A neural network to classify auditory signals for use in autonomous harvester control systems

by

Avery Simundsson

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

In partial fulfilment of the requirements of the degree of

MASTER OF SCIENCE

Department of Biosystems Engineering

University of Manitoba

Winnipeg

TABLE OF CONTENTS

List of Figures

List of Tables

1 ABSTRACT

As agricultural machinery moves into the digital era, significant developments in the available technology make autonomous farm vehicles more feasible, affordable, and desirable. One of the challenges of effective autonomous vehicle control specific to agriculture is the ability of the vehicle to interpret and adapt to constantly changing conditions. There are many types of sensors able to identify specific changes in conditions (elevation, temperature, image etc.), but a single indicator to signal a variety of changes in operating conditions would be beneficial to a remote human operator, particularly in triggering an automatic shutdown to prevent machinery damage. Auditory information is a primary indicator of changing conditions to an in-cab operator, particularly in detecting mechanical overload in a combine. This paper explores the potential for auditory information, which has proven valuable to an in-cab operator, to be used in autonomous vehicle control. Sound was recorded at a sampling rate of 48 kHz near the combine chopper for three different operating modes during the same harvest day for canola. Samples from each clip were segmented and analyzed using the Fast Fourier Transform (FFT) in MATLAB. The FFT generated a power spectral density (PSD) function for each segment, from which eight features were extracted and labelled to create feature vectors. These vectors were used to create a classifier using a feedforward pattern recognition neural network. The network used scaled conjugate gradient backpropagation training to achieve accuracy of 100%. The speed of sampling and analysis is sufficient to be used in real time machinery analysis and control.

2 INTRODUCTION

There is an increasing demand for quality food across the globe with simultaneous pressure to reduce the environmental effects of agricultural production. As the world population grows, natural resources such as water and land become increasingly valuable and there is a push to "produce more with less". Continued migration to cities reduces rural labour forces that are already scarce in some areas. Recent advances in cloud computing, data processing, and technology available in farm machinery have made the idea of autonomous farm vehicles an attractive potential solution to increase efficiency, management resolution, and decrease dependency on seasonal labour.

Autonomous farm vehicles have been designed and built by several companies and universities around the world but are generally not yet commercially available for reasons including difficulty operating in dynamic environments with uneven terrain, and adequate control systems to account for the intricacies of specific tasks. As these obstacles are overcome, humans will likely transition from machinery operators to a more supervisory role, potentially for multiple autonomous vehicles simultaneously. An in-cab operator can extract visual, olfactory, and auditory information to quickly discern the state of mechanical operation. In the supervisory role that may be off-site, these channels of information may be harder to access without creative and innovative monitoring methods.

Many operations in agriculture are time-sensitive and breakdowns can be very costly in terms of both time and service to what is often a rural, possibly remote, area. It may be difficult to monitor operations with many moving parts, such as a harvester, without the

benefit of sensory information that can be rapidly processed by the human brain. A few seconds of delay in processing information may be the difference between preventing a failure and a serious harvest delay.

Translating operational information in traditional ways (rpm, engine temperature, and other numerical representations of information) requires a large amount of data that may be difficult for a human to process continually for repetitive operations over long periods of time. To effectively supervise multiple operations under these conditions, it is necessary to carefully convey essential information to avoid sensory overload, as well as provide a rapid alert system for failure or mis-operation. Translating valuable sensory information, which has been fundamental to human perception of the environment throughout our evolution, rather than mechanical information (temperature, pressure, rotational speed, etc.) may be a relatively inexpensive and reliable method of failure prevention.

This thesis will explore the possibility of using the auditory information produced by a combine to determine changes in the mode of operation. Though this information may not be enough to determine exactly what is occurring, it may simply provide information that can be rapidly processed to alert a supervisor that something in the operation has changed. If this information is more effective than more common numerically represented information from gauges or monitors as is standard today, it may prove a valuable tool for remote operators. There may also be further opportunity to refine the processing of auditory information to detect threshing load, engine RPM, or any number of other important operational factors from a single point of reference. This

would mean significantly less bandwidth would be required to transport data to the supervisory area, though it may require more processing power at the supervisory site. This would be a significant advantage for many farms given the limited service available for data transfer in most rural areas.

3 LITERATURE REVIEW

3.1 Evolution of technology in agriculture

The advent of precision agriculture technologies in the past decades has increased our ability to cope with on-farm and in-field variability and to incorporate a much larger volume of information into management decisions (Fountas, et al., 2015). The ISOBUS standard has increased vehicle monitoring and the collection of information to optimize field operations and maximize productivity. Extensive, geo-referenced databases are being built for the purposes of providing enhanced decision support for greater management ability. This has caused a shift in farmer responsibility from vehicle operation to monitoring and strategic management.

The ISOBUS protocol also facilitates the exchange and storage of information collected by sensors with controllers and software packages (Fountas, et al., 2015). The main challenge at this point in time is to integrate this data, which is generally heterogeneous in nature, into a functional farm management system. Data conversion, formats and interfaces are often incompatible between applications which introduces significant complications impeding effective use of data. In addition, compliance with legal regulations and evolving agricultural production standards increase the complexity of a

proper automated management system if innovations such as machine-readable decision rules and compliance checking is to be implemented.

Fountas et al. (2015) describe a shift in thinking from considering the farmer as the center of the management system to a machine-centric approach, where the information flow from machinery provides an automated part of the decision-making process (Fountas, et al., 2015). To be successful, such a system needs to be: (1) in situ sensor-based; (2) automated for real-time or near real-time computer processing and transformation into knowledge for decision-making (3) packaged so that sensing and processing of information are a part of the equipment used to accomplish the required management action, and (4) transparent to the operator/manager for decision evaluation and confirmation.

In a study of driver perception response time, Olson and Sivak (2017) found that the average time of a person to sight an obstacle and apply the brake was 1.6s for 95% of testing drivers, regardless of age (Olson and Sivak, 1986). A similar study showed that the 85th percentile of people have reaction times of 1.3-3.6 s depending on the driving conditions (night/day, moving vs stationary obstacle, etc.) (Triggs, et al., 1982). Though these studies take place on highways in personal vehicles, they cover a variety of driving conditions and we can assume that reaction times to stimulus while driving farm machinery would be similar. Therefore, any system that can provide a real-time reaction time (time from sensing the issue to implementing a response mechanism) of less than 1 s can be considered to be faster than a human response, and sufficient for a vehicle control system.

3.2 Autonomous agricultural vehicles

Autonomous agricultural vehicles have been in development for several decades and significant advances have been made in guidance systems. It is nearly universal on modern farms to have some form of autonomous guidance in any vehicle perform a field operation to the obvious benefits of reduced overlap and operator fatigue. However, a vehicle with an autonomous guidance system still requires a human operator to control the actual function of the machine and adjust parameters as necessary. In a combine, it may be necessary to adjust the concave clearance, cylinder speed, and fan speed to maximize harvest efficiency while minimizing losses and reducing seed damage. As conditions change throughout the day, these parameters should continue to be adjusted. Being able to make these adjustments automated, or even to alert an operator to changing conditions requiring adjustment, would be a significant step forward in a fully autonomous harvest machine.

A significant amount of research around autonomous vehicles has been focused on guidance systems. Automated guidance systems have been sought after in agriculture for several decades. In the 1980's, advances in computers and image sensors spurred research in guidance systems based on machine vision (Reid, et al., 2000). There were also successful programs in automated harvesting of oranges at the University of Florida during this decade. The sensors that have proved to be most important include those for machine vision, mechanical feelers, and GPS sensors.

Part of the reason for a high interest in developing sophisticated navigation systems for agricultural vehicles is that they often operate on uneven and changing terrain. The

ability of a vehicle to avoid low-lying wet areas or adjust operation for variations in topography would be beneficial for some farming systems (Reid, et al., 1999).

Recent developments in agricultural autonomous vehicles Human operators must use a significant amount of intelligence and judgement to rapidly process both visual and audio cues, motion sensations, and then react accordingly depending on the required vehicle operation or maneuver (Reid, et al., 1999). Combined with successive long days and highly repetitive tasks, this creates an environment conducive to fatigue, decreased efficiency, and potential safety hazards. Introducing greater levels of autonomy, particularly in guidance and navigation, can reduce the intensity of work and lower the number of sensory distractions. Vision sensors and GPS sensors are the most recent additions to autonomous guidance and GPS guidance is now considered a standard feature on agricultural equipment in many parts of the world.

No individual sensing technology can be considered ideal for all agricultural vehicles in all situations (Reid, et al., 1999). However, having multiple sensors provide data yields superior results, such as combining a fiber optic gyroscope (FOG) with a real time kinematics DGPS. The FOG sensor provides instantaneous heading information while the DGPS provides offset information related to its desired path. This combination provides an accuracy of 3 cm at a rate of 5 Hz (Reid, et al., 1999). Fusion of these sensors minimizes the effects of their individual weaknesses.

The idea of autonomous agricultural vehicles has been of interest for many years with early examples of 'driverless tractors' appearing in the 1950s and 1960s using leader

cable guidance systems (Hidalgo, et al., 2008). Guidance systems have been the first part of machinery operation to be automated with guidance systems using GPS guidance considered the norm in North America. While these systems are generally not fully autonomous, researchers have paired GPS guidance with numerous sensors (inertial, laser-based, geomagnetic direction sensors, etc.) to build systems that could function autonomously. Other guidance systems that have been explored include machine vision, particularly for row-cropping.

Extensive research around the use of autonomous vehicles in agriculture has been performed, but there are still extremely limited commercial options widely available (Bechar, et al., 2016). Much of the difficulty in making these robots commercially available lies in the fact that they required sophisticated, intelligent algorithms for sensing, planning, and controlling various aspects of the implement or machine in an unstructured and changing environment. Productivity has increased substantively from other innovations in precision agriculture, such as precision seed placement during seeding and optimization of in-season nitrogen applications. As farm size grows across North America, skilled and even minimally trained workers become scarcer. Incorporation of autonomous machines would reduce the overall need for labour (though creating a need for a different kind of labour) and increase efficiency, reliability, and precision while reducing the need for human action.

There are many potential benefits of incorporating autonomy into vehicle operations, such as increased productivity, application accuracy, and operational safety (Li, et al., 2009). However, development can be difficult due to the unique set of circumstances

that agricultural vehicles must operate in. Operating areas are large, surfaces can be uneven, and environmental conditions can be harsh. It is also necessary to make these machines affordable as the size of individual agricultural operations may limit their ability to accept the risk of a new technology or pay a high price for the latest technology (Li, et al., 2009). This is in contrast to many industrial applications where autonomous machines are currently used where tasks are often repetitive, simple, well-defined with distinct boundaries or tolerances in stable and controlled environments (Bechar, et al., 2016).

Although there are challenges, there are also advantages to developing autonomous vehicles for agriculture compared to other industries (Li, et al., 2009). The physical working environment does not generally change, as field boundaries, topography, and obstacles are usually static features. There are a limited number of plants used in a rotation, making their identification less complex.

