cab of heavy equipment, or the aural response that would be experienced by the machine operator. The benefits of accurately simulating the acoustic response of a machine cab are twofold: first, the acoustic response can be used in a broader audiovisual simulation that is used to train operators to use the machinery, and second, the simulations can be used as a design tool to help achieve a desired sound.

The present research focuses on both the creation of the acoustic simulations and the evaluation of the simulations' fidelity. The next section summarizes some common methods used to create acoustic simulations of coupled vibroacoustic systems, motivating the creation of the new hybrid simulation method developed here. The following section then reviews methods commonly used to evaluate audio fidelity, motivating the novel machine learning approach developed here. An overview of the remaining chapters is then provided.

1.1 Acoustic Simulation Methods

Accurately simulating the vibroacoustic response of a system is valuable in many industries and applications. Various techniques are used to model a system. Lumped parameter models, which simplify the systems into discrete mass, stiffness, and damping elements and are one of the more basic methods. This is often the first type of model one learns in an introductory physics or engineering course on vibration [1]. Despite its relative simplicity, it is frequently used to accurately model vibroacoustic systems. Fahnline and Koopmann use a lumped parameter method to model the acoustic power radiated from a vibrating structure [2]. Karnopp develops lumped parameter models of acoustic filters similar to those found in exhaust systems [3]. Beranek and Mellow [4], as well as Tilmans [5], detail lumped parameter models of various transducers including loudspeakers and microphones.

Lumped parameter estimates are not appropriate for all systems, and there are numerous methods for modeling continuous vibroacoustic systems. Analytical solutions can be derived for some simpler systems, although the extent of this approach is fairly limited [6]. Numerical methods have been developed and are commonly used for vibroacoustic applications where analytical solutions are not feasible. Finite element analysis (FEA) is a method where a larger system is modeled by breaking it into smaller continuous pieces called finite elements, connected at nodes where all of the interactions occur [7]. This method can be computationally expensive, although advances in computer technology continue to lessen this limitation, and it has become common due to its accuracy and versatility [8]. While it is most common for purely structural/mechanical systems, FEA has been used to model many coupled vibroacoustic systems. Gan, et al. use FEA to model sound transmission through the human ear [9]. Nefske, et al. examine a FEA formulation for structural-acoustic analysis of the enclosed cavity of an automobile passenger compartment [10]. Everstine provides a review of several FEA formulations used to solve coupled fluid-structure acoustic problems [11].

The boundary element method (BEM) is a numerical method for solving boundary value problems of partial differential equations. It is, for the most part, more common than FEA for acoustic problems [8]. Kirkup provides a survey of research and applications of the BEM for vibroacoustic problems [12]. The boundary element method's efficiency for problems where there is a small surface/volume ratio make it ideal for modeling acoustic radiation from vibrating structures; however, FEA tends to be more appropriate for contained systems [13].

Statistical energy analysis (SEA) is another common method for modeling vibroacoustic systems. In SEA, the complex system is divided into subsystems, and the primary objective is to properly balance the distribution of energy among the subsystems. Within each subsystem, the energy is assumed to be equally distributed among the resonance modes, and the resonances are assumed to be uniformly distributed in frequency within specified frequency bands [14]. Thus, SEA only provides an average level in each subsystem and is more accurate at higher frequencies where a higher modal density occurs [15]. Statistical energy analysis is well established and continues to play a role in ongoing research. Price and Crocker use SEA to model sound transmission between rooms through double panels [16]. Chen, et al. propose an affine interval perturbation SEA method to reduce uncertainty in models of a plate-cavity coupled system as well as a simplified launch vehicle fairing [17].

Each of the above methods has its own advantages/disadvantages, but the overall issue common to all methods is that they are only valid or feasible for limited frequency ranges. For many numerical methods, computation time significantly increases for higher frequency ranges – for example computation time is proportional to frequency cubed for rectangular acoustic cavities in a modal analysis [18]. Such methods are termed "low-frequency" because they are often not practical for obtaining a response up to higher frequencies (higher frequencies being defined separately for each unique system depending on the geometry and material/fluid properties). In contrast to these low-frequency methods, which become impractical with increasing frequency, energy-based methods such as SEA typically improve with increasing frequency due to higher modal densities. However, these energy-based methods typically are not able to obtain good low-frequency results where the modal density becomes low.

The research presented here seeks to create a hybrid method, combining a low-frequency method and a high-frequency method to obtain a broadband acoustic response of a vibrating system. The hybrid method is intended to create auralizable responses that can be assessed by listening. The details of this method are provided in Chapter 2.

1.2 Evaluating Audio Fidelity

Audio fidelity and sound quality are important across many industries, from digital applications such as virtual reality, live streaming, and algorithm development to physical applications such as loudspeaker and product design. The two are not necessarily synonymous: audio fidelity denotes an accurate reproduction of sound, while sound quality deals with human preferences and perceptions. However, in some cases, they can be essentially the same thing. For example, many audiophiles would consider fidelity to be equivalent to sound quality. In general, sound quality tends to be broader than fidelity and its specific definition is application dependent [19]. The two will be used somewhat interchangeably here, since the main metric used in this work contains elements of both – accurate reproduction as perceived by human listeners.

Subjective and objective measures are both commonly used to evaluate audio fidelity or sound quality. Subjective measures usually take the form of listening tests. A group of people will listen to the sounds of interest and then will give the sounds some sort of preferential or perceptual rating. This is generally regarded as the best way to get an accurate representation of how humans will actually perceive sounds, and it is used in various industries for audio evaluation. Rydén proposes a listening test method for appropriately assessing audio codecs [20]. Gabrielsson and Lindström [21], as well as Toole [22], show that loudspeaker sound quality can be reliably evaluated through listening tests, although both note the importance of careful experimental design for obtaining repeatable results. Unfortunately, designing and conducting such listening tests is often difficult, time-consuming, and expensive. Objective measures, on the other hand, are things that can be physically measured such as a spectrum or distortion. They are typically more efficient to obtain than subjective measures, but often fail to capture all the nuances of human hearing. Despite being less accurate than subjective measures, objective measures are often used to approximate human preferences and sound quality perceptions when efficiency is important. Campbell, et al. summarize the perceptual evaluation of audio quality (PEAQ) algorithm, one of the only standardized methods for objective audio quality assessment [23]. They note several shortcomings of the PEAQ algorithm and provide a review of more recent advancements in objective audio quality assessment. Gabrielsson, et al. [24] and Rumsey, et al. [25] show that there are correlations between common objective measures and subjective perceptions, but predicting subjective responses based solely on objective measures is only marginally accurate. Deciding whether to use subjective measures or objective measures remains a tradeoff in accuracy vs. efficiency. Bech and Zacharov provide an in-depth discussion of both theory and methodology of perceptual audio evaluation, with the conclusion that many systems are accurately evaluated only through listening tests [26]. Citing the ineffectiveness of objective measures, Cartwright, et al. attempt to improve the efficiency of listening tests through crowdsourcing [27]. Their method significantly decreases the time and cost of performing listening tests, but they acknowledge that the method reduces the amount of control investigators have, adding new variability and potential biases.

This research presents a new approach to evaluate fidelity or sound quality, using machine learning to capture the benefits of both subjective and objective methods. Carefully training a machine learning model on an initial set of listening tests allows it to accurately predict human ratings. Once the model has been trained, it is able to supplant future listening tests, matching the efficiency of traditional objective measures while maintaining the accuracy of listening tests. Some work has been done using machine learning to evaluate sound quality and has proven quite successful, but it remains unstudied and unused in all but a few specific industries. Machine learning techniques for sound quality evaluation are most common in vehicular industries. Huang, et al. use deep belief networks to assess the sound quality of interior car noise [28]. Fang and Zhang use a support vector machine based model to investigate the sound quality of electric vehicle powertrains [29]. Lopes, et al. use artificial neural networks to assess interior sound quality of propellor aircraft [30]. Pietila and Lim provide an in-depth review of machine learning concepts for sound quality evaluation [31]. The large majority of their referenced applications are also automotive or aviation related, with a few other product quality applications such as vacuum and refrigerator noise annoyance. Sottek and Henrique introduce a new family of metric models called AI-SQ Metrics (Artificial Intelligence in Sound Quality Metrics) to evaluate product sound quality [32]. Giraldo, et al. use machine learning techniques to evaluate the tonal quality of live violin performances [33]. Lemaitre and Susini cite the “huge potential” of machine learning techniques for sound quality evaluation, acknowledging the limited scope of applications where it has been tested [34]. Although research in this specific area remains relatively sparse, other applications of using machine learning to classify audio are ubiquitous.

Some well-studied areas include speech recognition and speaker recognition, environmental sound classification, and identification of music genres, instruments, and songs [35].

1.3 Overview

The proposed methods are discussed separately in the next two chapters. Chapter 2 describes the development of the hybrid simulation method for coupled vibroacoustic systems. It begins by providing some additional background about work leading to the hybrid method. Details about the hybrid method follow, along with references to other research attempting to create broadband responses of vibroacoustic systems. An experimental setup is used to validate the hybrid method and results are presented comparing the hybrid method to measurements.

Chapter 3 details the development of a machine learning model for predicting human perceptions of audio fidelity. The development section includes a description of the dataset used, a comparison of various machine learning algorithms, and a summary of various audio features considered. The performance of the fidelity prediction model is then presented. An additional application is then investigated, involving the evaluation of loudspeaker sound quality, and is shown to further validate the methods developed.

Final remarks are given in Chapter 4. These include analysis and conclusions about both of the methods developed in the previous chapters, as well as recommended directions for further work.

Chapter 2 Hybrid Simulation Method

The systems simulated in this research are meant to represent basic structural/acoustic coupling found in heavy equipment, but the methods developed are generally applicable for many applications. As previously mentioned, the ultimate objective is to be able to simulate the acoustic response inside the cab of a machine, or the aural response that would be experienced by the operator. As such, the method needs to create a broadband response of the vibroacoustic system that can then be auralized and evaluated by listening, unlike many other methods that are only evaluated graphically. The measure of success is then directly related to the perception of the simulated sounds, with the goal of creating sounds that are perceived as "realistic", not necessarily perfect, and avoiding an overall perception of artificialness. The next section describes the development of a hybrid method that combines two traditional methods to achieve this goal. An experimental setup, consisting of a rigid acoustic cavity coupled to a vibrating plate, is used to validate the hybrid method. Results are presented comparing the hybrid method to measurements, showing that the hybrid method is able to create realistic acoustic simulations.

2.1 Development

Simplicity and efficiency are two criteria that guided development of the simulation method. Prior to creating the hybrid method, various methods were tested for creating simple approximations of a machine cab response. Two methods emerged as desirable solutions based on these criteria: classical modal analysis (CMA) and SEA. Neither method is sufficient to create the desired broadband response on its own, since CMA is only feasible at lower frequencies and SEA is only valid at higher frequencies. However, the hybrid method shows how the two can be combined to create a unified broadband response suitable for auralization.

In a classical modal analysis (CMA), the in vacuo structural modes and rigid boundary acoustic modes are first determined analytically. These independent analytical modes are then combined through spatial coupling coefficients. The final response is then obtained by summing over the total number of modes to be used for the desired frequency bandwidth [36, 37]. A convenient matrix formulation of CMA was developed by Kim and Brennan that is based on the impedance/mobility approach [38]. This allows for efficient calculation of coupled responses when the independent analytical modes can be determined. The low-frequency responses presented later in this paper were obtained using this matrix formulation of CMA. If determining the analytical modes is infeasible, FEA could be used to determine the low-frequency responses.

SEA was used to create the high-frequency responses. SEA is very computationally efficient, making it an ideal candidate for the simple approximate model developed in this paper. One major limitation of SEA is that, since it results in average levels, the final responses do not contain the phase information necessary to create the impulse response needed for creating auralizations. The hybrid method outlined below seeks to overcome this limitation and allow for SEA responses to be combined with a low-frequency response to create auralizable broadband acoustic responses.

2.1.1 Hybrid Method

There are many applications where only considering either low or high frequencies is not enough, and broadband responses are required. This is particularly true when the final response will be auralized, since human hearing spans approximately 20 Hz – 20 kHz, and a reduced frequency range is often perceived as unnatural. Significant research has been done investigating ways to achieve broadband simulations involving acoustic excitation from coupled vibrating structures. No universal method has been found and it remains an active area of research [39]. Many of the proposed methods are quite complex and/or application specific, keeping them from being more widely adopted. Wang, et al. use a hybrid approach combining a node-based smoothed finite element method (FEM) and SEA and show good results for several theoretical systems; however, they apply the different methods to separate subcomponents of the system, leaving the SEA portions absent of any phase information [40]. Chronopoulos, et al. incoorporate a wave FEM with SEA to better account for dispersion in curved shells [41]. Yotov, et al. introduce a non-parametric stochastic FEM allowing them to accurately model responses of spacecraft in high-frequency ranges where structures begin to exhibit chaotic behavior and element-based techniques are typically unreliable [42]. Aretz, et al. combine FEA, image sources, and stochastic ray tracing to simulate broadband impulse responses [43]. This work is most similar (in objective, not method) to the research presented in this paper, but the method does not achieve the simplicity aimed for here and would be difficult to implement in more complex systems. They provide additional references to similar work, citing limitations and unsatisfactory results in most cases.

There are some established methods that combine low- and high-frequency methods to compute the response of vibroacoustic systems. Certain computer software packages, VA One [44] for example, will simulate complex systems by combining individual components that are each modeled by either FEA or SEA. The user determines which method (FEA or SEA) will be used for each component depending on its geometry and material properties. Such computer simulations can provide accurate responses for complex systems; however, they can become extremely computationally expensive/time consuming when a large frequency range is desired. Additionally, the energy-based portion of the solution only provides an average level, and it does not capture any resonance or phase information. This becomes problematic if one desires to auralize the simulated response.

There are two main drawbacks in many of the existing broadband solutions. First, the methods are often quite complex. They either require significant computation time or they involve complicated mathematical techniques that are only applicable in specific situations. As previously stated, the goal of this project is to create a simple method to model vibroacoustic systems that is both computationally efficient and simple enough to easily change and apply in various configurations. Of course, there must be a tradeoff here: the simpler the model, the less likely it will be able to capture all the complexities of the system. Accordingly, the measure of success is creating a method where the resultant models sound “realistic”, not perfect.

The second drawback is related to the way that many of the existing broadband solutions are evaluated. Plotting the magnitude of the frequency response of the system is the most common way that model accuracy is evaluated. Even when different methods are used in different frequency ranges, the results are often just plotted side by side, without providing any real way to combine the results into a single overall response [45]. This may be sufficient in many instances; however, our main concern is about how the simulated sounds are perceived compared to real sounds. Therefore, our method needs to produce a result that can be auralized. It must be a single response that contains both magnitude and phase information across the frequency range of interest.

The hybrid method developed here seeks to overcome these two problems. In the end, it creates a simple model that produces auralizations that are reasonable approximations of how the real system sounds. There are four steps in the process: 1) creating a separate low-frequency modal response and a high-frequency SEA response of the system, 2) interpolating between the two responses to get a single broadband magnitude response, 3) adding amplitude modulation to the SEA portion of the response, and 4) calculating approximate phase information. Each of these steps is discussed below and pictured in Fig. 2-1.

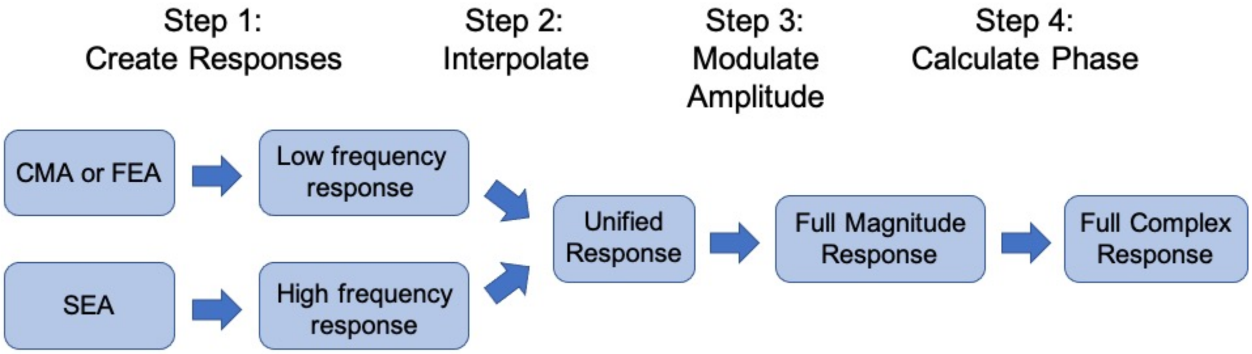

Figure 2-1. *Diagram representing the steps in the hybrid model process: Step 1) creating a separate low-frequency response and a high-frequency response, Step 2) interpolating between the two responses to get a single broadband magnitude response, Step 3) adding amplitude modulation to the high-frequency portion of the response, and Step 4) calculating approximate phase information.*

First, two separate responses are calculated, one using a low-frequency method and one using a high-frequency method. For this project, classical modal analysis, based on a matrix formulation developed by Kim and Brennan, was used to calculate the low-frequency response [38]. Although FEA is probably more commonly used due to its accuracy and ease of implementation with modern software packages, CMA was chosen because of its simplicity and computational efficiency. By way of illustration, a model of an acoustic cavity coupled on one side to a simply-supported vibrating plate (see experimental setup section below) was created with both FEA and CMA; the full finite element forced response mode superposition model took about 44 minutes to run while the classical modal model took only about 1.2 minutes to obtain a frequency response with the same frequency resolution and bandwidth, showing the benefit of using the CMA approach. The high-frequency response was obtained by building an SEA model in the computer program VA One. The SEA response was calculated in one-third octave bands. SEA also meets the simplicity and efficiency criteria.

Second, a single magnitude response was created by interpolating between the separate low and high-frequency response magnitudes. At this point, only the magnitude response can be obtained because the SEA portion of the response does not contain any phase information. Built in MATLAB interpolation methods were used to obtain the single unified response. It is important to choose the interpolation method carefully to avoid unexpected results (MATLAB documentation recommends using interp1 with the 'pchip' interpolation method when the signal x is not slowly varying). Determining the crossover frequency, or the point at which to switch from the modal response to the SEA response, is another important consideration in this step. Various crossover frequencies were tried, and it was found that examining the number of modes per frequency band is useful in determining an appropriate crossover frequency. This is discussed further in the results section, and guidance is given on how many modes should be present in a one-third octave band before crossing over to SEA. Once the crossover frequency was determined, each of the individual responses was truncated; everything above the crossover frequency was discarded from the modal response, and everything below the next one-third octave band center frequency was discarded from the SEA response, leaving a gap between the crossover frequency and the next one-third octave band center frequency. This gap allowed for a smoother transition between the separate low- and high-frequency responses. The two separate responses were then combined via MATLAB's interp1 function, and the resulting unified response was resampled to a 1 Hz frequency resolution to match the resolution of the SEA portion to that of the modal portion.

Third, the SEA response only captures the average level across frequency, it does not capture any information about resonances/antiresonances. This makes for a very smooth unrealistic response. Of course, it is unknown where the resonances/antiresonances would have occurred – a classical modal model or finite element model would be required to know. However, a more realistic response can be obtained by randomly adding amplitude modulation to the SEA response. Although randomly modulating the response will not create peaks at the exact same frequencies as the real system, it was found that it is sufficient to create a more realistic sounding response. This is because the SEA response is only used in a frequency range where the modal density is high. In this frequency range, the exact location of the peaks is less important than in the lower frequency range covered by the modal model. The amplitude modulation also has the added benefit of helping create a more realistic phase in the next step. There are two important considerations when creating the amplitude modulation: the magnitude of the modulation and how rapidly the modulation occurs along the frequency axis. The magnitude of the modulation is representative of the damping in the system. Large amplitude modulation represents a system with little damping and results in a high-pitched "metallic" ringing sound in the final simulation. On the other hand, low amplitude modulation represents a system with high damping and results in little to no ringing in the final simulation. As expected, modulating the amplitude of the SEA response only affects the high-frequency ringing (or the ringing in the frequency range where the SEA response is used), while low-frequency ringing is determined by the modal response. Determining the appropriate amplitude to modulate the signal can be challenging since it is often difficult to predict the damping in a complex system. One method is to use the magnitude of the peaks and dips in the low-frequency portion of the response to estimate the amplitude by which to modulate the SEA portion. This was done by visually inspecting plots of the magnitude of the low-frequency portion, although an algorithmic method could be implemented to streamline the process. It was found that it is better to overestimate the damping (underestimate the amplitude of modulation) in the SEA portion of the response, because extensive high-frequency ringing tends to cause the simulated sounds to be perceived as artificial sounding. The amplitude modulation formula used for the results presented in this paper is given by

$$A' = A * \mathrm{lognrand}(\mu, \sigma) \tag{2.1}$$

where A is the original amplitude, A' is the modified amplitude, and lognrand() is a MATLAB function producing lognormal random numbers with parameters μ (mean of logarithmic values) and σ (standard deviation of logarithmic values). The parameter values used to produce the results presented in this paper were $\mu = 0$ and $\sigma = 0.5$. Determining how rapidly to modulate the amplitude along the frequency axis is a second concern. Although the exact resonances of the coupled system are not known, the uncoupled natural frequencies of the dominant components can be used to estimate an appropriate density of peaks and dips in the frequency response. In the plate-cavity system described in the experimental setup section below, the resonance frequencies of the plate served as an appropriate approximation.

Fourth, in order to auralize the response it needs to have phase information as well as magnitude. As mentioned before, an SEA response does not contain any phase information. Therefore, to finalize the response, an approximate phase needs to be calculated. Significant time was spent experimenting with various ways of creating this approximate phase. Some things that were tried include random modulation similar to the modulation added to the magnitude, calculating a minimum phase via the Hilbert transform, setting the phase at each peak/dip in the magnitude response to π or $-\pi$ respectively and interpolating in-between, and extrapolating from the unwrapped phase of the low-frequency modal response. Although some of the other methods appeared visually better when plotted, the minimum phase was the only method that did not create noticeable artifacts when calculating an impulse response and auralizing the results. The process of using the Hilbert transform to create the final response consisted of three parts. First, the magnitude response calculated in the previous step was used to create a two-sided spectrum, since the Hilbert transform expects negative frequencies. Second, the Hilbert transform was used to calculate a minimum phase for the given magnitude response. The formula for calculating the minimum phase is given by

$$\phi(\omega) = -\mathcal{H}\left[\ln\left(G(\omega)\right)\right], \tag{2.2}$$

where ϕ is the minimum phase, $\mathcal{H}$ represents the Hilbert transform, and G is the two-sided magnitude response. Third, the final complex frequency response was calculated according to

$$\hat{G}(\omega) = G(\omega) * e^{j\phi(\omega)}, \tag{2.3}$$

where $\hat{G}$ is the two-sided complex frequency response, G is the two-sided magnitude response, and ϕ is the minimum phase. The minimum phase was used across the entire frequency range. This proved less problematic than attempting to interpolate between the existing low-frequency phase and the calculated minimum phase at high frequencies.

An Inverse Fast Fourier transform (IFFT) was then applied to the complex frequency response to obtain an impulse response. The impulse response was convolved with various excitation signals to create auralizations, so the validity of the approach could be assessed by listening.

2.2 Experimental Setup

A simple coupled structural-acoustic system was built to validate the hybrid method. The system consisted of a rectangular acoustic cavity with five rigid walls and one flexible wall. Similar systems have been studied extensively and used many times to validate new methods [37, 46, 47]. The rigid walled acoustic cavity was built with a similar method to that used by Kim and Brennan [38], and the flexible wall was constructed to mimic a simply-supported plate, based on a method proposed by Robin et al. [48]. Details of the system are provided below.

A diagram of the experimental setup is show in Fig. 2-2. Two five sided boxes were constructed using ½ inch medium-density fiberboard (MDF), one larger box and one smaller box designed to sit inside the larger box with a 10 cm gap on all sides. The 10 cm gap between the boxes (including the bottom) was filled with sand so that the inner box acted as a rigid walled acoustic cavity. The inner dimensions of the smaller box were 48 cm x 42 cm x 110 cm. A microphone was located at (20 cm, 18 cm, 63 cm) according to the coordinate system marked in red.