3.3 Auditory information & control

Unmanned aerial vehicles (UAVs) are a prominent example of the success of autonomous vehicle use in agriculture. While traditional farm vehicles are not commercially available as an autonomous version, unmanned aerial vehicles (UAVs) are. The success and commercialization of UAVs may provide insight into strategies for wide-scale deployment of autonomous ground vehicles. Currently, one UAV is managed by one operator, or a team of operators, but there is a desire to reverse this trend so that one operator is able to control multiple UAVs. Donmez, Cummings, and Graham (Donmez, Cummings, & Graham, 2009) have studied the effectiveness of

sonifications (continuous auditory alerts). UAV supervisors are required to monitor a number of tasks, and information is generally conveyed visually and supported with discrete auditory alerts. Visual overload can be draining to an operator and perceptual offloading to other senses can be beneficial. When sonifications were compared with discrete auditory signals, sonifications produced faster reactions (19% enhancement). Sonifications were most beneficial in aiding operators in predicting future states but may potentially interfere with monitoring tasks that require divided attention.

Other studies have tested the use of sound more specifically in agriculture (Karimi, Mondor, & Mann, 2008). Pilot testing on the effect of auditory cues on operator performance showed that it was suited for some tasks, but not for others. Steering was found to be performed best as a purely visual task, while monitoring of an external task (e.g. an implement) benefited from auditory cues.

Sound can sometimes be a distracting noise, and some "noise design" may be necessary to be used effectively in autonomous vehicles (Barrass, 1997). Stephen Barrass describes auditory information design as "a mapping of information to perceptual relations in the acoustic domain (sounds) to meet the information requirements of an information processing activity". The approach to designing a useful form or representation of auditory information includes an analysis of the information requirements of the activity, and the design of auditory representation that meets those information requirements.

Auditory information in mechanical systems Sound can be a key part of mechanical analysis in a conventional setting (Donmez, Cummings, & Graham, 2009). Many mechanics use sound as a tool for preliminary diagnosis, and even individuals with little mechanical aptitude understand that an unusual sound in a vehicle is a signal of an engine malady. In the past, combine operators have relied on sound as an indicator of over-capacity when threshing (Donmez, Cummings, & Graham, 2009). Even with extensive visual displays of information in modern combines, audible cues to overload may allow for quicker response times (Donmez, Cummings, & Graham, 2009). In the case of autonomous vehicles, it is unlikely that a human controller will be directly listening to each machine. However, it is important to evaluate how else sound, which has proven to be an invaluable tool to an operator in the cab, can be used in autonomous vehicle control for a remote operator.

Though there are several monitors and alerts in modern combines to provide information on the threshing process, vibration (sound) is generally one of the primary indicators to the operator that a change in threshing load has occurred. It is possible that this signal could be used in autonomous vehicle control to signify changes in the operating conditions. This signal could be expanded upon to indicate the actual current machine load, adjust threshing parameters to increase efficiency and reduce losses, and prevent overloading and plugging which can cause machinery damage and expensive harvest delays. If a single signal could provide information on multiple machinery parameters and indicate necessary adjustments faster than a human operator, this would reduce losses, increase efficiency and mitigate machinery damage or delays due to overloading.

3.4 Sound analysis & classification

Processing of auditory information falls into the category of signal processing, which encompasses the analysis, synthesis, and modification of signals. Sound waves can be represented in a number of ways, but in order to analyze or manipulate them, they are represented in the form of an electrical quantity (Priemer, 1990). Sound analysis encompasses a wide range of activities, but this paper will focus on methods of classifying sounds through signal analysis for identification.

Classification of music is an excellent example of processing audio signals for rapid identification. Music Information Retrieval (MIR) is a field of science that is becoming increasingly important as consumers become increasingly accustomed to tailored experiences in everything from movie selection to curated playlists of new songs. In 2005, the International Society of Music Information Retrieval (ISMIR) launched a competition called Music Information Retrieval Evaluation eXchange (MIREX) to evaluate algorithms in 17 categories, including melody extraction, cover song identification, or query by singing (Gwardys, et al., 2014). Classification is accomplished through different methods of extracting and evaluating specific 'feature sets' to determine the optimal method for recognizing a signal.

Classification, at its essence, is simply pattern recognition. A pattern recognition algorithm is one that takes examples of complex signals with the correct decisions made for them, and then make decisions for future signals based on the original set of data (Ripley, 2007). Pattern recognition is something that humans tend to be relatively good at, but as machines grow more sophisticated, tasks involving this skill are being

transferred to machines. Problems in which pattern recognition by machines have been implemented include diagnosing disease, classifying galaxies by shape, selecting moves in games such as chess, and reading hand-written symbols.

The goal of pattern recognition is to create a classifier which can analyze specific features of an item as its input and return a label or value indicating grouping to which the item belongs (Mahana, et al., 2015). The specific patterns that can be recognized by the algorithm are based on features of the signal. A feature is a distinctive measurement, transform, or structure component extracted from a pattern that is distinguished from a regular vector, to be used for classification. The features are the inputs to the algorithm that are expected to be predictive of the outcome. The purpose of feature extraction is to identify information that is most useful for determining the classification of the signal. An example outside of music retrieval is classifying brain electrical activity from an electroencephalogram (EEG) signal to provide a diagnosis (Al-Fahoum, et al., 2014).

Looking at the example of music classification, the most popular methods used to extract engineered features from music include spectrogram, zero crossing rate, spectral centroid, and fundamental frequency. Since these methods have been successfully used to classify music, which varies widely in properties, it is likely they can successfully classify sounds in other realms and have been used successfully in speech processing, noise cancellation, data compression and transmission applications.

More sophisticated methods of analysis include Chromograms or Mel Frequency Cepstrum Coefficients (MFCC). Both the Chromagrams and MFCC are perceptually motivated, meaning they are designed specifically to process sound in a similar manner to the human ear. For example, perceptual experiments have shown that the human ear derives more information from lower frequencies in a sound as opposed to higher ones (Prahallad). Therefore, the Mel Frequency Analysis technique uses more filters in lower frequency regions, and less in high frequency regions. MFCC is often used in high-end speech recognition systems.

Spectral analysis is another method of representing auditory signals that has been used successfully in classification applications. Spectral analysis is a process used to estimate the power spectrum of a signal based on a time-domain representation (MathWorks, 2017). The spectrum density provides insight into the frequency content of a signal. The spectrum decomposes the signal into its various frequencies and detects periodicities. Non-parametric spectral analysis involves breaking the signal down into segments based on the time domain, applying a Fourier transform to each segment, and performing computational processes to sum and average the transform. A Fourier transform reveals the frequency contents of a signal for analysis and discovery (Nisar, et al., 2016). From this model, the algorithm calculates the implied power spectrum.

A spectrogram is a spectro-temporal representation of a sound (Raina, et al., 2014). Dominant frequencies can be visually identified on a spectrogram, as darker areas of the spectrogram represent higher energy densities. Spectrograms have been used by companies such as Spotify to aid in classifying music into different genres (Dieleman,

2014), though it has been a larger challenge to use this audio signal for actual recommendations based on listener preference. Identifying music genres is done with the use of mel-spectrograms in convolutional neural networks.

Most large mechanical systems tend to be in the low-frequency range (Boyce, 2012) which makes the Mel Frequency Analysis an interesting candidate for this research. Chromograms may also be useful, considering the motivation for this research is the ability of humans to recognize patterns in machinery sound. Zero-crossing rate, often used to distinguish between voice and non-voice sounds, is a method of measuring the frequency content of the signal (Zero crossing rate and Energy of the Speech Signal of Devanagari Script, 2014) which may be indicative of operating mode. Spectral centroid and formant (dominant) frequency are also methods of measuring frequency characteristics.

These descriptions are indicative that there are likely multiple methods of successfully gathering useful information from machinery sounds for use in autonomous control systems (or other applications). Because this research focuses on using feature extraction from raw audio signals with minimal processing, frequency characteristics (spectral centroid and fundamental frequency) will be explored in more detail.

3.5 Audio features

Spectral Centroid The spectral centroid of a signal is used to detect the "center of mass" of the spectrum and can be perceptualized as the "brightness" of a sound (Perceptual effects of spectral modifications on musical timbres, 1978). In this calculation, the spectrum is considered as a distribution of values, which represent the frequency, and the probabilities to observe these values are the normalized amplitude (Peeters, 2004). Sub band spectral centroids have been proven to be closely related to spectral peak positions in both clean and noisy signals in speech recognition applications (Robust Speech Recognition in Noisy Environments Based on Subband Spectral Centroid Histograms, 2006). In fact, when compared with MFCC as a speech recognition feature, sub band spectral centroid histograms (SSCH) outperformed MFCC features in the presence of additive white noise, and also in more complex recognition tasks. SSCH features were also shown to be considerably more robust in complex speech recognition tasks than zero-crossing with peak amplitudes feature in the same study, except when human babble was introduced as noise.

Formant Frequency Two critical characteristics in sound analysis are bandwidth and formant frequency. The frequency bandwidth of a sound, typically measured in Hz, is the range of the upper and lower frequencies in a continuous bandwidth. A formant frequency is a concentration of acoustic energy around a particular frequency and can be defined as a "spectral peak" of the sound spectrum. Identification and analysis of formant frequencies is a technique used in speech and speaker recognition, biomedical signal analysis, and musical instrument analysis (Nisar, et al., 2016). Since formants are most often used in linguistic analysis, they generally refer to sounds made by the human voice (Abhang, Gawali, & Mehrotra, 2016) and can be used to

differentiate between various sounds made in speech. Several formants may exist in a sound as a relative maximum or minimum, or as amplitude peaks in the frequency spectrum. Identifying the formants in a particular sound identifies the frequency at which sound energy is concentrated and the dominant frequencies of the sound. Determining the power of the formant bandwidth provides another measure, or feature, of the sound. Using dominant frequency information for feature extraction in automatic speech recognition tasks has been shown to increase the robustness when background noise is added (Robust Speech Recognition in Noisy Environments Based on Subband Spectral Centroid Histograms, 2006).

Formants represent the harmonic or dominant frequencies of a sound and provide the information necessary for the human brain to distinguish between vowel sounds (Raina, et al., 2014). Vowel sounds can be represented purely quantitatively by formants. Though vowels always have four formants and some can have more than six, the first four formants are generally used for identification with the first two being the most important (Raina, et al., 2014). The lowest frequency formant is notified as F1, the second lowest as F2, and so forth. This indicates that identifying formants in other sounds may also be useful in distinguishing nuance between similar sounds, particularly the first and second formants.

Despite a wide amount of variation in human anatomy, the sounds made by the human vocal tract produce similar patterns for specific sounds regardless of who they are made by, particularly in vocal tract resonance ratios (Ross, et al., 2007). The fact that the frequency relationships of the first two formants in vowel phones represent all

12 intervals of the chromatic scale has been used to explain how cultures across the world and through history have preferred music using pitch intervals dividing octaves into the 12 tones of the chromatic scale.

Fundamental frequency and frequency ratios (F1 and F2) have also been used to understand why certain musical patterns can elicit different human emotions (Bowling, et al., 2010). Bowling et. al showed that the spectra of major musical intervals are more similar to spectra of excited speech compared with minor musical intervals, which are more similar to subdued speech. Listening to these musical sounds connect our brains to the respective emotion, which is one reason why songs can be categorized into 'sad', 'wistful', and 'energizing'.