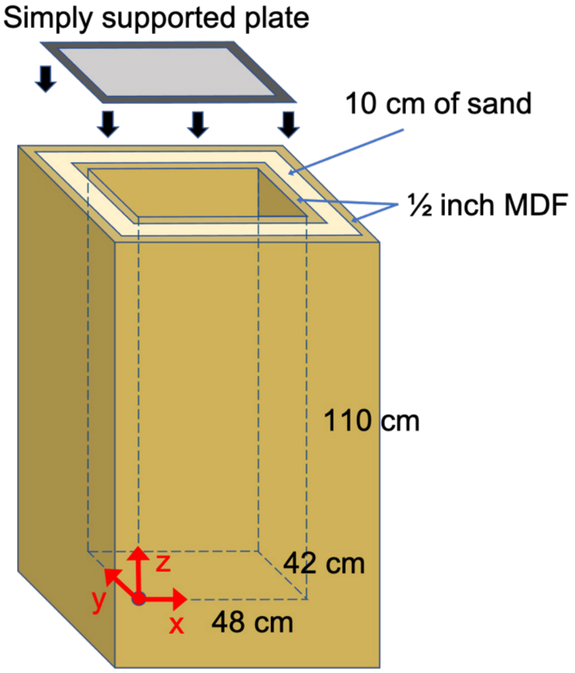

Figure 2-2. *Diagram of the experimental setup, a simply supported plate coupled to an acoustic cavity.*

An aluminum simply supported plate, mounted to a steel frame, was placed on top of the cavity to create the flexible wall (Fig. 2-3). The plate and cavity were designed to minimize any gaps, but prevent touching on the sides, once the plate was placed atop the cavity. This was done so that the plate dimensions and x-y dimensions of the cavity could be assumed equal when modeled, while preserving the simply supported nature of the plate. The plate was measured to be 3.15 mm thick. The plate was excited by a mechanical shaker at (20 cm, 18 cm, 110 cm), directly above the microphone. A force sensor (not pictured) was attached between the shaker and the plate. The transfer function was measured between the force on the plate and the microphone in the cavity.

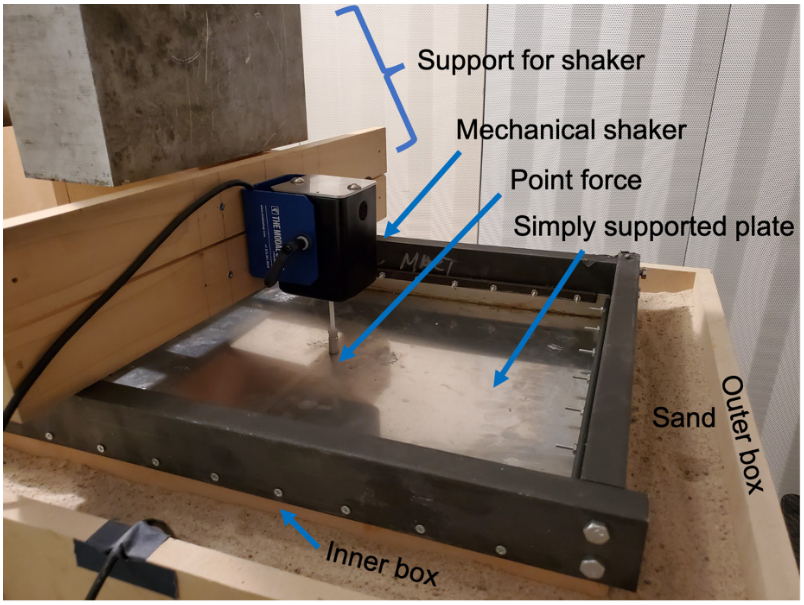

Figure 2-3. *Photograph of the simply supported plate excited by a mechanical shaker.*

2.3 Results

A model of the plate/cavity experimental setup was built using the hybrid method. The shaker was modeled as a point force and the microphone was modeled as a point acoustic sensor. For the low-frequency portion of the response, the matrix modal formulation was used [38]. The modal response was calculated up to 2 kHz. VA One's default material properties for aluminum were used to be consistent with the SEA model: density = 2700 kg/m^3, Poisson's ratio = 0.33, and Young's modulus = 7.1e10 Pa [44]. An airborne sound speed of 340 m/s was used, and the density of air was assumed to be 1.21 kg/m^3, consistent with lab conditions of a room temperature of 20 ºC and an elevation of 1400 m. A damping ratio of 0.01 was used, determined by comparing to the measurement since it can be difficult to estimate damping accurately. The high-frequency portion of the response, above 2 kHz, was obtained by creating a SEA model in

VA One, using all the same parameter values. The low-frequency CMA and high-frequency SEA responses are shown alongside the measured response in Fig. 2-4.

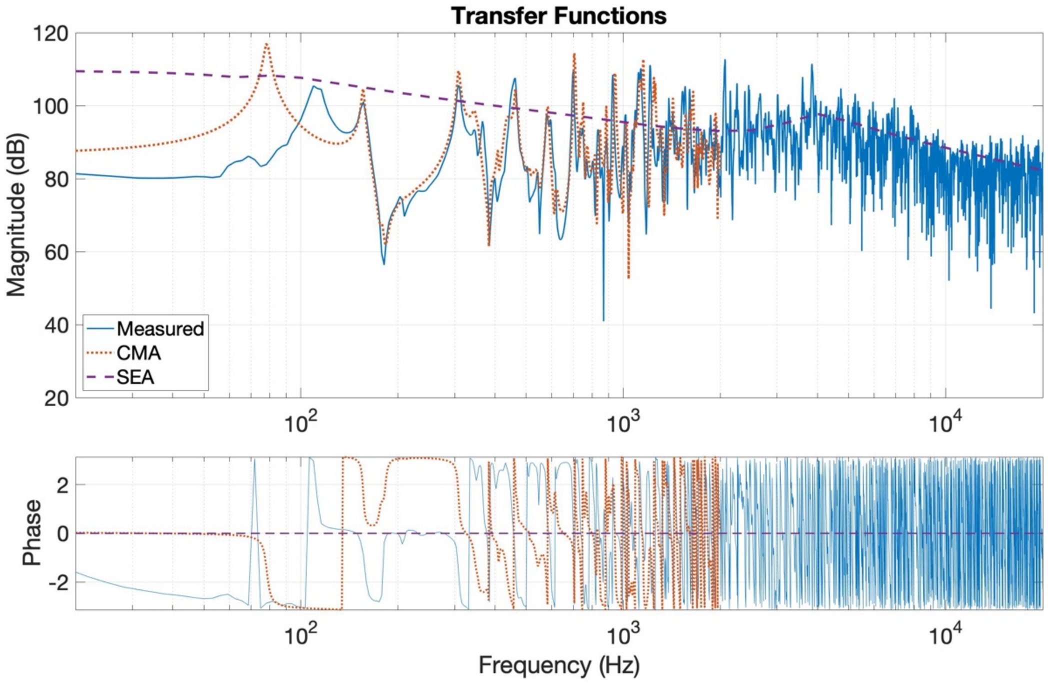

Figure 2-4. *Classical modal analysis (CMA) response and statistical energy analysis (SEA) response compared to measured response of the experimental setup. The transfer functions go from the force on the plate to the microphone in the cavity.*

The individual CMA and SEA responses were combined using the hybrid method introduced in Section 2.1. The result from the hybrid model is compared to a measurement of the experimental setup in Fig. 2-5. The pictured response is the transfer function from the input force on the plate to the microphone in the acoustic cavity. These transfer functions were used to calculate impulse responses, which were convolved with various excitation signals (recordings of engine noise and other sounds of interest from heavy machinery). Listening to these auralizations is the main way that the validity of the approach was evaluated. However, presenting audio recordings is not possible in a written format, so the frequency responses will be discussed.

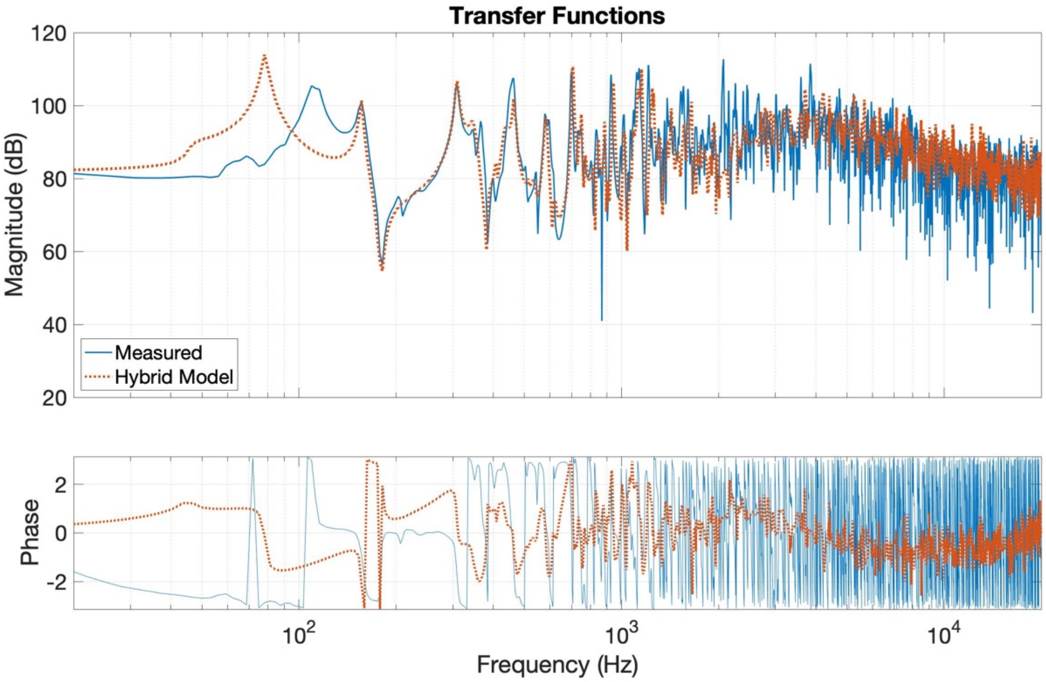

Figure 2-5. *Full hybrid model result compared to experimental measurement. The transfer functions go from the force on the plate to the microphone in the cavity.*

One of the most notable features in the frequency response is the mismatch in the frequencies of the lowest peak between the hybrid model and the measurement. The model predicts a peak at 78 Hz while the measurement showed a peak at 109 Hz. This was concerning and somewhat perplexing considering how well the two match after the second peak at 156 Hz. Examining the natural frequencies of the plate and cavity individually, one finds that the first peak corresponds exactly to the 1-1 mode of the plate. This is because the 1-1 plate mode is lower in frequency than any of the modes of the acoustic cavity (the first acoustic cavity mode occurs at 155 Hz), so there is no coupling between structural/acoustic modes in this frequency range. The theoretical natural frequency of the 1-1 mode of a simply supported plate with the given material properties is 78 Hz, matching the hybrid model. The discrepancy with the measured response was reconciled by looking at previous measurements of the physical plate separately, not coupled to the acoustic cavity. Those measurements had revealed that the natural frequencies of the experimental plate closely matched those of a theoretical simply supported plate for all the higher modes, but not for the 1-1 mode. The natural frequency of the 1-1 mode was measured to be 109 Hz, exactly matching the measured resonance in the coupled system. Therefore, the discrepancy did not come from an error in the model, but from the inadequacy of the experimental setup in replicating simply supported boundary conditions at these lower frequencies. It is not surprising that the physical boundary conditions do not match the theoretical ones exactly, and it is worth noting that the developers of the method used to construct the simply supported plate also found the largest percent error with the 1-1 mode [48]. Identifying these results as the source of the discrepancy alleviated concerns and attempting to fix the plate was deemed unnecessary. Applying a high-pass filter at 100 Hz to the auralizations proved sufficient in minimizing the differences caused by these mismatching fundamental frequencies.

As previously stated, a full complex frequency response, including both magnitude and phase information, is necessary to transform to the time domain to obtain an impulse response for auralization. Although both are necessary, the magnitude portion of the responses tends to dominate human perception of sound, while the phase plays a secondary rule. For example, imagine a musical note played from a pair of loudspeakers. Shift the frequency or change the amplitude and people are bound to notice, but shift the phase and there is likely to be no perceptible difference (except in specific circumstances where significant interference occurs). This means that matching the magnitude portion as closely as possible is vital, but finding an appropriate approximation of the phase can be sufficient. By no means does this imply that any old phase will do. Out of the infinite number of possible phases, only a small subset will approximate reality close enough to sound natural. It was previously discussed that many ways of constructing an approximate phase were tested, and nearly all of them introduced undesirable artifacts into the final auralizations. Using a minimum phase was the one method that preserved the naturalness of the auralizations. While most physical systems are not truly minimum phase (though many approximate minimum phase at low frequencies), it has nice properties such as preserving causality and invertibility that allow it to produce auralizations without introducing such artifacts. The minimum phase calculated for the hybrid model shown in Fig. 2-5 is sufficient to create a natural sounding auralization for the system of interest, confirmed via listening tests. To further test the viability of the minimum phase, a new response was created by combining the magnitude of the measured response and the phase of the hybrid model. Auralizations created with this new response were not perceptibly different than those created from the full measured response. This shows that the minimum phase is indeed a good approximation of the real system. This may not be the case for all systems, and further investigation would be appropriate examine the generalizability of using the minimum phase, as well as investigating other methods for calculating an approximate phase.