The dominant frequencies of audio collected from machinery may be equally indicative of machinery operating mode as they are of vowels and consonants in human speech and will be explored in this research as a defining feature. Other features that may be extracted based on formant are the distance of frequency bin between dominant formants, (i.e. F1-F2, F1-F3, etc.) or the ratio of power between formants (i.e. F1/F2, F1/F3, etc.).

3.6 The Fourier Transform

Time-frequency analysis is commonly used to characterize phenomena such as vibration, music, and biomedical signals (Nisar, et al., 2016). Fourier transforms are typically used to gain useful information from these phenomena, though this method ignores all time-related information. The Fourier Transform operates on the principle that

every signal in the time domain is composed of a number of weighted sum of sines and cosines of different amplitudes and frequencies, which also means that any signal can be decomposed into a number of different sine waves. Each sine wave can then be analyzed to understand the different frequencies that are present in the original signal, thus converting the signal from the time domain to the frequency domain.

Ubiquitous use of the Fourier transform in signal process and analysis, and unanimous acceptance as a valued function (Kumar, Singh, & Saxena, 2011), make it an obvious candidate for evaluation in this study. According to Vijay Madisworetti, "the Fourier transform is a mathematical tool that is used to expand signals into a spectrum of sinusoidal components to facilitate signal representation and the analysis of system performance" (Madisetti, 2009). It is used for spectral analysis, spectrum shaping, and decomposition of input signals. There are several different forms of the Fourier Transform, including the short-time Fourier Transform (STFT), the continuous-time Fourier series (CT), the discrete-time Fourier transform (DTFT), the discrete Fourier transform (DFT), and the fast Fourier transform (FFT).

The short-time Fourier Transform (STFT) analyzes small sections of a signal at a time rather than the signal as a whole. In order to be successful using the STFT, proper window selection is essential. The chosen window size should be large enough to localize the frequency domain, but small enough to ensure that the input signal falling within it should remain stationary.

The CT method is useful in the analysis and design of linear continuous-time systems, particularly in cases where the frequency domain is used to specify characteristics (e.g. linear filtering) (Madisetti, 2009). Useful characteristics of CT make it an excellent method for shifting spectral energy among different frequency bands, solving differential or integral equations, of specifying systems directly by impulse or frequency response.

The discrete Fourier transform (DFT), which works by mapping a sequence *x(n)* into the frequency domain, is the most widely used method in signal processing (Rao, Kim, & Hwang, 2010). This method represents the signal in the frequency domain components in discrete values called frequency bins (National Instruments, 2019). DFT is used for a wide variety of applications, including multiple time series analysis and filtering, filter simulation, and channel separation and combination. Using DFT requires many complex arithmetic operations, but this process can be simplified through efficient algorithms (Rao, Kim, & Hwang, 2010).

The DTFT shares many properties of CT Fourier transform (Madisetti, 2009). However, time domain differentiation and integration are not defined for DT signals. This requires DT users to manipulate difference equations in the frequency domain, making linearity and index-shifting very important.

The Fast Fourier Transform (FTT) is not technically a standalone Fourier transform, but rather an efficient algorithm for computing DFT (Madisetti, 2009). The reduction of computational effort in FFT makes real-time DFT analysis practical in situations when it

would otherwise be unfeasible. The FFT is a one-push algorithm that allows efficient implementation of the DFT. The FFT allows for simplification, reduced storage requirements, and reduced computational error. The FFT performs best when processing stationary signals as compared to non-stationary signals (Al-Fahoum, et al., 2014). It is particularly appropriate for narrowband signals (such as a sine wave) and has superior speed over almost all other methods for real-time applications. A stationary signal is one in which the statistical properties of a random process do not depend on the time index.

Because of its ubiquitous use in signal processing, error reduction, speed of processing and low storage requirements, the FFT is most suitable for real-time analysis and will be used in this research.

FFT is commonly used for spectral analysis analog waveforms sampled and recorded over a finite period of time (Madisetti, 2009). The FFT considers the data set as one period in a periodic sequence. If the waveform being analyzed is not periodic, there is a danger of harmonic distortion. This is due to sharp discontinuities at boundaries of the data segment in the periodic waveform created by the FFT. Windowing the data so that segment boundaries are tapered to zero can mitigate harmonic distortion, as can removing the mean of the data.

Windowing Window functions are used in harmonic analysis to reduce spectral leakage, which occurs when fine spectral lines appear to spread into wider signals (Kumar, Singh, & Saxena, 2011). A window function in signal processing is a

mathematical function that gives a value of zero outside of a specified interval (Bojkovic, Bakmaz, & Bakmaz, 2017) and therefore reduces the amplitude of the discontinuities at the boundaries of each finite sequence. There are several standard windows used to optimize applications of signal processing, but all are designed to reduce side lobes of the spectral output of the FFT at the expense of resolution. Choosing a window that best suits the application minimizes this effect, based on the parameters required to appropriately process the signal in question. Most window functions place more emphasis on the center of a data set compared to the edges, which results in loss of information on both sides of the data set. This can be overcome by overlapping individual windows of data.

Window types to reduce side lobes include Hamming, Hann, Chebyshev, and Taylor (Basit, Qureshi, Khan, Rehman, & Khan, 2017). Out of these, the Hamming window is simple to apply and produces relatively sharp main lobe width. The Hamming window in particular is noted for reasonable frequency resolution and acceptable noise performance in smoothing a function prior to applying Fourier analysis (Bojkovic, Bakmaz, & Bakmaz, 2017). The Hann and Hamming windows are both acceptable for most signal processing tasks (National Instruments). However, the Hann window eliminates any discontinuity by zeroing at both ends; the Hamming signal results in slight discontinuity. This means that the Hamming window is marginally more effective at cancelling the nearest side lobe.

The window size represents a number of samples in a particular window and is dependent on frequency, intensity, and changes in the signal. Knowledge of the signal

characteristics is necessary to determine the correct window size. Selecting an appropriate window size based on signal characteristics is critical for proper application of the Fourier method (Nisar, et al., 2016). Even with no window specifically applied, the FFT still provides a windowing effect. No window is often called the "uniform" or "rectangular" window. Because the audio signals captured from large machinery are typically considered broadband (Lee, et al., 2017), the rectangular window will be used in this research.

3.7 Classification

Classification is simply the process of classifying something according to shared qualities or characteristics. Classification is used in biology to differentiate between mammals, reptiles, birds and amphibians based on whether they fly, are warm-blooded, give birth to live young or lay eggs, etc. Some of these categories overlap. For example, reptiles, birds and amphibians lay eggs whereas mammals give birth to live young. This single descriptor can distinguish animals which fall under the category of 'mammal, but further descriptors, or features, are needed to distinguish between the remaining classes. 'Able to fly' may be a second feature that can specifically distinguish 'bird' from the remaining classes, but a third (or more) feature(s) are needed to differentiate, with confidence between reptiles and amphibians. These are examples of 'yes/no' features in which the answer to 'able to fly' falls into one of two categories: yes or no. Features such as 'number of toes" have a much wider variety of responses. Determining the correct features to compare in an efficient manner is a critical component of classification.

Classification can be binary (i.e. male or female) or multi-class as in the example of animal classification. There are many methods of classification such as Naïve Bayes, k-Nearest Neighbours, Decision Trees, Support Vector Machines and Random Forests. Each have unique methods of putting items in their respective classes and the best method is dependent on the number of classes, desired accuracy, amount of training data, and a number of other factors.

3.8 Machine learning for classification and audio signals

Audio classification and retrieval is a popular machine learning application, particularly in the areas of music and voice recognition (Mahana, et al., 2015). Both images and waveforms produced from audio signals can be used, depending on the method selected. The work done in classifying music into genres or styles can be leveraged for other audio classification applications, such as identifying specific operating characteristics of machinery.

Audio signal classification is fundamentally a pattern recognition problem (Mahana, et al., 2015). Classification involves extracting features that are expected to be optimal predictors of class for a signal, and then selecting a classifying tool to develop or apply algorithms by recognizing patterns. Machine learning algorithms are particularly useful in creating a classifier based on features.

Pattern recognition can be explained as an attempt to infer predictability in a set of known data or example solutions and to use this predictability to solve new instances of a problem (Bishop, 1995). These example solutions are known as a training set and

consist of correlated sets of input and output values. This set of values is used to define the discrepancy between the predictions made by the network compared to the desired outputs, called the error function. Once the network has been "trained" so that the error function is minimized, it can be used to predict the output values for a new set of inputs.

There are three main groups of machine learning algorithms (Mahana, et al., 2015):

Supervised: Training data is labelled with the correct result to provide feedback to the algorithm. Classification is based on prior knowledge. The classification function is derived by analyzing the relationship between attributes/features of an object and its grouping. This model is then used to classify future samples.

Unsupervised: No desired output exists, and no error signal is generated. This method is used to find hidden structures in unlabeled data.

Reinforcement: The algorithm learns its behaviour based on feedback from the environment. There is little need for the supervisor to have domain knowledge in the application area.

All data used in this research will be labelled with a specific class and data for the research is ubiquitous and easy to collect. Though all three algorithms can likely be useful in classifying machinery operation, this research will focus on supervised algorithms. Unsupervised and reinforcement algorithms may be used in the future to

understand what other mechanical factors may be identified through audio classification, or for improving the algorithm under various operation conditions.

In supervised pattern recognition, there is a set of *K* pre-determined classes and a desire to correctly label any example provided with its appropriate class (Ripley, 2007). Various measurements, called features, are used by the classifier algorithm to make these labelling 'decisions'. Along with being classified correctly, other outcomes include incorrect classification, or unknown classification. The error tolerance of the classifier is dependent on the cost of a mistaken or unknown classification.

Neural nets Once the features have been extracted from a signal, the next step is to select a method of classification, such as linear analysis, nonlinear analysis, adaptive algorithms, clustering and fuzzy techniques, and neural networks (Al-Fahoum, et al., 2014). Though it is possible that any of these may work, feedforward networks are most commonly used in pattern recognition applications (Use of Artificial Neural Networks in Pattern Recognition, 2010) due to their ability to ingest complex, non-linear data, use sequential training procedures, and adapt to the input (Use of Artificial Neural Networks in Pattern Recognition, 2010). Neural networks require a large amount of data to ensure accuracy. In regard to continuous machinery operation, data is relatively easy to collect and accumulate. Data within a class are relatively uniform meaning large amounts of data are not necessary compared to other applications where data may be widely variable within a class. For these reasons, neural networks will be explored as a methodology for this research.

A feedforward network accepts a set of inputs and maps them to corresponding output values and can be understood as a graphical representation of a parametric function (Bishop, 1995). Feedforward means there is no feedback to the input. Providing feedback can increase the performance of neural networks, but generally requires more complex construction. Feedforward networks consist of a series of connected layers between the network input and output. The output of each layer is used as the input to the subsequent layer from start to finish. Applications include fingerprint identification, recognition of handwritten numbers and letters, and medical image screening.