Many of the simulated auralizations sounded similar to the measurements, although the exact level of similarity was somewhat difficult to assess. Listening tests were conducted to evaluate the similarity, focusing on realism/artificialness. Eleven listeners participated in the listening tests to capture a variety of perceptions and opinions. Two trends became apparent when examining listening test responses. First, the perceived pitch of the sounds was dominated by the peaks with the highest magnitude in the frequency response. This had a significant impact on how similar the simulations were perceived compared to the measurements because pitch is one of the main perceptual traits that people tend to focus on when comparing two sounds.

However, even though the pitch differences significantly affected perception of the overall similarity of the sounds, they did not significantly affect the perceived realism of the sounds. For example, the peak at 2069 Hz in the measured response shown in Fig. 2-5 is notably missing from the response of the hybrid model because a 2000 Hz crossover frequency was used for the model. This caused a significant difference in the pitch of the measured sounds vs. the sounds created from the hybrid model, resulting in lower ratings for overall similarity. Increasing the crossover frequency to 2100 Hz allowed the model to capture the 2069 Hz peak, resulting in noticeably more similar pitches and therefore better ratings of overall similarity. Despite better ratings for overall similarity between the measurement and the model, there was no difference in the perceived realism of the simulated sounds for the 2100 Hz crossover compared to the 2000 Hz crossover. This shows that exactly matching the dominant peaks is not necessary to create realistic simulations, even if pitch differences are introduced.

Second, the excitation signal had a significant impact on whether the sounds were perceived as realistic or artificial. In particular, it was found that sounds created with input signals containing transients were much more likely to be perceived as realistic, while sounds created from entirely steady state input signals were more likely to be perceived as artificial. This was the case across the board, for both measured sounds and simulated sounds; even measured sounds were more likely to be perceived as artificial if the excitation signal contained no transients.

The crossover frequency, or the point at which to switch from a low-frequency method to a high-frequency method, is one of the main considerations in the proposed hybrid method. Low-frequency methods usually provide a more accurate result, so theoretically the crossover frequency should be as high as possible for the best response. However, as mentioned before, low-frequency methods take considerably more computation time. If a simple and efficient model is the goal, the question that naturally arises is, “How low of a crossover frequency is acceptable?” In order to address this question, models were created with crossover frequencies varying from 16 Hz to 2000 Hz at the one-third octave band center frequencies. These models were used to create auralizations which were listened to and rated according to perceived realism/artificialness. Although there was some variation, and the perceived realism of the sounds depended on the excitation signal as discussed above; it was found that the lowest “good” crossover frequency was about 630 Hz as shown in Fig. 2-6. (Good crossover frequencies were defined based on the ability to produce auralizations that on average sounded more natural than artificial). This does not mean that 630 Hz was necessarily the best or optimal crossover frequency; models with higher crossover frequencies were rated better. Rather, it provides a lower bound for the crossover frequencies that can be used and still retain a sense of realism in the auralizations. The 630 Hz limit is only for this particular setup and would certainly change for other configurations, but it is reasonable to assume that the lower limit would be related in some way to a modal density. The plot in Fig. 2-6 shows that crossing over at a frequency where each component has 3+ modes per one-third octave band preserves the realism of the final auralizations. This seems reasonable since SEA gives a more accurate result when there are multiple modes per frequency band. Therefore, when computation time is a concern, the number of modes per frequency band can be examined for each component of the system and used to determine an appropriate crossover frequency.

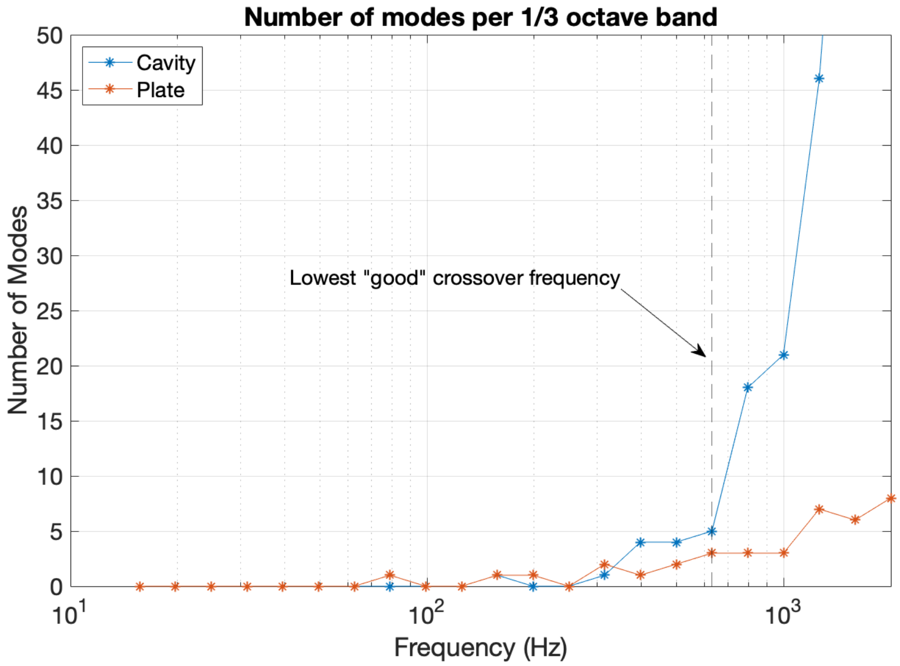

Figure 2-6*. Number of theoretical modes per 1/3 octave band for the simply supported plate and the acoustic cavity. The lowest "good" crossover frequency for the hybrid model is shown to be approximately 630 Hz.*

Chapter 3 Machine Learning Model

This chapter describes a method developed to evaluate the fidelity of the acoustic simulations previously discussed. A machine learning approach is used to match the efficiency of traditional objective measures, while maintaining the accuracy of subjective listening tests. The development section includes a description of the dataset used, a comparison of various machine learning algorithms, and a summary of various audio features considered. The performance of the machine learning model is then examined, showing that it is able to accurately predict human perceptions of audio fidelity. The method is then tested on an additional application, predicting human perceptions of loudspeaker sound quality. The results achieved in this additional application further validate the use of a machine learning approach for predicting human perceptions of audio, although limitations are seen.

3.1 Model Development

Using machine learning to assess audio fidelity is a novel application for intelligent sound quality evaluation. This chapter details the development of the novel audio fidelity model. An overview of the developed model is shown in Fig. 3-1. Each step is discussed in the following sections. First, the audio clips making up the dataset are described, along with details about how each clip was classified. Second, various machine learning algorithms are discussed and compared to find one suitable for the problem at hand. Third, numerous audio features are considered to determine the optimal set for implementation. Fourth, the performance of the model is presented and analyzed. Finally, an additional application is shown to further validate the methods developed.

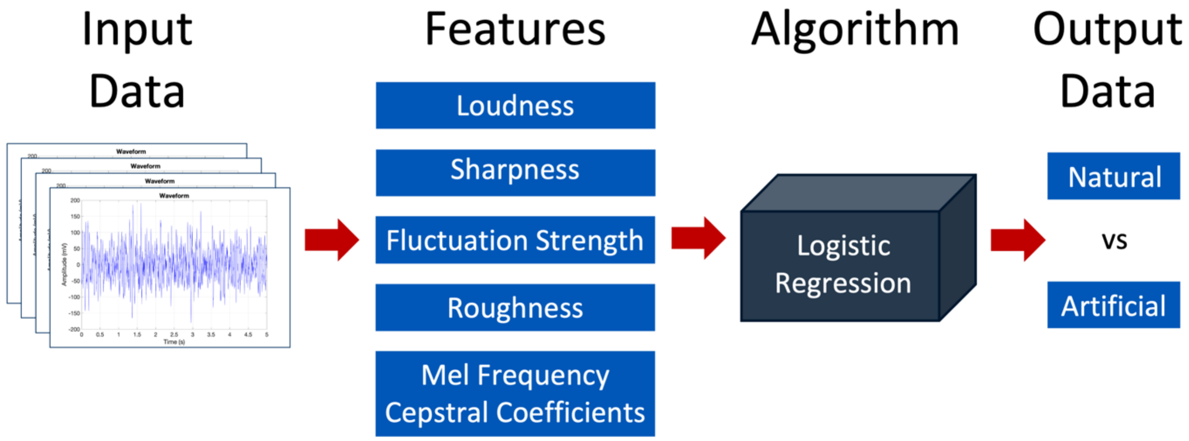

Figure 3-1. *Diagram of the model developed to evaluate the fidelity of simulated audio.*

3.1.1 Dataset

The machine learning model was developed in parallel with the acoustic simulations, so there was no library of sounds initially. It was necessary to build a dataset of sounds to develop and test the model. The primary concern when evaluating the quality of the simulations was that they would sound artificial rather than natural. Therefore, the goal was to build a model that could distinguish between artificial and natural sounding audio clips. It was found that audio compression algorithms sometimes produced a similar artificial sound as was heard in early simulations. A dataset was then created from a library of recordings of machinery, applying a discrete cosine transform (DCT) based audio compression algorithm [49] with 5 distinct levels of compression: keeping 99.99%, 99.90%, 99.0%, 95.0%, and 90.0% percent of the original energy. The uncompressed audio tracks were also included, resulting in a dataset of 618 audio clips.

The machine learning model was intended to be able to learn and predict human responses. This indicates a supervised learning approach, where the model is trained on examples of input-output pairs, as opposed to unsupervised learning, where no output is specified during training and the model determines patterns on its own. The library of compressed audio clips made up the inputs, and the desired output would be some sort of "artificialness" rating for each clip. These ratings could either be on a scale, resulting in a regression problem, or the ratings could be nominal groups, resulting in a classification problem. Although the perceived artificialness of sounds is certainly a spectrum, a binary classification, "artificial" vs. "natural", was decided upon to simplify the rating process for listeners and since it matched well with the overall goals of the project. Each audio clip was then listened to and assigned one of these two labels, resulting in 354 "artificial" sounding clips and 264 "natural" sounding clips. Only a single researcher listened to the audio clips in order to streamline the listening process. This was deemed sufficient for the initial dataset of compressed audio clips because this dataset was only intended to aid in development and prove the viability of predicting human perceptions of audio.

Once the simulation methods had been further developed, a library of 160 simulated audio clips was created. Each of these audio clips was similarly listened to and classified as either artificial or natural. In this case, a group of 11 listeners was used to better capture the perceptions of a variety of listeners, and the final label was determined by majority. The model development described below was conducted using the initial dataset of compressed audio clips. The model was then retrained and retested for the new dataset of simulated audio clips. Final results are presented for both datasets.

3.1.2 Initial Algorithm Comparison

There are many types of machine learning algorithms and model structures used for audio data, with recurrent neural networks (RNN), convolutional neural networks (CNN), and support vector machines (SVM) being some of the most common [50, 51, 52]. Recurrent neural networks and convolutional neural networks are both deep learning algorithms, a subfield of machine learning that uses large neural networks that were originally designed to mimic the human brain. While deep learning is certainly a broad and complex field, a simple description captures the essence of deep learning: extremely large amounts of data can lead to exceptional results, and generally, results continue to improve as more data is added [53]. In addition to a high level of accuracy, another benefit of deep learning is that it often does not require manual feature extraction. Normally, important features of the raw data are extracted before being input into a machine learning algorithm (for example, the overall sound pressure level might be one important feature of an audio recording). The features to use are determined by an engineer/researcher depending on what characteristics might be important for the given application. When done well, this makes the task of the machine much easier and simple algorithms can produce accurate results. However, human deficiency in choosing appropriate features can be detrimental. Deep learning methods can remove the need for feature extraction and raw data can be used as the input; essentially the machine does its own feature extraction. This can lead to even better results and has proven exceptionally successful in many audio and image data applications [54, 55, 56]. Unfortunately, it is often only effective when large amounts of data are available. If few data are available to train the machine, then it is not able to learn which features are important in the raw data. This is one of the reasons that machine learning has become thought of as a solution for “big data” problems.

Older machine learning algorithms such as logistic regression (LR), support vector machines (SVM), k-nearest neighbors (KNN), and random forests (RF) don’t do well extracting important features from raw data. However, if appropriate features are extracted beforehand, they can produce accurate results with limited amounts of data. For this reason, these algorithms were the focus when developing the machine learning model for this research. A dataset of only

618 samples is rather small in the context of machine learning. Each of these algorithms was tested throughout the development process to see if any outperformed the others. A final comparison of these algorithms is provided in Section 3.1.4.