Conjugate Gradient Methods are accepted to be an optimization method that is well-suited to large-scale problems (Moller, 1993). However, standard CGMs increase the complexity of calculations in each learning iteration due to several required calculations of either the global error function or its derivative. A variation of a the CGM which allows the method to avoid these calculations is called the Scaled Conjugate Gradient Method (SCG). Scaled Conjugate Gradient (SCG) is a supervised learning algorithm for feedforward neural networks, which is faster than standard backpropagation and other Conjugate Gradient Methods.

Back-propagation training is a feedback mechanism that enables the network to adjust the connection weights back through the network layers. This trains it in response to representative examples. The back propagation algorithm allows the network to learn and store very large amounts of mapping relations of input-output model without the requirement of providing details of the mathematical equations that describe these

relations (Artificial Neural Network Application in the Diagnosis of Disease Conditions with Liver Ultrasound Images, 2014).

Creating a neural network is divided into three steps:

Training: Inputs and their corresponding correct outputs are presented to the network during training, and the network is adjusted according to its error to minimize the global error function (Moller, 1993). Training multiple times will generate different results due to different initial conditions and sampling.

Validation: These are used to measure network generalization, and to halt training when generalization stops improving (prevents overfitting). Training automatically stops when generalization stops improving, as indicated by an increase in the cross-entropy error of the validation samples.

Testing: These outputs have no effect on training of the model and so provide an independent measure of network performance during and after training. The results of testing provide an indication of the accuracy of the network.

4 OBJECTIVE

There has been a significant amount of research in recent years in how to use sound in classification and identification systems, particularly in music retrieval. There are a variety of ways to do this, but one of the simplest and most robust is to take the Fast

Fourier Transform of a signal and seek out specific features for identification. If characteristic features can be identified, they can be used to create a classification system through various methods. Recent advances in computing power have made neural networks an excellent candidate for training classification models, particularly if lots of data is available.

Using sound information in classification models has not been widely applied to mechanical systems, though changes in pitch, frequency, and volume caused by changes in machinery operation would make sound an excellent information source for these types of problems. The increasing demand for autonomous vehicles necessitates more sophisticated control systems which allow the machine to quickly and accurately assess the immediate environmental conditions and react appropriately. Agricultural vehicles, in particular, are an excellent candidate in which to incorporate automation due to the prohibitive costs of the machinery, difficulty in finding qualified labour to run them, and the repetitive natures of tasks. Human operators already rely on sound as a primary indicator of changes in these machines, particularly in combines.

The goal of this thesis is to see if the methodology used in other sound classification applications can also be used to classify machinery operations. Particularly, it will focus on the ability of a neural network to correctly identify the operation condition of a combine with the goal of creating an analysis technique that could be used for real-time monitoring and control (less than 1 second to detect the stimulus and implement a response).

5 METHOD

Sound recordings were taken from harvest video feed during the 2017 canola harvest in East Selkirk, MB. The canola was harvested with an S680 John Deere combine and video was captured with a GoPro Hero Session. The recordings were taken from the rear of the combine (see Figure 1) near the straw chopper, always when the combine was in forward motion. Though there were several GoPros placed in various locations around the combine (on the header, in the cab, etc.), the camera located at the straw chopper provided the best context as to operating mode. Recordings from the other cameras were not analyzed in this study. All recordings were taken in the same field on the same day from the same machine. Sound, sampled at a rate of 48 kHz with AAC compression and automatic gain control, was lifted from the video and converted to .wav files for analysis.

Figure 1. Location of the GoPro during operation, from which sound was lifted

Sounds recordings were isolated into three different operating modes (classes) of the combine.

Mode 1: The combine's engine is running, and mechanized threshing is not engaged ("Empty")

Mode 2: The combine's engine is running, and mechanized threshing is engaged with no actual threshing being performed ("Engaged")

Mode 3: The combine's engine is running, and mechanized threshing is engaged and utilized at approximately 80% capacity ("Full")

A short clip of sound (30-36 s) was taken from each operating mode as a representative audio sampling for that operating mode. Each operating mode was assigned a class (1, 2, or 3). Each recording was considered a stationary signal, independent of time, and the combine ran in a steady state during each clip. Each clip was segmented into identically sized segments of 2^n samples and truncated to remove any hanging samples. Segments of length 2^n were used as the FFT works most efficiently with samples groups of 2^n size. The goal was to find the smallest segment size that would yield favorable classification results, as this would result in the largest group of samples and the shortest time for analysis. Since this experiment involves a finite amount of data available to test and train, a larger segment size reduces the overall number of samples available for training. In real-time classification, too large a segment would also mean a longer processing time and therefore a longer reaction time for any automated response. The entire experiment was repeated for each segment size from 2^9 (512 samples, or 0.011 seconds) to 2^{14} (16,384 samples, or 0.341 seconds). The length of the FFT was kept constant throughout the experiment at 2^11 (2048).

The power spectral density (PSD) estimation was calculated for each segment using the FFT and the resulting PSD were grouped by class. All PSDs from a class were overlaid in a single chart to search for visual patterns within a class, and for differences between classes, as seen in Figure 2, Figure 3, and Figure 4. This visual analysis was used to select features that may be extracted for a machine learning model. Other features commonly used in audio signal feature extraction determined through a literature review were also selected for analysis, as discussed in the previous sections.

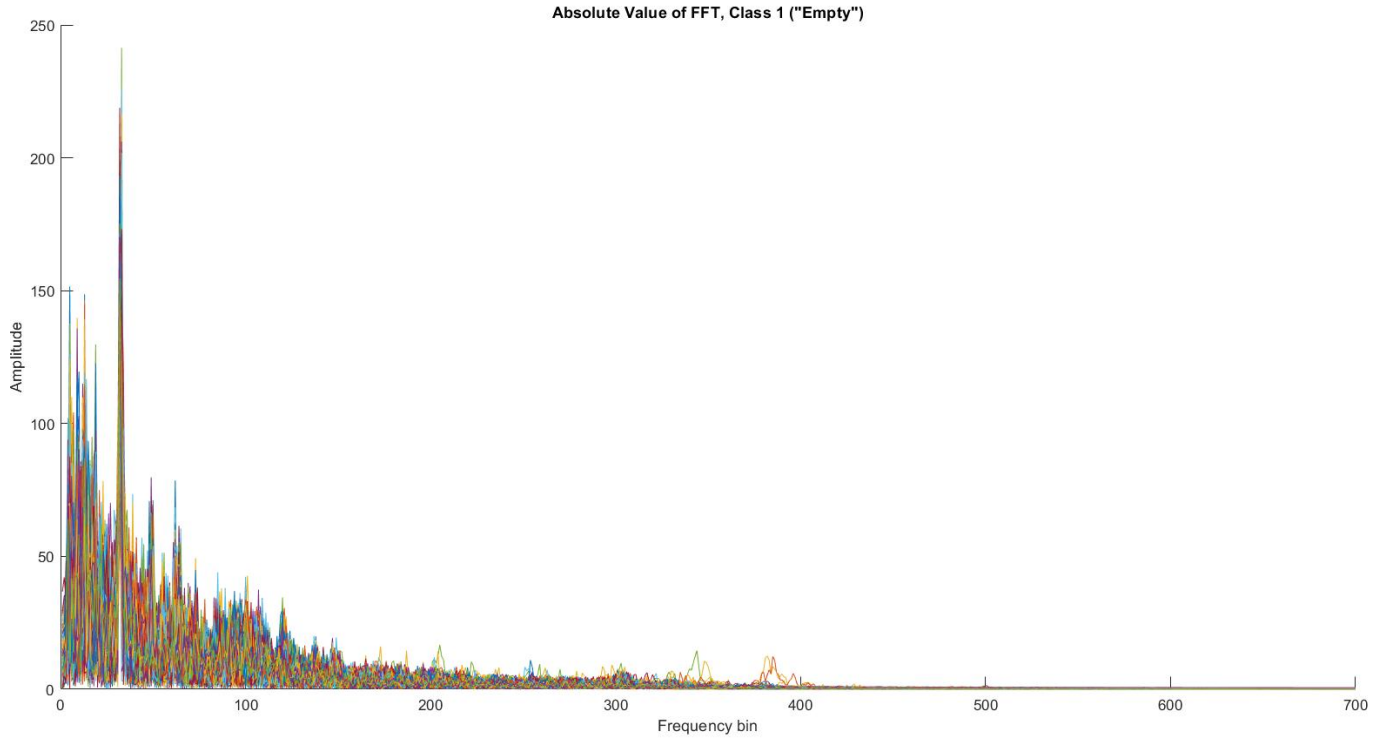

Figure 2. Absolute Value of FFT for samples from Class 1 "Empty"

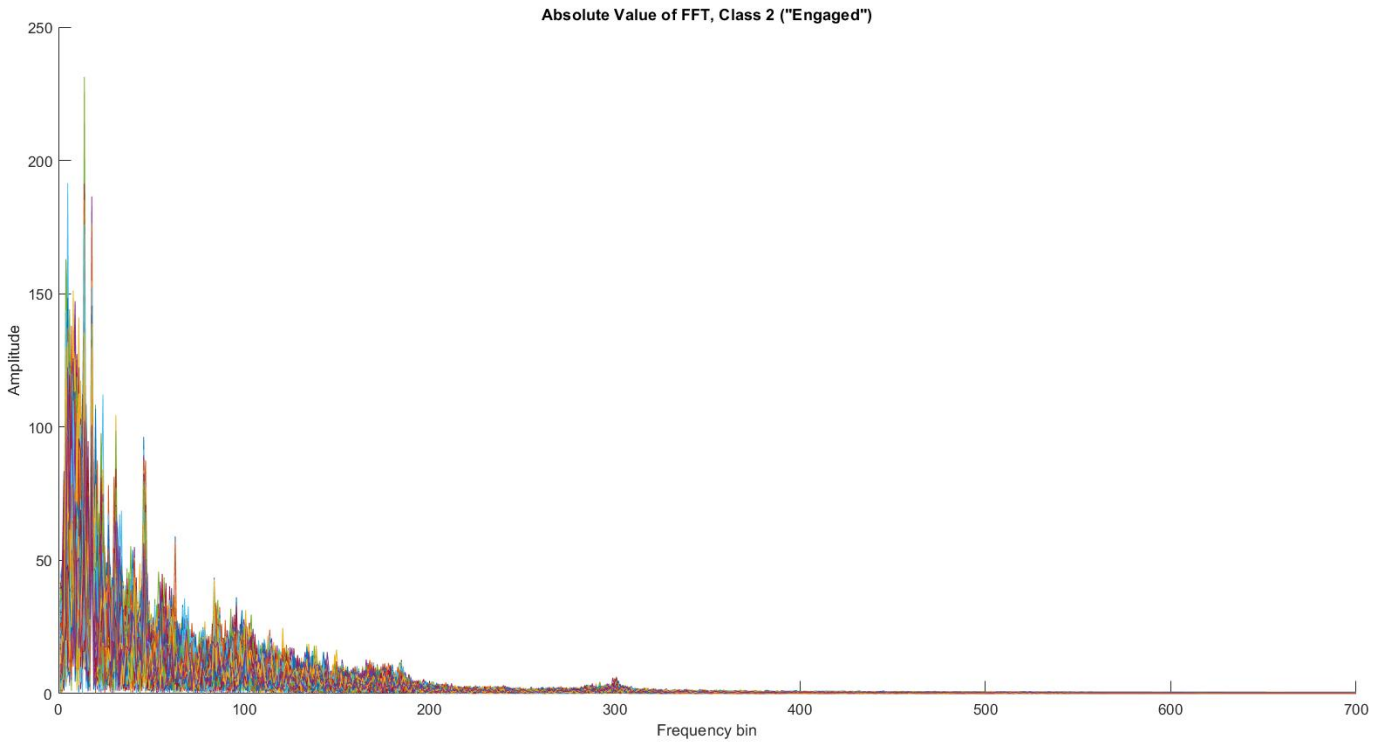

Figure 3. Absolute Value of FFT for samples from Class 2 "Engaged"