3.1.3 Feature Extraction

There are numerous methods for audio feature extraction, most of which are based on established signal processing techniques [59]. Theoretically, any characteristic of the audio signal could be used as a feature, making a comprehensive list of potential features near impossible. Some of the most common features used for audio in machine learning were considered for the research in this paper, along with a few that are not as common in machine learning but that are frequently used in sound quality assessment. A complete list of these features, including short descriptions of each, is provided below:

1) Mel-frequency cepstral coefficients (MFCCs) are likely the most commonly used audio feature in machine learning [60]. They have proven particularly useful in areas such as speech recognition and are integral parts of the algorithms that allow lingual interaction with modern cell phones and smart devices [61, 62, 63]. Mel-frequency cepstral coefficients are best understood by examining the steps used to calculate them. First, a spectrogram of the audio data is calculated. Second, the frequency axis is adjusted according to the Mel scale, which represents the frequency dependence of human hearing. Third, the discrete cosine transform is performed over frequency to decorrelate the data and allow for compression. Finally, only the first approximately 13 coefficients are retained while the others are discarded. More detailed descriptions of this process are provided in the References [61, 64, 65].

2) Delta and delta-delta are approximations of the first and second derivatives of a given metric. Delta and delta-delta cepstral coefficients are evaluated based on MFCCs and are used to capture the dynamic nature of a signal that isn't fully captured in MFCCs alone. They often improve the performance of models using MFCCs [66]. A delta coefficient is defined as,

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta (c_{t+\theta} - c_{t-\theta})}{2 \sum_{\theta=1}^{\Theta} \theta^2}, \tag{3.1}$$

where $c_{t+\theta}$ is the static coefficient (MFCC) at time $t + \theta$ and Θ is the size of the delta window. Delta-delta would be calculated from the same formula but by replacing the static coefficients with the corresponding deltas.

3) The spectral centroid is essentially the "center of mass" of the power spectrum. It is a weighted average of the energy distribution across frequency. It is commonly used to measure the brightness of a sound [65]. It is defined as,

$$\mathrm{SC} = \frac{\sum_{k=0}^{N_{FT}/2} f(k) P_s(k)}{\sum_{k=0}^{N_{FT}/2} P_s(k)} \tag{3.2}$$

where $f(k)$ is the frequency of the k^{th} bin, P_s is the estimated power spectrum, and N_{FT} is the size of the discrete Fourier transform (DFT).

4) Zero-crossing rate is a count of the number of times a waveform crosses zero in a given time interval. The count is normalized by the length of the input signal. It is particularly useful for identifying percussive and fricative sounds, commonly used to aid in classifying voiced/unvoiced speech as well as music genres [65]. It is defined as,

$$\mathrm{ZCR} = \frac{1}{2}\left(\sum_{n=1}^{N-1}\left|\mathrm{sign}\big(s(n)\big) - \mathrm{sign}(s(n-1))\right|\right)\frac{F_s}{N} \tag{3.3}$$

where $s(n)$ is the signal, N is the number of samples in the signal, F_s is the sampling frequency, and $\mathrm{sign}(x)$ is defined as,

$$\mathrm{sign}(x) = \begin{cases} 1 & if\ x > 0 \\ 0 & if\ x = 0 \\ -1 & if\ x < 0 \end{cases} \tag{3.4}$$

5) Spectral flux is the average variation of the signal amplitude spectrum between adjacent frames. This measure of local spectral change is useful for separating speech and music, as well as detecting events in a recording [65]. It is defined as,

$$\mathrm{SF} = \frac{1}{LN_{FT}}\sum_{l=0}^{L-1}\sum_{k=0}^{N_{FT}-1}[\log(|S_l(k)| + \delta) - \log(|S_{l-1}(k)| + \delta)]^2 \tag{3.5}$$

where $S_l(k)$ is the discrete Fourier transform (DFT) of the l^{th} frame at frequency k , N_{FT} is the order of the DFT, L is the total number of frames in the signal, and δ is a small parameter to avoid calculation overflow.

6) The spectral roll-off point is the frequency below which a specified percentage of the spectral energy lies. A value of 85% is commonly used. This metric helps characterize the overall shape of the spectrum and is useful in speech recognition [65]. It is defined as,

$$\sum_{k=0}^{K_{roll}}|S(k)| = 0.85\sum_{k=0}^{N_{FT}/2}|S(k)| \tag{3.6}$$

where $S(k)$ is the DFT of the signal, N_{FT} is the size of the DFT, and K_{roll} is the frequency bin corresponding to the estimated roll-off frequency.

7) The chroma feature is a vector that characterizes pitch according to the 12-tone equal temperament common in western music. This helps the machine learn how different tones might be perceptually related (for example, a doubling in frequency is perceived as the same note). It is most commonly used in the field of music information retrieval and genre classification. The chroma vector is defined as,

$$\text{Chroma}(n) = \sum_{i=1}^{nPeaks} w(n, f_i) \cdot a_i\,, \qquad n = 1, 2, \ldots, 12 \tag{3.7}$$

where $w(n, f_i)$ is the weight of the signal at frequency f_i for the semitone n, f_i is the frequency of the i^{th} peak of the signal, and a_i is the amplitude of the i^{th} peak of the signal. Calculation of the weight $w(n, f_i)$ is described by Shi, Li, and Tian [67].

8) Short-time energy is a representation of the energy in a time-varying signal. A window is applied to examine the energy in frames rather than averaging across the entire signal [68]. It is useful in classifying dynamic and transient signals. Short-time energy is defined as,

$$E_n = \sum_{m=n-N+1}^{n} [s(m)w(n-m)]^2 \tag{3.8}$$

where $s(m)$ is the signal, $w(n-m)$ is the window, n is the sample that the analysis window is centered on, and N is the window length.

9) Loudness is a psychoacoustic sound quality metric that is a measure of how loud a sound is as perceived by humans. It is more representative of human perception than purely physical quantities such as sound pressure level or sound intensity level, because it accounts for aspects of human hearing such as frequency dependence and masking. Unfortunately, such adjustments are not simple mathematical equations, and the process

to calculate loudness (as well as the other psychoacoustic sound quality metrics sharpness, roughness, and fluctuation strength) is quite involved. There are several methods for modeling loudness, none of which is universally accepted. This paper uses the Zwicker method for calculating the loudness of time-variant sound as outlined in DIN 45631 [69]. More information about loudness and other psychoacoustic sound quality metrics can be found in Reference [70].

10) Sharpness is a psychoacoustic sound quality metric that relates the amount of high-frequency energy to total energy in a sound. Sharpness is one of the most significant indicators of the pleasantness of a sound; less sharp sounds are generally perceived as more pleasant [70]. It is somewhat similar to the spectral centroid and spectral roll-off point in that it gives a general representation of the relative frequency content of a sound, but it accounts for the frequency dependence of human hearing and particularly identifies the presence of disagreeable high-frequency content.

11) Roughness and fluctuation strength are psychoacoustic sound quality metrics that characterize the presence of amplitude modulation in a sound. Auditory perception of amplitude modulation is divided into three regions dependent on the frequency of the modulation. Below about a 15 Hz modulation frequency, the amplitude modulations are perceived as a slow up and down change in loudness. This sensation is referred to as fluctuation and the sensation is maximum at a modulation frequency of 4 Hz. Above 15 Hz, the modulation is no longer perceived as slow changes in loudness, and a new sensation of roughness is introduced. The roughness sensation reaches its maximum at a modulation frequency of about 70 Hz and lasts up to about 300 Hz. Above 300 Hz, the

different amplitudes are likely to be perceived as separate sounds. Methods to quantify the strength of fluctuation and roughness in a sound are described in Reference [70].

A backward elimination method was used to determine a suitable subset of these features for implementation. Backward elimination is a common variable selection process used in both statistics and machine learning [71]. The process involves creating a model including all possible explanatory variables (features), then removing the least significant variable and checking to see if model performance degrades. This process continues iteratively until no more variables can be removed without adversely affecting model performance. It was found that a combination of MFCCs, loudness, sharpness, roughness, and fluctuation strength performed the best, achieving a high accuracy with a relatively low number of parameters. Interestingly, the features that performed well all account for human hearing in some way, while the features that did not perform as well are merely characteristics of the signal itself and do not account for human hearing (except for perhaps the chroma feature, which doesn't necessarily account for human hearing, but does account for tonal/pitch association that may be caused by a familiarity with western music).

Initially, the overall loudness, sharpness, roughness, and fluctuation strength values, along with the time averaged MFCC values, were extracted from each audio clip and input to the model. A 10-fold cross validation was used to test the performance of each algorithm, using these features as inputs. Cross validation is a process by which the data are split into k separate groups (in this case 10). The model is then trained on k-1 of those groups and tested on the final group. This is repeated for every combination of training/testing sets. It accounts for the possibility of testing the model on unrepresentative subsets of the data and has become the "gold standard" for testing in machine learning [57]. In the case of the dataset consisting of compressed audio tracks, the 618 tracks were split into 10 nominally equal sized groups, containing either 61 or 62 audio tracks. The algorithms were each then trained on 9 of those groups (~556 tracks) and tested on the 10th group (~62 tracks). The accuracy for the test group, or the number of correct predictions out of 62, was then recorded for each algorithm. This process was repeated, training on 9 different groups and testing on the 10th, for all possible combinations of the groups. This resulted in 10 distinct tests, each with a recorded accuracy for each algorithm. The best results achieved using this method amounted to an 87% mean accuracy averaging across the 10 tests, with a standard deviation of 16%.

To improve these results, a reference audio clip was introduced. This reference was a real life (uncompressed) audio recording of the system to be simulated. The features were then calculated as the difference between the original and the reference. Therefore, a positive loudness input would indicate that the audio clip to be evaluated was louder than the reference, a negative loudness input would indicate that it was less loud, and a zero loudness input would indicate the same loudness, with similar interpretations for each of the other audio features. Scatterplots of these features, calculated from each of the 618 audio tracks, are plotted against each other in Fig. 3-2 and Fig. 3-3. Examining scatterplots such as these provides a glimpse of the distribution of the data in the high dimensional feature space. Of particular interest is comparing the distributions of the distinct classes (marked by different colors on the plots). Visualizing these distributions is useful for a few reasons. First, the overall amount of overlap between the classes indicates the potential ability of the features to discriminate between the classes. If there is too much overlap between the classes, then the features may not be sufficient to discriminate between the classes and yield accurate predictions. If there is little to no overlap, then a smaller subset of the feature set may suffice, allowing for a simpler model. The features shown do start to separate the two classes, but there are no pairs of features that separate the classes completely. Second, the scatterplots give an idea of which features are most significant. The plots show that loudness, sharpness, and MFCC0 have the most significant ability to separate the two classes, whereas roughness, fluctuation, and the higher order MFCCs do not separate the classes as much (though removing these less significant features from the model does degrade performance, so they are still providing useful additional information not captured by the more significant features). Third, the scatterplots can provide additional insight into which machine learning algorithms might be most appropriate. The two classes appear to be linearly separable for the most part, especially with respect to the most significant features. This impacts which of the machine learning algorithms perform the best, as seen in Section 3.1.4.

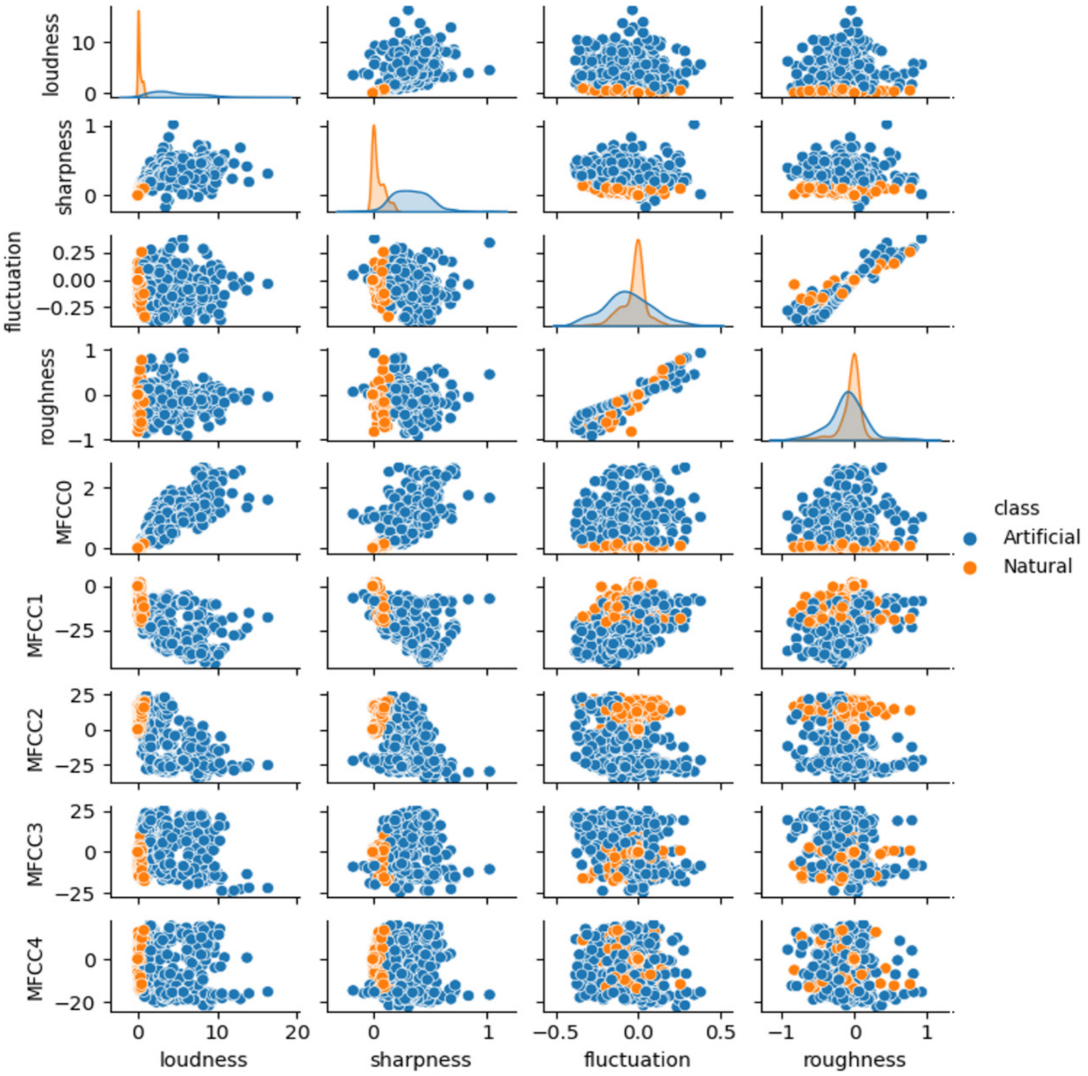

Figure 3-2. *First half of a matrix of scatterplots showing features plotted against each other, with distributions of each feature on the diagonal. A total of 13 MFCCs were used, but only the first (most significant) 5 MFCCs are shown.*

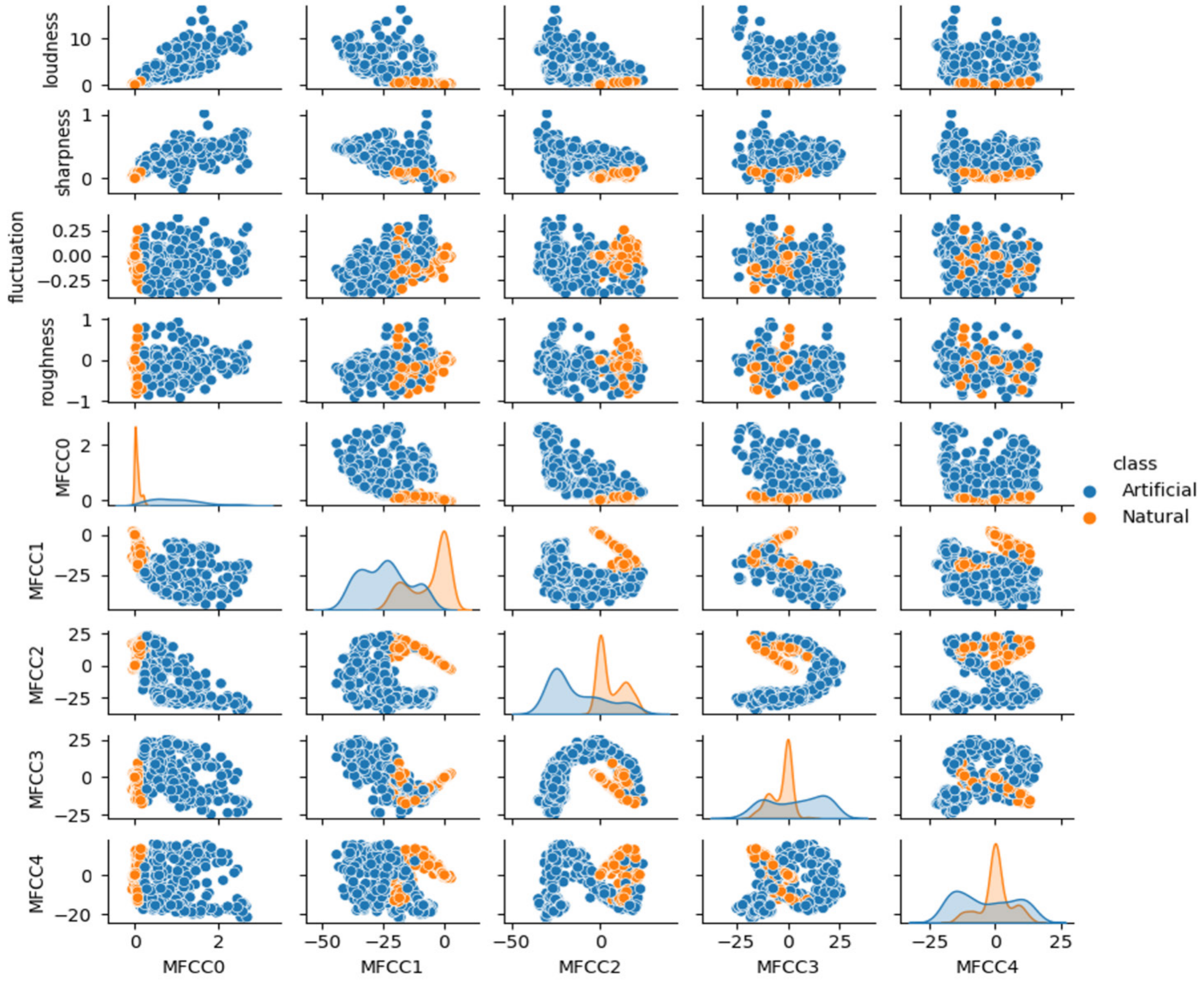

Figure 3-3. *Second half of a matrix of scatterplots showing features plotted against each other, with distributions of each feature on the diagonal. A total of 13 MFCCs were used, but only the first (most significant) 5 MFCCs are shown.*

Much better results were achieved using the reference method, as discussed in the following sections. Of course, the need for a reference signal does limit the applicability of the model. However, such a model can still be useful, essentially predicting how similar two sounds will be perceived by human listeners. In the context of this project, using the reference method still achieved the goal of providing a quick and efficient way to determine whether simulations were realistic, and it is later shown how such a model can be useful in other applications such as the evaluation of loudspeaker sound quality.

3.1.4 Algorithm Selection

After extracting appropriate features, it is necessary to revisit the machine learning algorithms to determine which is best for implementation. As discussed in Section 3.1.2, logistic regression, support vector machines, k-nearest neighbors, and random forests were each tested throughout development to monitor performance. A final comparison of the accuracy of these algorithms, using the features and reference method discussed in Section 3.1.3, is shown in Fig. 3-2. The accuracies were obtained through a 10-fold cross validation, resulting in 10 distinct tests, each with a recorded accuracy for each algorithm. The distributions of the accuracies from these 10 tests are represented in the box plots in Fig. 3-4. It was found that logistic regression performed the best for the binary classification problem, with 9 out of 10 accuracies exceeding 98%, a median accuracy of 100%, and only one test yielding an accuracy of about 94.5%. This is not surprising since logistic regression fits to the sigmoid function $\mathrm{sig}(x) = \frac{1}{1+e^{-x}}$ which is bounded between 0 and 1 and is ideal for binary classification. The other algorithms had larger spreads, with accuracy percentages extending down into the lower 90s, and median accuracies of 98%. Logistic regression was chosen for implementation in the fidelity prediction model because of its better performance, along with the added benefits of simplicity, ease of interpreting parameter estimates, and it appeared to have the least tendency to overfit. Support vector machines (SVM) proved more appropriate for the additional application discussed later in Section 3.3, where a binary classification was no longer used. Unlike logistic regression, SVM classifiers do not fit to a specific function, but rather they attempt to find optimal hyperplanes to separate the classes [58]. These hyperplanes are similar to two-dimensional planes that bisect a three-dimensional space, but they are separating an N-dimensional space, where N is the number of features or inputs into the algorithm.

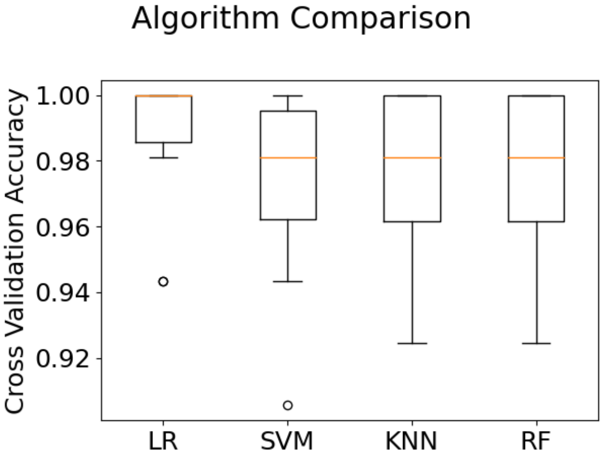

Figure 3-4. *Box plots comparing a 10-fold cross validation accuracy of logistic regression (LR), support vector machines (SVM), k-nearest neighbors (KNN), and random forests (RF). The box extends from the lower to upper quartile values of the data, with a line at the median. The whiskers extend to the minimum and maximum values, excluding the outliers marked as circles (an outlier is defined if its distance from the box is more than 1.5 times the total width of the box).*

3.2 Results

The model was initially tested on the dataset consisting of 618 compressed audio clips. This dataset was split into two groups, one containing 528 audio clips that was used to test the model throughout development, and one containing 90 audio clips that was kept separate to validate the model after development and to check for overfitting. The second group was made up of audio clips from completely different machines (for example a grader vs. a bulldozer) than the first group to best test the generalizability of the model. For the first group of audio clips, the final model (which used the features and reference method described in Section 3.1.3, along with a logistic regression algorithm) achieved 98.3% mean accuracy averaging across a 10-fold cross validation, with a standard deviation of 1.9%. The model was then trained on the entire first group of 528 audio clips and tested on the second group of 90 audio clips, resulting in a 96.7% accuracy. This provides evidence that the model is not overfitting since it was still able to perform well on completely new data. The accuracy from the second group is slightly lower than the mean accuracy from the cross validation, but it is within one standard deviation, so the difference is not significant enough to cause concerns about overfitting. Details about the predictions made on the second group are provided in the confusion matrix in Fig. 3-5 and the classification report in Table 3-1.

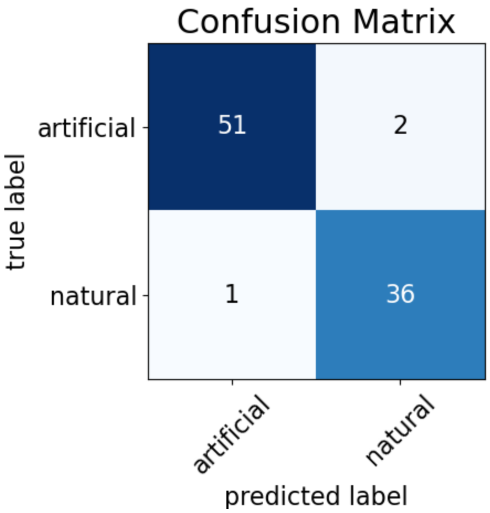

Figure 3-5. *Confusion matrix for the group of 90 audio clips used to test for overfitting.*

Table 3-1. *Classification report for the group of 90 audio clips used to test for overfitting.*

	PRECISION	RECALL	F1 SCORE	SUPPORT
ARTIFICIAL	0.981	0.962	0.971	53
NATURAL	0.947	0.973	0.960	37
ACCURACY			0.967	90

The confusion matrix is a common way of visualizing classification results. The rows represent the true labels, in this case the labels given by the participants in the listening tests. The columns represent the predicted labels, or the labels predicted by the model. The diagonal represents all of the correct predictions, and anything off of the diagonal represents an incorrect prediction. The confusion matrix provides a quick and easy way to see how many of the model's predictions are correct/incorrect for each class. It is rather simple for a binary classification but can be used for problems with any number of classes. In addition to visualizing final results, it can be useful during model development to show where incorrect classifications are occurring (maybe data points from class 1 are consistently misclassified as belonging to class 3), suggesting where further investigation and adjustments may be needed.