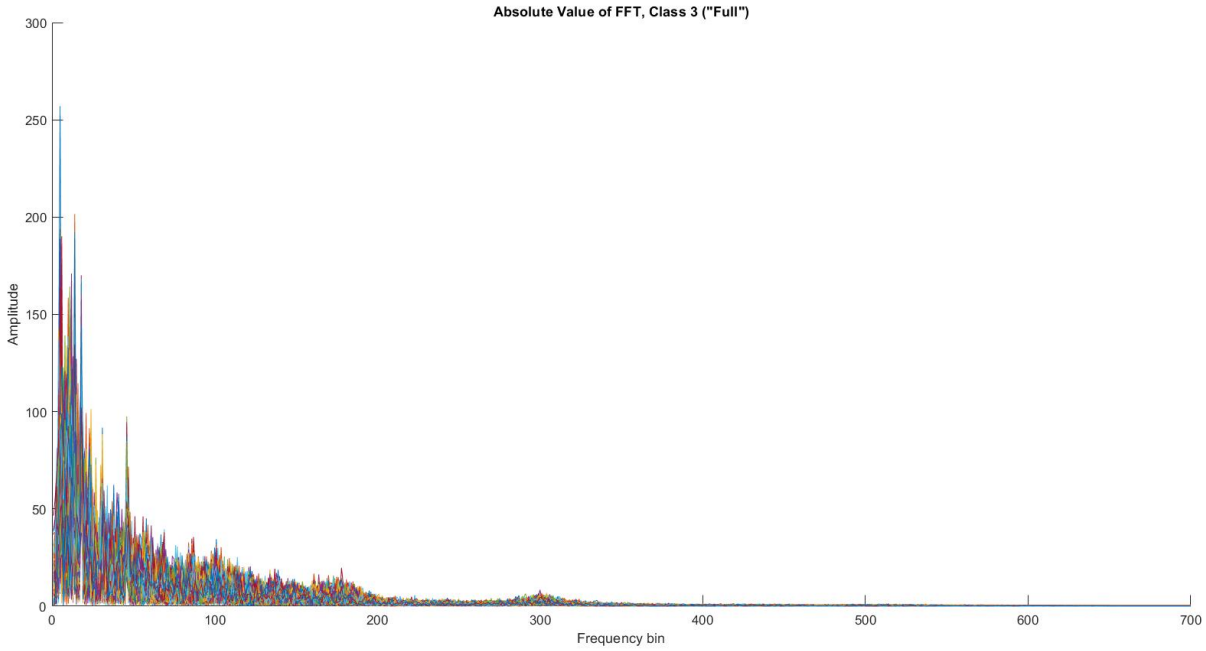

Figure 4. Absolute Value of FFT for samples from Class 3 "Full"

After a visual inspection and a literature review of potential features, eight features were selected to be used to build a classification model:

- The frequency bin of the first dominant peak (P1)
- The frequency bin of the second dominant peak (P2)
- The frequency bin of the third dominant peak (P3)
- The ratio of the magnitude of the first and second dominant peaks (P2/P1)
- The distance between the first and second dominant peaks
- The mean of the signal amplitude at specific frequency bins (200-300)
- The variance at higher frequencies (500-600)
- The center of gravity (spectral centroid) at higher frequencies (500-600)

These features were extracted for each segment and all eight features were used to create a feature vector labelled with Class 1, 2, or 3 depending on the operating mode. These feature vectors were used to create a pattern recognition neural network with the goal of using it for real-time classification.

Tools in MATLAB were used to create a feedforward, pattern recognition neural network. The network used scaled conjugate gradient backpropagation training. The data samples were selected randomly to be used for training, model validation, and testing with the following allocations: 70% of the samples were used for training, 15% of samples were used for validation, and 15% of samples were used for testing. A goal of 1×10^{-7} for the mean squared error (MSE) was set, and all other training parameters used MATLAB default values (10Appendix I – Matlab Default Training Values). The network was designed as a two-layer network with the hidden layer consisting of N neurons. Both the number of neurons (N) and the segment size (2^n) were varied to achieve optimal classification results.

The processing time of feature extraction and classification were recorded to evaluate the potential of this technique to provide real-time analysis.

6 RESULTS

When the number of neurons in the hidden layer was 5 and the segment size of samples set to 2^{11} (2048 samples per segment), 100% accuracy of classification could be achieved. Based on the sampling frequency (48,000 Hz) of the original audio, segments

of 2048 samples represented 0.0427 seconds of audio, which is an acceptable sample size for real-time monitoring. Table provides a record of segment size and the corresponding accuracy of the model. Each time a neural net is created, random samples are selected for training, validation, and testing, meaning that each new neural net may have different classification results. For each segment size, 10 neural nets were created and the accuracy for each was recorded. The average accuracy of the model created with that segment size is shown in Table . A segment size of 2048 (2^11)produced a model with 100% accuracy. Increasing the segment size increased the accuracy, but only marginally. The increase in accuracy in segment sizes beyond 2048 is considered negligible.

Table I. Segment size of sample and average accuracy

	SEGMENT SIZE					
NEURAL NET	**512**	**1024**	**2048**	**4096**	**8192**	**16384**
	Accuracy (%)					
1	88.7	93.1	99.7	99.3	100	100
2	87.5	94.6	100	100	100	100
3	90.9	94.9	100	100	100	100
4	87.7	94.6	99.3	100	100	100
5	85.9	94.6	100	100	100	100
6	88.6	97.5	99.3	100	100	100
7	86.8	94.8	100	100	100	100
8	89.1	67.7	99.7	99.3	100	100
9	88.8	93.4	99.3	98.6	100	100
10	86.9	92.9	99.3	100	98.6	100
AVERAGE ACCURACY	**88.09**	**91.81**	**99.66**	**99.72**	**99.86**	**100**

Using a segment size of 2048 resulted in 775 samples from Class 1, 550 samples from Class 2, and 645 samples from Class 3 for a total of 1970 samples which is sufficiently large to provide a high degree of confidence in the results. Of these, 1378 were used for training, 296 for validation, and 296 for testing. The relative value of each feature for each sample segment is show in Figure 5 through Figure 12.

In each Figure, the separation of classes indicates whether the feature described in that figure can be useful in determining which class it belongs to. A figure where one class has unique and predictable values compared to the other classes indicates that that feature is useful in predicting the class a sample belongs to. If values assigned to samples in different classes overlap or show an unpredictable pattern, this indicates that the feature described in the figure is likely to be less useful in determining which class the sample belongs to. These figures are for the purposes of visual assessment and understanding. Further insight through a classification model is necessary to truly determine which features are required/useful for classification.

In all Figures, Blue = Class 1, Red = Class 2, and Yellow = Class 3.

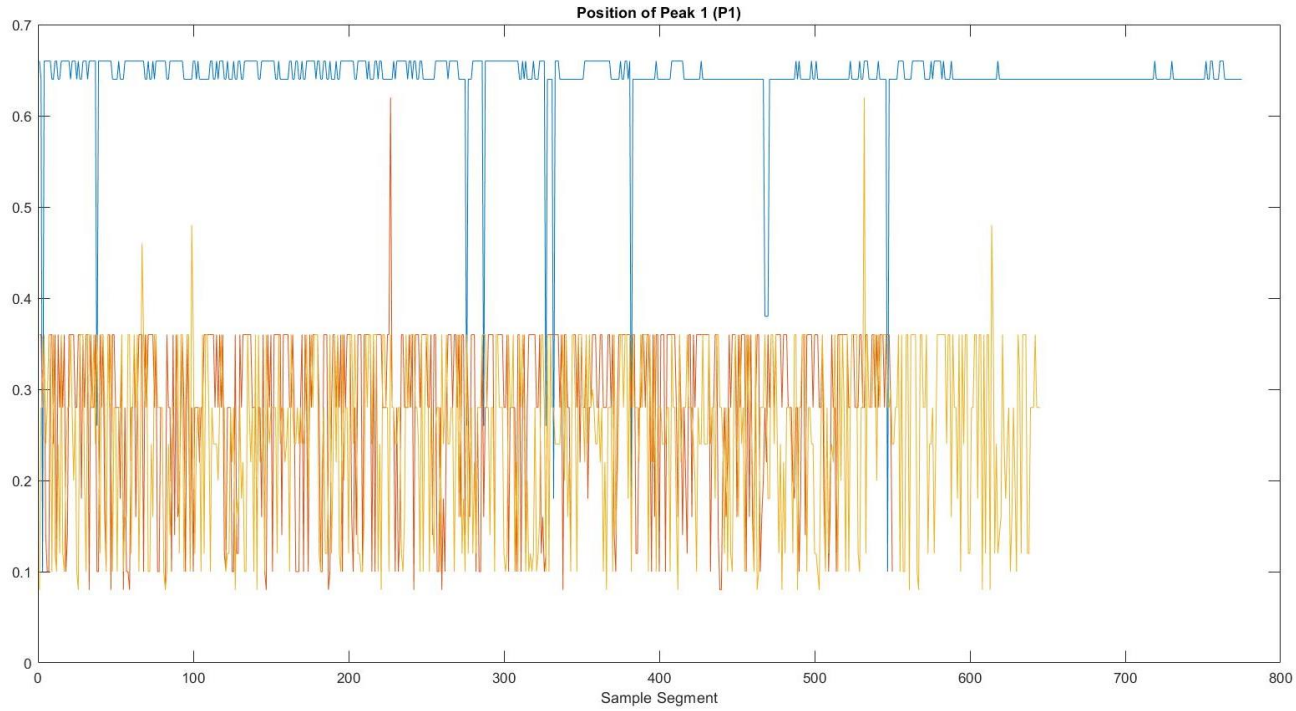

Figure 5. Relative position of P1

In Figure 5, there is a clear distinction in amplitude value between Class 1 and the remaining samples, indicating that the position of the dominant peak in the FFT analysis is an excellent indicator of which sample belong to Class 1. The lack of variation in the position of P1 among samples from Class 1 also indicates a high predictability for samples from that class, making this a strong feature for classification. However, there is a high degree of overlap between the samples representing Class 2 and Class 3, indicating that the position of the dominant peak does not provide information to differentiate between the two classes.

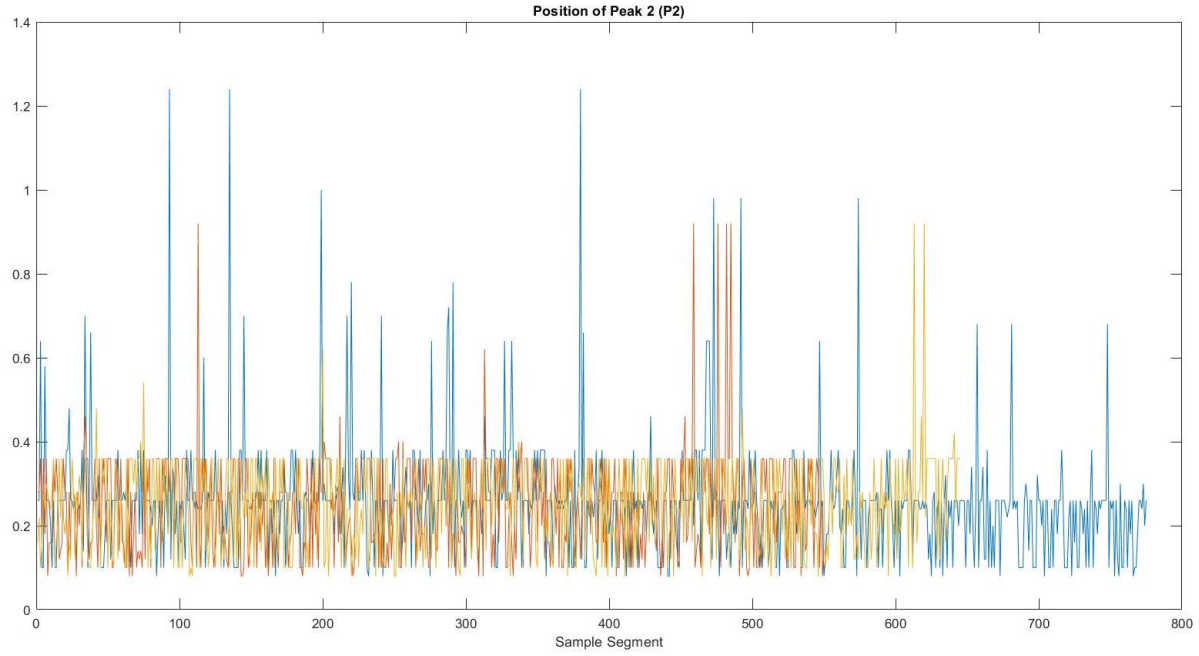

Figure 6. Relative position of P2

In Figure 6, the position of the second dominant peak (P2) was very similar for all classes with no clear distinction (visually) in the mean value between Classes 1, 2, and 3, or range of values. This indicates that the position of the second dominant peak may provide less valuable information in determining the class of an individual sample.

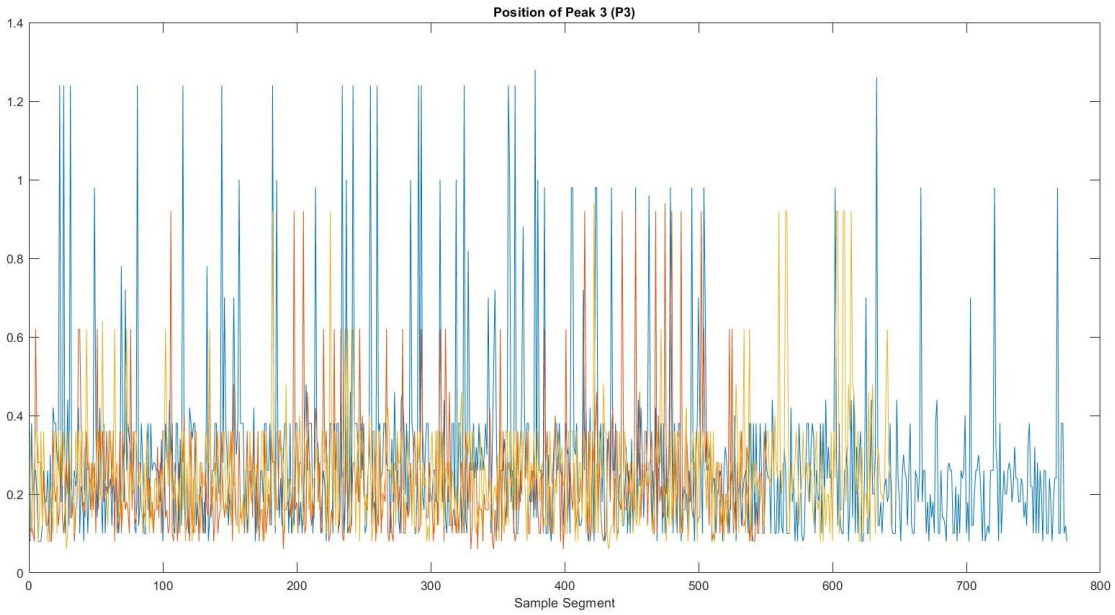

Figure 7. Relative Position of P3

Figure 7 shows similar results to Figure 6, in that values representing the location of the third dominant peak (P3) do not differ significantly between Classes 1, 2, and 3. However, it appears that Class 1 samples are more likely to be large peaks compared with Classes 1 and 2. It is difficult to visually determine how useful this feature will be in determining classification.

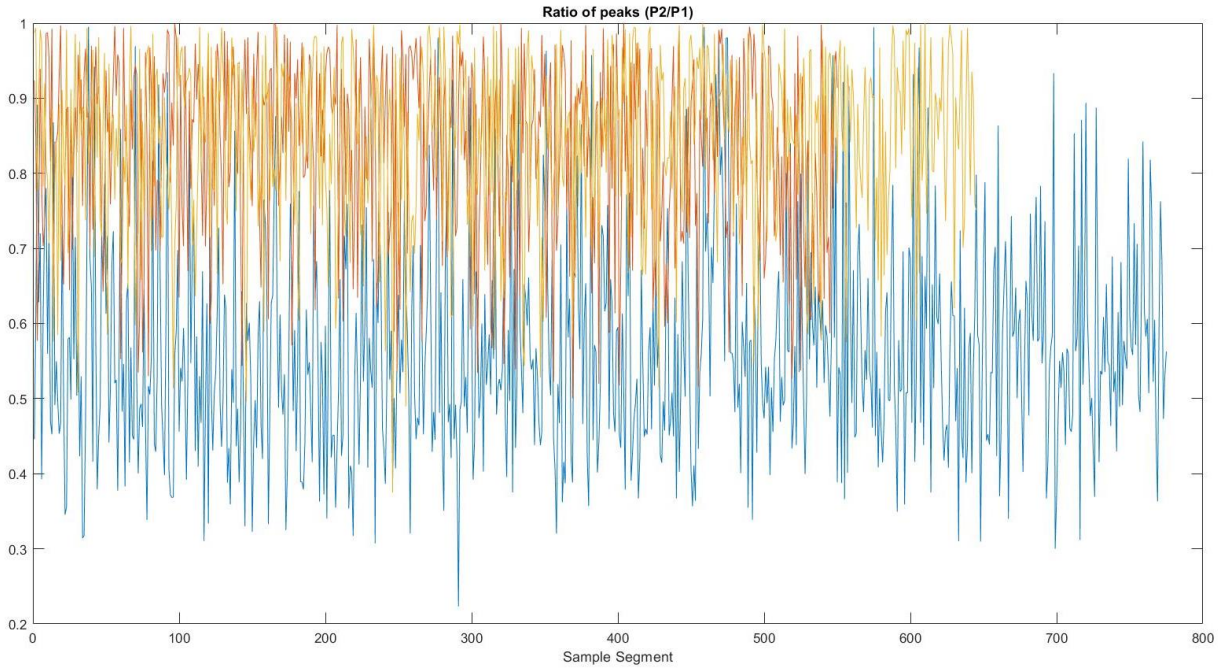

Figure 8. Ratio of P1 and P2

In Figure 8, we can see that the values of the ratio of P2 and P1 are generally much higher for samples in Class 2 and 3 compared with samples from Class 1. This indicates that the ratio of peaks (P2/P1) is likely useful in identifying samples in Class 1, but is unlikely to be useful in distinguishing between Classes 2 and 3. The large difference in value indicates that this feature may be a very strong candidate to aid in classification. However, samples beyond 200 from Class 1 indicate a fair amount of overlap with samples from Class 2 and Class 3. This indicates a wide range of possible values for P2/P1 for Class 1 samples, which reduced confidence in this feature to classify samples accurately.

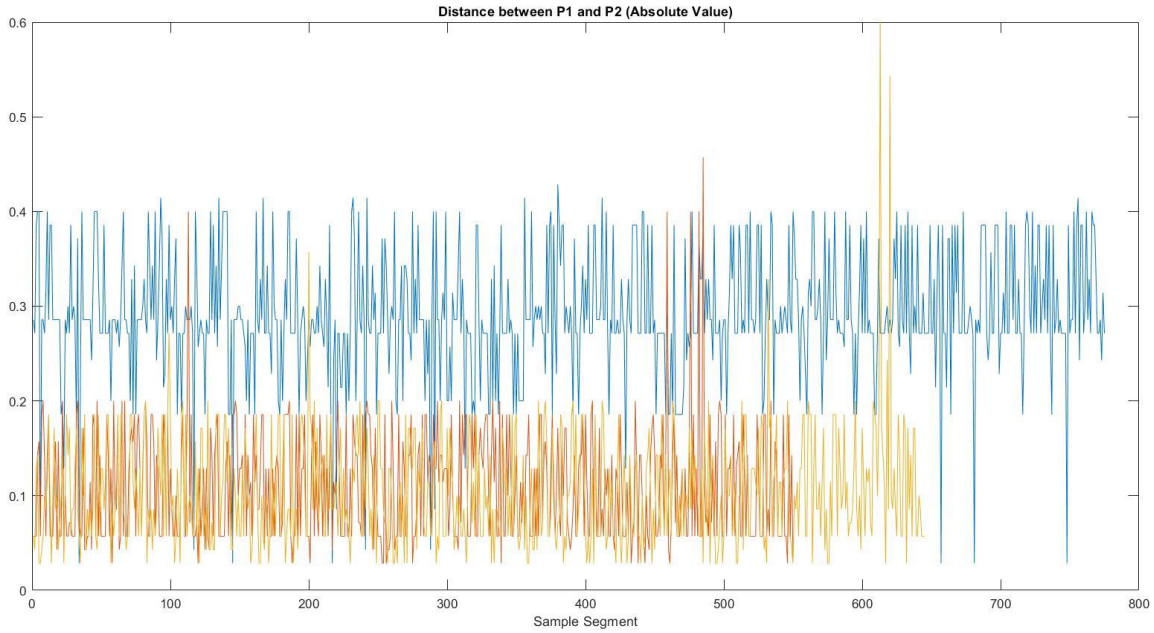

Figure 9. Distance (frequency bins) between P1 and P2

In Figure 9, the distance (frequency bins) between the first (P1) and second (P2) dominant peaks is higher for samples from Class 1 compared to Classes 2 and 3. This indicates that the distance between the first and second peaks may be a critical features to distinguish between samples from Class 1 and Classes 2 and 3. It is unlikely to be useful in classifying between Class 2 and Class 3 samples.