The classification report provides four key metrics for each class: precision, recall, F1 score, and support. Precision, recall, and F1 score can take multiple examples to fully understand. A good additional explanation is given by Koehrsen [72]. Precision is defined by

$$\text{precision} = \frac{\#\text{ true positives}}{\#\text{ true positives} + \#\text{ false positives}} \tag{3.9}$$

True positives for a given class are data points that are correctly predicted to be in the class. False positives for a given class are data points incorrectly predicted to be in the class. Therefore, precision is the percentage of data points predicted to be in a specific class that actually belong to that class. This is best understood through an example. The precision for the "artificial" class is given by

$$\text{precision(artificial)} = \frac{\#\text{ artificial sounds predicted to be artificial}}{\#\text{ artificial sounds predicted to be artificial} + \#\text{ natural sounds predicted to be artificial}} \tag{3.10}$$

The true positives for the “artificial” class are data points that the model correctly predicts as “artificial”, the false positives for the “artificial” class are the data points that the model incorrectly predicts to be “artificial”, and the precision for the “artificial” class is the percentage correct of the total number predicted to be “artificial”. Comparing Eq. (3.10) to the confusion matrix and the classifications report, one can see that 51 tracks were correctly predicted to be artificial and 1 track was incorrectly predicted to be artificial, leading to a precision of 0.981 for the artificial class.

$$\text{precision(artificial)} = \frac{51}{51+1} = 0.981 \tag{3.11}$$

The second metric in the classification table is recall. Recall is defined as

$$\text{recall} = \frac{\text{\# true positives}}{\text{\# true positives} + \text{\# false negatives}} \tag{3.12}$$

True positives for a given class are the same as described above. False negatives for a given class are data points incorrectly predicted not to be in the class. Therefore, recall is the percentage of data points belonging to a specific class that are correctly predicted by the model. This is best understood through an example. The recall for the “artificial” class is

$$\text{recall(artificial)} = \frac{\text{\# artificial sounds predicted to be artificial}}{\text{\# artificial sounds predicted to be artificial} + \text{\# artificial sounds predicted to be natural}} \tag{3.13}$$

The true positives for the “artificial” class are data points that the model correctly predicts as “artificial”, the false negatives for the “artificial” class are the data points that the model incorrectly predicts to be “natural” (not “artificial”), and the recall for the “artificial” class is the percentage of “artificial” sounds that the model correctly predicts. Comparing Eq. (3.13) to the confusion matrix and the classifications report, one can see that 51 tracks were correctly predicted to be artificial and 2 tracks were incorrectly predicted to be natural, leading to a recall of 0.962 for the artificial class.

$$\text{recall(artificial)} = \frac{51}{51 + 2} = 0.962 \tag{3.14}$$

Sometimes the precision of a specific class is most important: perhaps it is imperative that all sounds predicted to be “natural” are indeed “natural” sounding, but it doesn’t matter as much if some “natural” sounds are incorrectly predicted to be “artificial”. Other times, the recall of a class is most importation: perhaps it is vital that all “natural” sounds are predicted to be “natural”, but it doesn’t matter as much if some “artificial” sounds are incorrectly predicted to be “natural”. However, often both precision and recall are equally important, in which case both metrics are combined into the F1 score. The F1 score is the harmonic mean of precision and recall, defined by

$$\mathrm{F}_1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}, \tag{3.15}$$

and is used to determine the balance between the precision and the recall. The harmonic mean is calculated by dividing the number of data points by the reciprocal of each number in the series. It is used in this case because it minimizes the effect of large outliers, which is not the case with an arithmetic mean. The F1 score is particularly useful when there are not equal numbers of data points in each class, because it penalizes the model if it merely predicts the majority class. For example, in the set of 90 audio clips, 53 were labeled as artificial and 37 were labeled as natural. This means that the model could achieve 53/90 = 59% accuracy just by predicting every sound to be artificial. However, such predictions would give an F1 score of zero, showing that the model is not actually providing useful predictions. This becomes even more important when the class distribution is significantly uneven, and high accuracies could be obtained just by predicting the majority class. Table 1 shows F1 scores that are not significantly different than the overall accuracy, indicating that this is not a problem for the model developed here.

The final two metrics provided in the classification report are an overall accuracy and the support in the final column. The overall accuracy is simply the percent of correct predictions for all classes out of the total number of data points. The support is the number of data points belonging to each class: 53 artificial and 37 natural for a total of 90 data points.

Finally, the model was retrained and tested on the dataset of 160 simulated audio clips. The model was retrained because the initial model trained on the compressed audio clips was never intended to generalize to the simulated audio clips. Rather, it was meant to aid in development and prove the viability of predicting human perceptions of audio. Testing on the dataset of simulated audio clips, the retrained model achieved 98.1% mean accuracy averaging across a 10-fold cross validation, with a standard deviation of 2.9%. This is on par with the performance seen with the first dataset. These results are important for two reasons. First, it shows that the model is capable of predicting human perception of the fidelity of simulated sounds. Second, it shows that the model is able to perform well under more difficult circumstances. With the initial dataset, the model had a much easier job. Essentially, less compressed audio sounds more natural and more compressed audio sounds more artificial, so the model simply had to find the cutoff. The simulated dataset presented a more difficult challenge because there was no obvious indicator about what made the final audio clips natural or artificial sounding.

3.3 Additional Application

A similar model was constructed and tested on an additional application. This application involved evaluating and comparing the sound quality of loudspeakers. One common method to evaluate loudspeaker sound quality is to conduct A/B listening tests, where subjects will blindly listen to two loudspeakers and record which they prefer, A or B. Normally this would be done in person, but due to the COVID-19 outbreak at the time of testing, many listening tests were conducted remotely via high quality audio recordings and calibrated headphones. This allowed access to a library of listening tests where both the listener responses and the audio signals were available.

The listening tests used in this research asked participants three questions: Do you prefer loudspeaker A or loudspeaker B for bass? Do you prefer loudspeaker A or loudspeaker B for clarity? Do you prefer loudspeaker A or loudspeaker B for overall sound quality? The outcome in each of these three categories was recorded for each listening test as “A” if there was a clear preference among listeners for loudspeaker A, “B” if there was a clear preference for loudspeaker B, and “N” if there was no statistically significant preference for either (based on a 90% confidence interval).

Three separate models (one for each of the categories bass, clarity, and overall sound quality) were then trained to predict the outcome of the listening tests when fed the audio recordings of the two loudspeakers. The models used the same features as discussed above, but a support vector machine (SVM) algorithm was used, proving more effective for this trinary classification. The models were tested via a cross validation process. To help prove generalizability and mimic real world use cases, the cross-validation groups were split by separate listening tests rather than randomly splitting the groups.

The models were trained and tested on two distinct datasets: one where the loudspeakers were noticeably different, and one where the loudspeakers were similar enough that only trained listeners could easily distinguish between them. The bass, clarity, and overall sound quality models were able to achieve mean accuracies of 93.6%, 95.0%, and 95.5% respectively on the first dataset, but only mean accuracies of 51.8%, 40.9%, and 35.5% respectively on the latter dataset. These results indicate that the model likely fit to the more significant factors that influence perception of loudspeaker sound quality, but failed to capture the more subtle influences. This allowed the model to accurately predict preferences of noticeably different loudspeakers, but performance degraded with very similar loudspeakers where preferences are entirely determined by more subtle factors.

Chapter 4 Conclusions

Two separate methods have been presented, both aiming to aid in the creation of realistic acoustic simulations. The first is a hybrid method that allows separate low- and high-frequency acoustic responses to be combined into a single broadband response suitable for auralization. The second is a machine learning approach for predicting human perceptions of audio fidelity and sound quality. This chapter analyses the implications of the results achieved for each method and suggests directions for further work.

4.1 Analysis and Conclusions

The proposed hybrid method successfully merged a low-frequency response and a high-frequency response into a single response that could be auralized. This approach allowed for a simple and efficient approximation of the desired acoustic response over a broad frequency range. The auralizations were able to retain a sense of realism, skirting some of the unnatural artifacts prevalent in audio simulation. However, some limiting factors impacted the level of realism achieved. First, matching the largest peaks in the frequency responses is necessary to create the same pitch, and it was found that pitch differences are a significant factor when listeners compare the similarity of two sounds (simulation vs. measurement). Second, the presence/absence of transients in the excitation signal significantly affected the perceived realism of the final auralization. While important considerations, neither of these two limiting factors are directly related to the ability of the hybrid method to produce realistic auralizations. Early on in development, artifacts introduced by the interpolation method or when calculating the approximate phase tended to dominate the perception of the sounds. The final method appears to have overcome these challenges, and they are no longer limiting factors in the achievable realism of the final result

The second part of the thesis evaluated the possibility of training a machine learning model to accurately predict human perceptions of audio fidelity and sound quality. The models were able to achieve ~98% accuracy when evaluating the fidelity of both compressed audio and simulated audio. The models were able to generalize to new data and retain a high level of accuracy. The efficacy of the developed model structure was tested on an additional application evaluating loudspeaker sound quality. Promising results of ~95% accuracy were seen when evaluating noticeably different loudspeakers, but performance declined significantly when evaluating very similar loudspeakers. This study reveals that the current method is not a universal solution and performance may depend on the specific situation. Despite this limitation, the current machine learning approach has proven highly accurate in multiple situations, displaying promise in its ability to supplant human listening tests. This potential has significant time and money saving implications for industries where listening tests are regularly conducted to evaluate audio quality.

4.2 Directions for Further Work

The hybrid method for acoustic simulations was only compared to a single experimental setup. However, some uncertainty exists about how well some of the assumptions will generalize to other situations. For example, minimum phase was used as the approximate phase for the final response. This assumption proved appropriate for the given setup, yielding no perceptible difference from the measured phase, but this may or may not be the case for other setups. Real world systems vary in their resemblance to minimum phase. Testing additional systems would help establish how well the minimum phase assumption generalizes.

Another assumption was that the modes of the overall system are dominated by the modes of the structural components at higher frequencies. This assumption was used when adding amplitude modulation to the high-frequency portion of the responses. The modal density of the plate was used to determine the density of the amplitude fluctuations. This choice proved appropriate for the given plate-cavity system, but further investigation would be required to determine if the assumption holds in general.

Overfitting is always a major concern in machine learning applications. Using the cross-validation process helps protect against overfitting; however, it is not infallible. Potential limitations include the small quantity of data available and that the listening tests were conducted with few participants. Additional testing with a dataset of new listening tests, preferably with new subjects, would confirm how well the models generalize in predicting human preferences.

It was also seen that the performance of the method can be situational, with poor results on the dataset comparing very similar loudspeakers. Such a limitation might be mitigated through additional feature extraction not discussed in this paper, and/or through the implementation of a deep learning architecture if enough data were available to merit such an approach.

References

[1] F. Fahy and D. Thompson, Fundamentals of Sound and Vibration, Boca Raton: CRC Press, 2015.

[2] J. B. Fahnline and G. H. Koopmann, "A lumped parameter model for the acoustic power output from a vibrating structure," *The Journal of the Acoustical Society of America,* vol. 100, no. 6, pp. 3539-3547, 1996.

[3] D. Karnopp, "Lumped parameter models of acoustic filters using normal modes and bond graphs," *Journal of Sound and Vibration,* vol. 42, no. 4, pp. 437-446, 1975.

[4] L. L. Beranek and T. J. Mellow, Acoustics: Sound Fields and Transducers, Waltham, MA: Academic Press, 2012.

[5] H. A. C. Tilmans, "Equivalent circuit representation of electromechanical transducers: I. Lumped-parameter systems," *Journal of Micromechanics and Microengineering,* vol. 6, no. 1, pp. 157-176, 1996.

[6] A. D. Pierce, Acoustics: An Introduction to Its Principles and Applications, Springer International Publishing, 2019.

[7] S. S. Rao, Mechanical Vibrations, Hoboken, NJ: Pearson Education, Inc., 2017, pp. 1013-1045.

[8] N. S. Gokhale, S. S. Deshpande, S. V. Bedekar and A. N. Thite, Practicle Finite Element Analysis, Maharashtra, India: Finite to Infinite, 2008.

[9] R. Z. Gan, B. Feng and Q. Sun, "Three-Dimensional Finite Element Modeling of Human Ear for Sound Transmission," *Annals of Biomedical Engineering,* vol. 32, p. 847–859, 2004.

[10] D. J. Nefske, J. A. Wolf and L. J. Howell, "Structural-acoustic finite element analysis of the automobile passenger compartment: A review of current practice," *Journal of Sound and Vibration,* vol. 80, no. 2, pp. 247-266, 1982.

[11] G. Everstine, "Finite element formulatons of structural acoustics problems," *Computers & Structures,* vol. 65, no. 3, pp. 307-321, 1997.

[12] S. Kirkup, "The Boundary Element Method in Acoustics: A Survey," *Applied Sciences,* vol. 9, no. 8, 2019.

[13] J. T. Katsikadelis, Boundary Elements Theory and Applications, Amsterdam: Elsevier, 2002.

[14] R. H. Lyon and R. G. DeJong, Theory and Application of Statistical Energy Analysis, Newton, MA: Butterworth-Heinemann, 1995.

[15] C. B. Burroughs, R. W. Fischer and F. R. Kern, "An introduction to statistical energy analysis," *The Journal of the Acoustical Society of America,* vol. 101, no. 4, pp. 1779-1789, 1997.