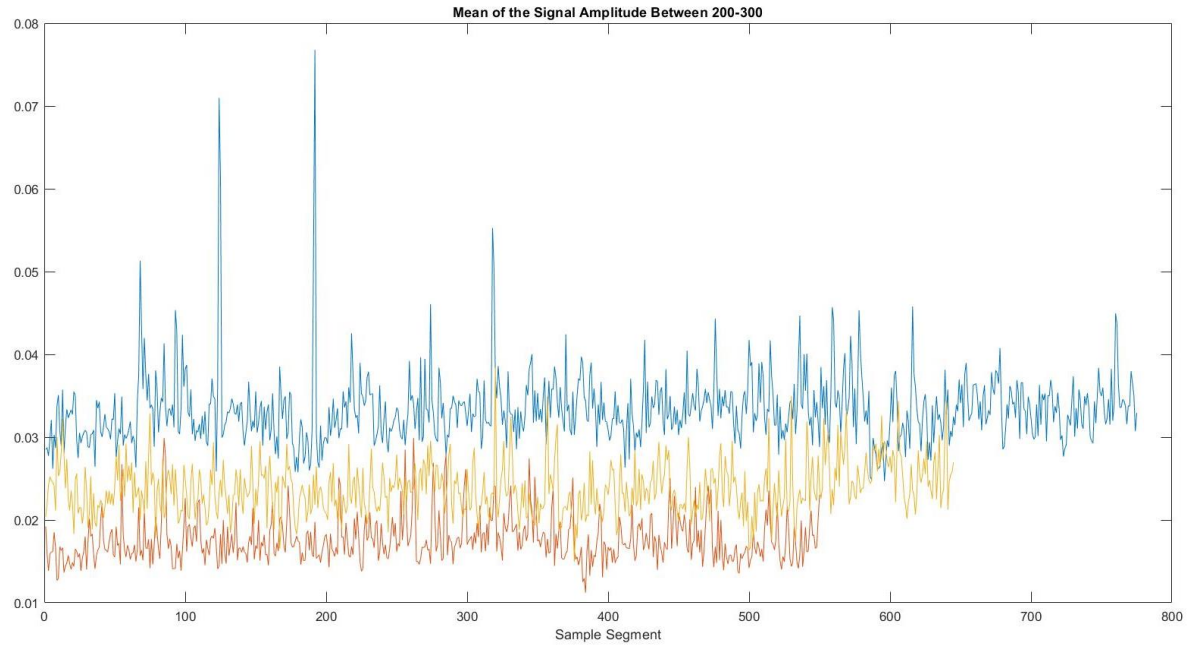

Figure 10. Mean of the signal amplitude between frequency bins 200-300

In Figure 10, the mean of the signal amplitude between frequency bins 200 and 300 fairly distinct for each class with samples from Class 1 having higher values than Class 3, and values from Class 3 higher than Class 2. This indicates that the mean of the signal amplitude for this range of frequency bins is likely a good feature for identifying signals from each class.

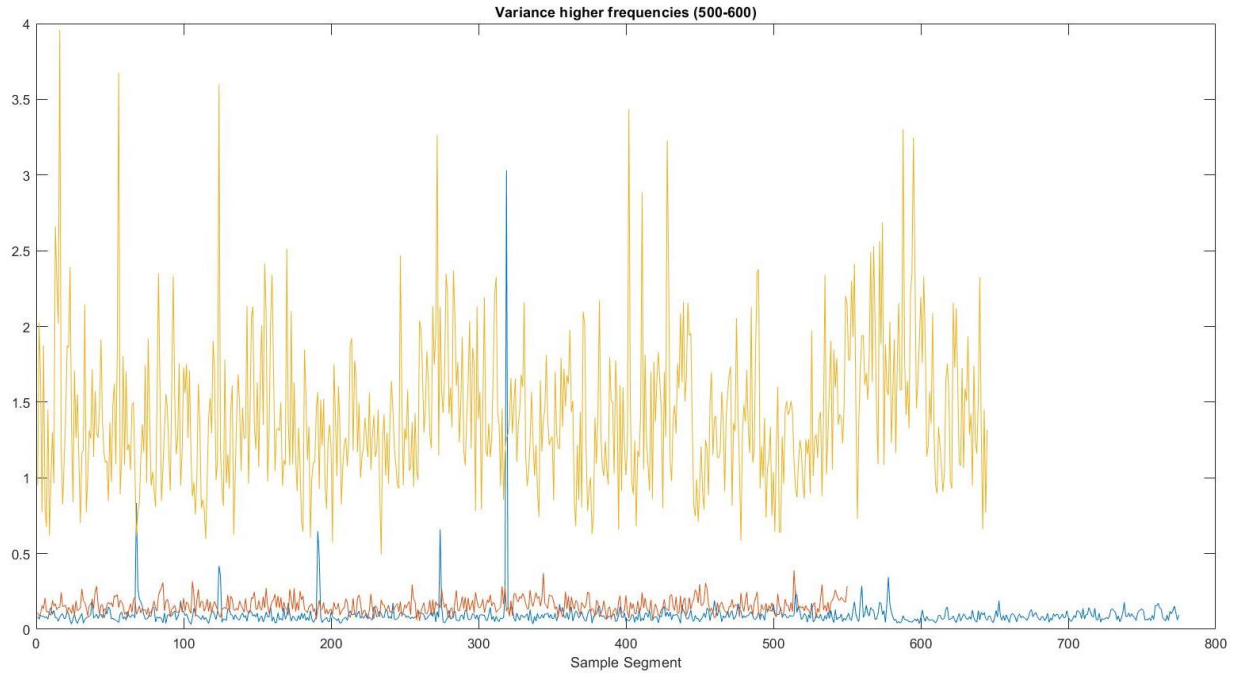

Figure 11. Variance of the signal at high frequency bins (300-400)

In Figure 11, the variance of the signal at higher frequency bins for samples from Class 3 is very clearly distinguished from samples from Classes 1 and 2. This indicates it is likely a excellent feature for identifying samples from Class 3. Visually, there is a small distinction between Classes 1 and 2 as well with samples from Class 1 generally being smaller values.

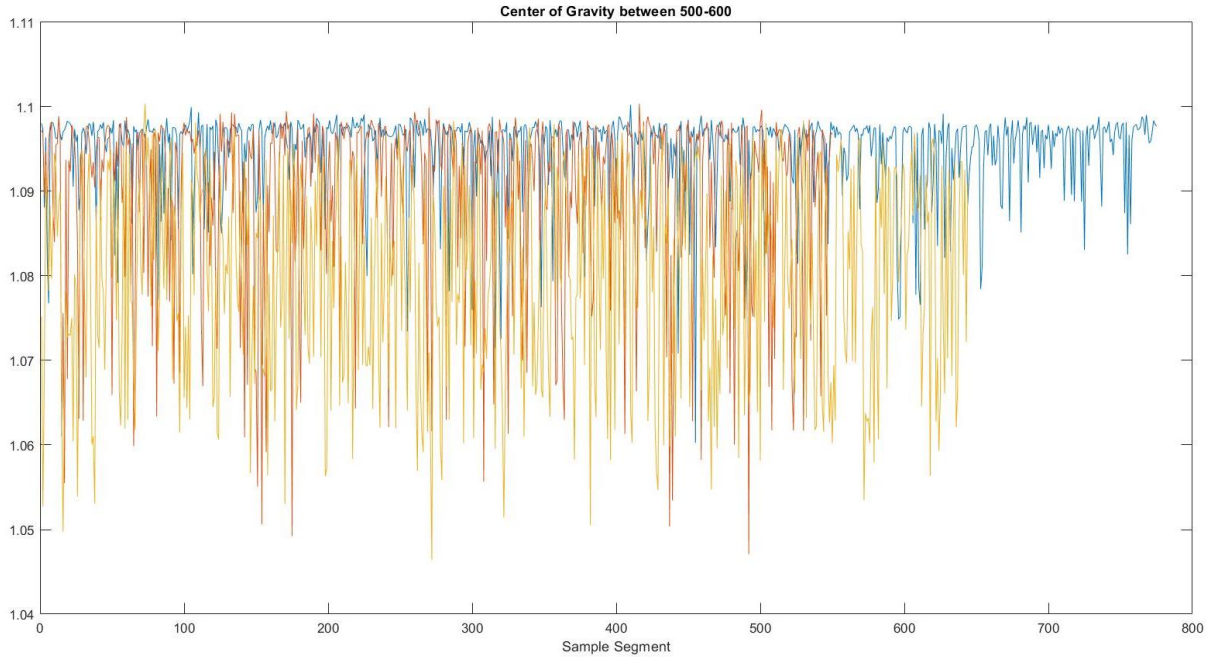

Figure 12. Location of the 'Center of Gravity' (Spectral Centroid)

In Figure 12, there is a wide range of values for the center of gravity for all samples, regardless of its Class. The value representing the center of gravity appears higher on average (and with less variation) for samples from Class 1, and there appears to be the most variation in samples from Class 3. This indicates that the center of gravity may support the classification of samples from Class 1 from Class 3.

The processing time to extract all eight features from a sample segment was between 1-2 ms, and to classify a sample averaged 50 ms. This processing time was recorded on a Lenovo T470 with an Intel Core i5-7300U CPU. This time of receiving and processing an appropriate response is much faster than the average human perceptual response

time while driving of 1-3 s depending on the situation. This indicates that this method is sufficiently fast for real-time analysis.

Features related to the dominant peaks (position, ratio, etc.) were critical in identifying Class 1 samples. However, energy and variance were critical in distinguishing Class 2 from Class 3. This balance of features indicates that high accuracy is possible. A confusion matrix is a tool that can be used to describe the performance of a classification model. There are three confusion matrices that can be examined: one for each phase of building the model (training, validation, and testing. Each matrix shows the possible output categories (Class 1, Class 2, or Class 3, and indicates how many of each Class were categorized correctly for that class, and how many were incorrectly categorized in that class.

Out of the 1970 samples used for training validation, and testing of the model, all were classified correctly (**Error! Reference source not found.**).

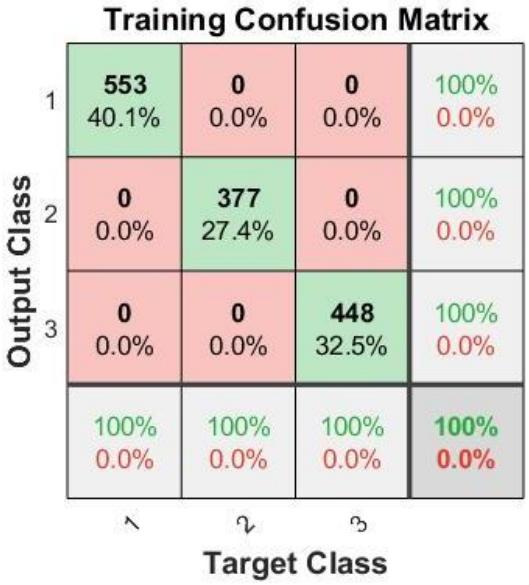

Figure 13. Training confusion matrix indicating an overall classification accuracy of 100%

Out of the 296 samples used for validation of the model (117 from Class 1, 80 from Class 2 and 99 from Class 3), all were classified correctly (Figure 14).