[16] A. J. Price and M. J. Crocker, "Sound Transmission through Double Panels Using Statistical Energy Analysis," *The Journal of the Acoustical Society of America,* vol. 47, no. 3, pp. 683-693, 1970.

[17] Q. Chen, Q. Fei, S. Wu and Y. Li, "Statistical Energy Analysis for the Vibro-Acoustic System with Interval Parameters," *Journal of Aircraft,* vol. 56, no. 5, pp. 1869-1879, 2019.

[18] R. D. Belvins, "Modal density of rectangular volumes, areas, and lines," *The Journal of the Acoustical Society of America,* vol. 119, no. 2, pp. 788-791, 2006.

[19] T. Letowski, "Sound quality assessment: concepts and criteria," *Journal of the audio engineering society,* 1989.

[20] T. Rydén, "Using Listening Tests to Assess Audio Codecs," *Journal of the Audio Engineering Society,* 1996.

[21] A. Gabrielsson and B. Lindström, "Perceived Sound Quality of High-Fidelity Loudspeakers," *Journal of the Audio Engineering Society,* vol. 33, pp. 33-53, 1985.

[22] F. E. Toole, "Subjective Measurements of Loudspeaker Sound Quality and Listener Performance," *Journal of the Audio Engineering Society,* vol. 33, pp. 2-32, 1985.

[23] D. Campbell, E. Jones and M. Glavin, "Audio quality assessment techniques — A review, and recent developments," *Signal Processing,* vol. 89, no. 8, pp. 1489 - 1500, 2009.

[24] A. Gabrielsson, B. Hagerman, T. Bech-Kristensen and G. Lundberg, "Perceived sound quality of reproductions with different frequency responses and sound levels," *The Journal of the Acoustical Society of America,* vol. 88, no. 3, pp. 1359-1366, 1990.

[25] F. Rumsey, S. Zieliński and R. Kassier, "On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality," *The Journal of the Acoustical Society of America,* vol. 118, no. 2, pp. 968-976, 2005.

[26] S. Bech and N. Zacharov, Perceptual Audio Evaluation –Theory, Method and Application, John Wiley & Sons, 2006.

[27] M. Cartwright, B. Pardo, G. J. Mysore and M. Hoffman, "Fast and easy crowdsourced perceptual audio evaluation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, 2016.

[28] H. B. Huang, X. R. Huang, R. X. Li, T. C. Lim and W. P. Ding, "Sound quality prediction of vehicle interior noise using deep belief networks," *Applied Acoustics,* vol. 113, pp. 149 - 161, 2016.

[29] Y. Fang and T. Zhang, "Sound Quality Investigation and Improvement of an Electric Powertrain for Electric Vehicles," *IEEE Transactions on Industrial Electronics,* vol. 65, no. 2, pp. 1149-1157, 2018.

[30] B. Lopes, C. Colangeli, K. Janssens, A. Mroz and H. Van der Auweraer, "Neural Network Models For The Subjective And Objective Assessment Of A Propeller Aircraft Interior Sound Quality," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, Madrid, 2019.

[31] G. Pietila and T. C. Lim, "Intelligent systems approaches to product sound quality evaluations – A review," *Applied Acoustics,* vol. 73, no. 10, pp. 987 - 1002, 2012.

[32] R. Sottek and T. Henrique, "AI-SQ Metrics: Artificial Intelligence in Sound Quality Metrics," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, Seoul, 2020.

[33] S. Giraldo, R. Ramirez, G. Waddell and A. Williamon, "A Real-time Feedback Learning Tool to Visualize Sound Quality in Violin Performances," in *10th International Workshop on Machine Learning and Music*, Barcelona, 2017.

[34] G. Lemaitre and P. Susini, "Timbre, Sound Quality, and Sound Design," in *Timbre: Acoustics, Perception, and Cognition*, Springer International Publishing, 2019, pp. 245-272.

[35] S. Bhattacharyya, V. Snasel, A. E. Hassanien, S. Saha and B. K. Tripathy, Deep Learning : Research and Applications, Walter de Gruyter GmbH, 2020.

[36] F. Fahy and P. Gardonio, Sound and Stuctural Vibration: Radiation, Transmission, and Response, Oxford: Elsevier, 2007, pp. 418-427.

[37] J. T. Du, W. L. Li, H. A. Xu and Z. G. Liu, "Vibro-acoustic analysis of a rectangular cavity bounded by a flexible panel with elastically restrained edges," *The Journal of the Acoustical Society of America,* vol. 131, no. 4, pp. 2799-2810, 2012.

[38] S. M. Kim and M. J. Brennan, "A compact matrix formulation using the impedance and mobility approach for the analysis of structural-acoustic systems," *Journal of Sound and Vibration,* vol. 223, no. 1, pp. 97-113, 1999.

[39] M. Vorländer, "Simulation and auralization of structure-borne sound," in *Auralization*, Springer, 2020.

[40] G. Wang, Y. X. Zhang, Z. B. Guo and Z. G. Zhou, "A novel hybrid deterministic-statistical approach for the mid-frequency vibro-acoustic problems," *Applied Mathematical Modelling,* vol. 83, pp. 202-219, 2020.

[41] D. Chronopoulos, B. Troclet, M. Ichchou and J. P. Lainé, "A unified approach for the broadband vibroacoustic response of composite shells," *Composites Part B: Engineering,* vol. 43, no. 4, pp. 1837-1846, 2012.

[42] V. Yotov, M. Remedia, G. Aglietti and G. Richardson, "Non-parametric stochastic FEM / hierarchical matrix BEM for efficient mid-frequency vibroacoustic response estimation," in *25th International Congress on Sound and Vibration*, International Institute of Acoustics and Vibration, 2018.

[43] M. Aretz, R. Nöthen, M. Vorländer and D. Schröder, "Combined Broadband Impulse Responses Using FEM and Hybrid Ray-Based Methods," in *EAA Symposium on Auralization*, Espoo, Finland, 2009.

[44] "VA ONE," ESI Group, 2020. [Online]. Available: https://www.esi-group.com/products/vibro-acoustics.

[45] R. Jiao and J. Zhang, "Vibro-acoustic modeling of a rectangular enclosure with a flexible panel in broad range of frequencies and experimental investigations," *Journal of Vibroengineering,* vol. 18, no. 4, pp. 2683-2692, 2016.

[46] J. Pan and D. A. Bies, "The effect of fluid–structural coupling on sound waves in an enclosure —Theoretical part," *The Journal of the Acoustical Society of America,* vol. 87, no. 2, pp. 691-707, 1990.

[47] H. Zhang, D. Shi, S. Zha and Q. Wang, "Vibro-acoustic analysis of the thin laminated rectangular plate-cavity coupling system," *Composite Structures,* vol. 189, pp. 570 - 585, 2018.

[48] O. Robin, J.-D. Chazot, R. Boulandet, M. Michau, A. Berry and N. Atalla, "A Plane and Thin Panel with Representative Simply Supported Boundary Conditions for Laboratory Vibroacoustic Tests," *Acta Acustica united with Acustica,* vol. 102, pp. 170-182, 2016.

[49] "MATLAB Documentation: DCT for Speech Signal Compression," Mathworks, [Online]. Available: https://www.mathworks.com/help/signal/ug/dct-for-speech-signal-compression.html. [Accessed 2019].

[50] P. Mahana and G. Singh, "Comparative Analysis of Machine Learning Algorithms for Audio Signals Classification," *International Journal of Computer Science and Network Security,* vol. 15, no. 6, pp. 49-55, 2015.

[51] D. Bertero and P. Fung, "Deep Learning of Audio and Language Features for Humor Prediction," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, Portorož, Slovenia, 2016.

[52] F. Rong, "Audio Classification Method Based on Machine Learning," in *International Conference on Intelligent Transportation, Big Data & Smart City*, Changsha, 2016.

[53] J. Brownlee, "What is Deep Learning?," Machine Learning Mastery Pty. Ltd., 16 August 2019. [Online]. Available: https://machinelearningmastery.com/what-is-deep-learning/. [Accessed December 2020].

[54] H. Purwins, B. Sturm, B. Li, J. Nam and A. Alwan, "Introduction to the Issue on Data Science: Machine Learning for Audio Signal Processing," *IEEE Journal of Selected Topics in Signal Processing,* vol. 13, no. 2, pp. 203-205, 2019.

[55] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss and K. Wilson, "CNN architectures for large-scale audio classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, New Orleans, LA, 2017.

[56] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," *Nature,* vol. 521, p. 436–444, 2015.

[57] B. Moews and G. Ibikunle, "Predictive intraday correlations in stable and volatile market environments: Evidence from deep learning," *Physica A: Statistical Mechanics and its Applications,* vol. 547, p. 124392, 2020.

[58] R. Gandhi, "Support Vector Machine — Introduction to Machine Learning Algorithms," Towards Data Science, 7 June 2020. [Online]. Available: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47. [Accessed January 2021].

[59] G. Sharma, K. Umapathy and S. Krishnan, "Trends in audio signal feature extraction methods," *Applied Acoustics,* vol. 158, p. 107020, 2020.

[60] J. Jogy, "How I Understood: What features to consider while training audio files?," Towards data science, 06 September 2019. [Online]. Available: https://towardsdatascience.com/how-i-understood-what-features-to-consider-while-training-audio-files-eedfb6e9002b. [Accessed December 2020].

[61] J. Lyons, "Mel Frequency Cepstral Coefficient (MFCC) tutorial," 2012. [Online]. Available: http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/. [Accessed December 2020].

[62] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 28, no. 4, pp. 357-366, 1980.

[63] L. Muda, M. Begam and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques," *Journal of Computing,* vol. 2, no. 3, 2010.

[64] T. Bäckström, "Cepstrum and MFCC," 21 May 2019. [Online]. Available: https://wiki.aalto.fi/display/ITSP/Cepstrum+and+MFCC. [Accessed December 2020].

[65] H. Kim, N. Moreau and T. Sikora, MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval, John Wiley & Sons, Ltd, 2005.

[66] M. A. Hossan, S. Memon and M. A. Gregory, "A novel approach for MFCC feature extraction," in *4th International Conference on Signal Processing and Communication Systems*, Gold Coast, 2010.

[67] L. Shi, C. Li and L. Tian, "Music Genre Classification Based on Chroma Features and Deep Learning," in *Tenth International Conference on Intelligent Control and Information Processing (ICICIP)*, Marrakesh, Morocco, 2019.

[68] M. Jalil, F. A. Butt and A. Malik, "Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals," in *The International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE)*, Konya, Turkey, 2013.

[69] "Calculation of loudness level and loudness from the sound spectrum - Zwicker method - Amendment 1: Calculation of the loudness of time-variant sound," 1 March 2010. [Online]. Available: https://standards.globalspec.com/std/1226735/DIN%2045631. [Accessed 2020].

[70] H. Fastl and E. Zwicker, Psychoacoustics: Facts and Models, Berlin: Springer-Verlag, 2007.

[71] F. R. a. D. Schafer, "Strategies for Variable Selection," in *The Statistical Sleuth*, Boston, Brooks/Cole, Cengage Learning , 2013, p. 358.

[72] W. Koehrsen, "Beyond Accuracy: Precision and Recall," Towards Data Science, 3 March 2018. [Online]. Available: https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c. [Accessed December 2020].

Appendix

Relevant files are provided in the zip file at

https://byu.box.com/s/rgzi2h0ad45nm8svc0jwyyh7vijsshf4

The contents are as follows:

Hybrid Method

- Modal_response.mat - contains the low-frequency modal response for the experimental setup.
- SEA_response.txt - contains the high-frequency SEA response for the experimental setup.
- plate_cavity_hybrid_model.m - is a MATLAB code example of the hybrid method. It imports the above responses and combines them into a single broadband response.
- wavs
    - original - contains the original audio files for aurlaizations.
    - auralizations - contains the saved audio files output by plate_cavity_hybrid_model.m.

Listening Test

- listening_test_app.mlapp - MATLAB app to run listening tests. Open within MATLAB to edit.
- results.xlsx - results from the listening test are saved here.
- wavs - wav files for the listening test.
    - intro - wav files used for the instructions tab of the listening test.
    - original - unaltered original wav files.
    - test - wav files to be tested, generated with the hybrid method.

Machine Learning Model

- CAT_model_compression.py - code for ML model for predicting perception of compressed audio.
- CAT_model_simulations.py - code for ML model for predicting perception of simulated (from hybrid method) audio.
- Data - contains the data that the ML models use (wav files, SQ metrics, and listening test results).

Sound Quality Metrics

- matlab_sq_functions - functions required to calculate SQ metrics.
- sq_metrics - script to run if one desires to calculate SQ metrics for a group of audio files.