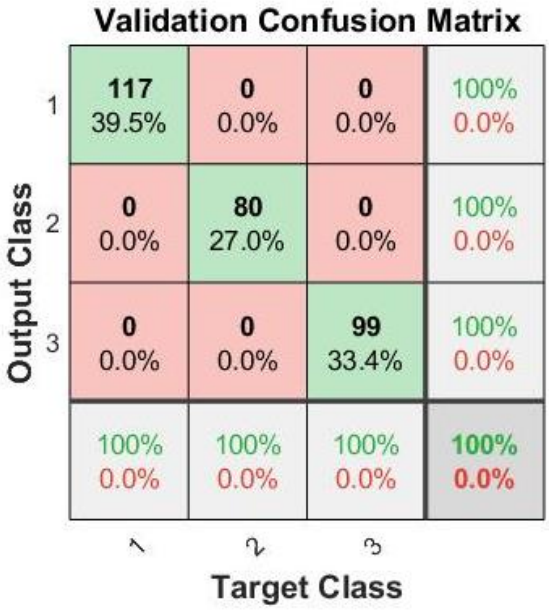

Figure 14. Validation Confusion Matrix indicating 100% accuracy

Out of the 296 samples used for testing the model (105 from Class 1, 93 from Class 2 and 98 from Class 3), all were classified correctly (Figure 15). This indicates the actual accuracy of the model in classifying samples without a label of Class 1, 2, or 3.

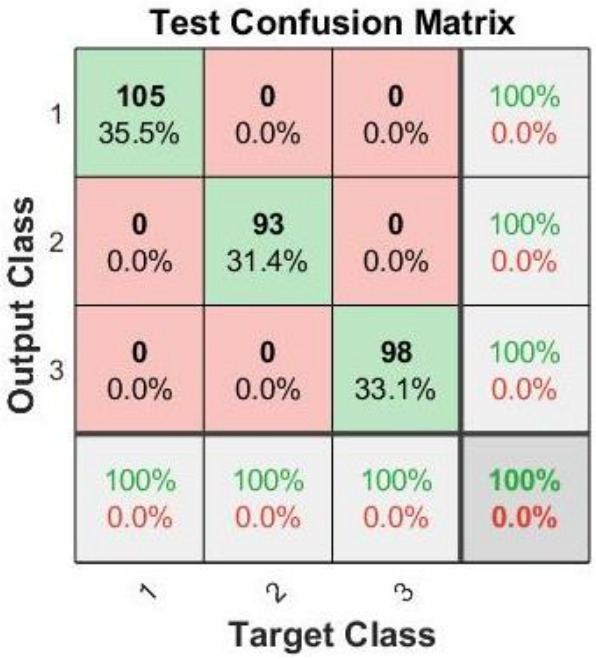

Figure 15. Test confusion matrix showing 100% accuracy

The model was trained in 93 Epochs with a mean squared error (MSE) of 0.2.4332 x 10^{-4} as shown in Figure 16. Each epoch represents an update to the weights assigned to each feature.

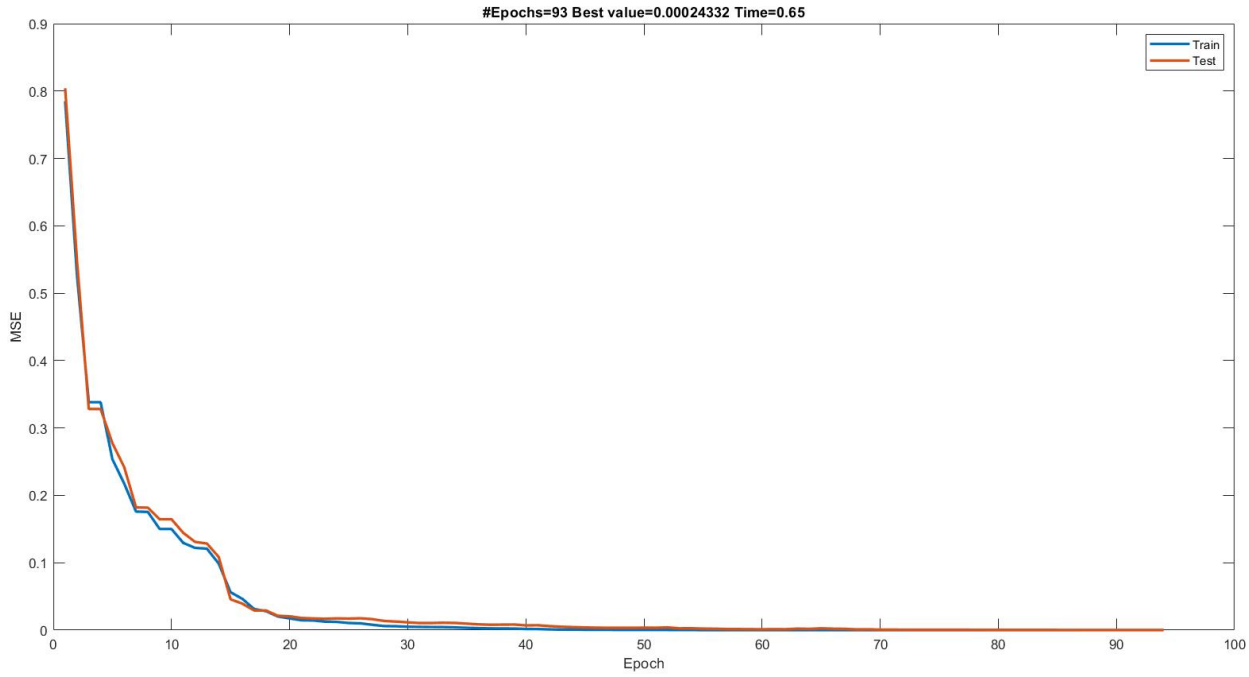

Figure 16. Mean squared error per training epoch

Cross-entropy returns a result that severely penalizes outcomes that are largely inaccurate and minimally penalizes outcomes that are nearly correct. Minimizing the cross-entropy leads to a better classifier. In this training process, a goal of 1×10^{-7} was proposed as the desired error and a cross entropy of 4.23×10^{-6} was reached.

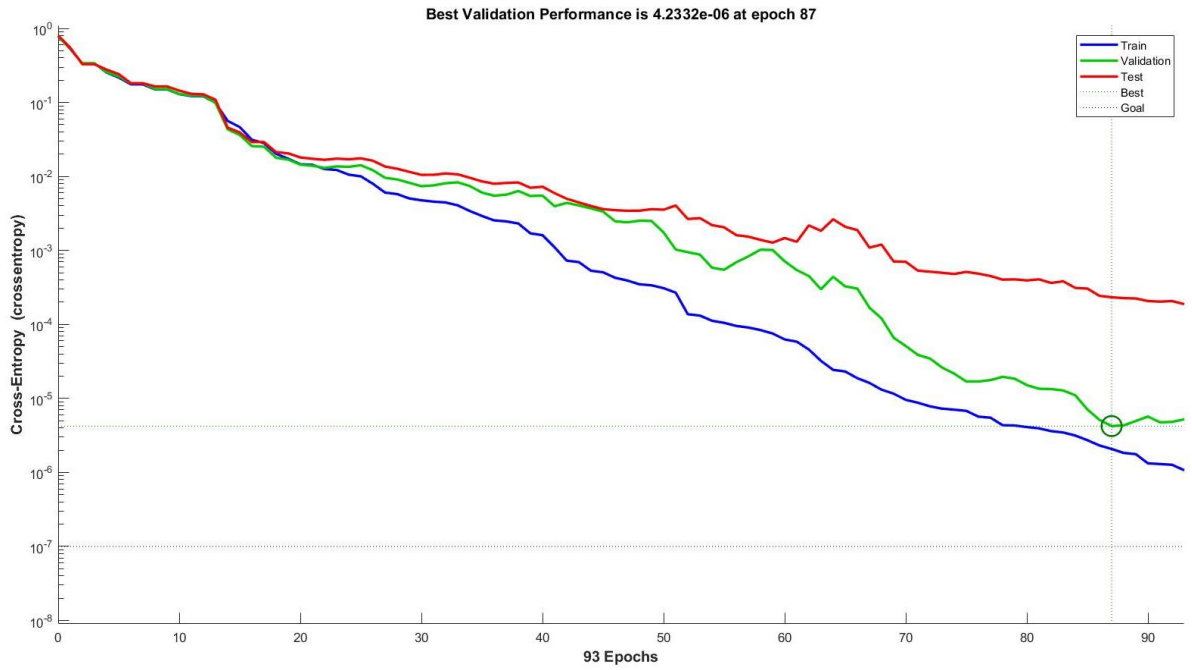

Figure 17. "Performance" on Neural Net Panel Results

7 CONCLUSIONS

In this paper, a method for creating a real-time classifier for the operating of a combine was presented. The technique used the Fast Fourier Transform to produce a power spectral density (PSD) function for a determined segment size of an audio signal. Eight features were extracted from the PSD and labelled with the operating mode (class) to create a feature vector. Feature vectors were created for three different operating modes of the combine – 1) Engine running with no threshing, 2) Engine running and threshing engaged, and 3) Engine running, and threshing engaged at 80%+ capacity.

The feature vectors were fed into a neural net using scaled conjugate gradient backpropagation training and achieved 100% accuracy. The network was designed as

a two-layer network, with the hidden layer consisting of 5 neurons. At an audio sampling rate of 48 kHz and a segment size of 2048 samples, a segmenting event takes 0.0425 seconds, indicating that this method could successfully be used for real-time analysis and control of the combine. With further refinement, this information may be used in the future to estimate and adjust the loading of the threshing system for optimal performance, reducing downtime and mechanical damage while increasing harvest efficiency.

8 RECOMMENDATIONS

Further work in this area should include a greater number of operating modes. In particular, it would be useful to identify events which may cause machinery damage or harvest delays, such as overloading the feeder house, which can cause lengthy delays due to the required shutdown and manual extraction of material. Increasing the dimensionality of the classifier would allow it to provide more information to a remote operator, an automatic controller, or provide useful data for diagnostics in machinery maintenance and repair. Future work can also include the use of spectrograms to train a classifier, different features, or combinations of features to optimally identify different operating modes, or sensor selection/placement to optimize data collection and reduce the initial and maintenance costs of this system.

Automatic feature extraction and selection could also be used with no need to visually inspect the absolute FFT for each class and manually select features. This functionality is offered in Matlab, Weka, and other machine learning tools and would likely provide and optimal set of features for efficient classification.

Unsupervised and reinforcement algorithms may also be used in the future to understand what other mechanical factors may be identified through audio classification, or for improving the performance of the algorithm under various operations conditions.

REFERENCES

Al-Fahoum, Amjed S. and Al-Fraihat, Ausilah A. 2014. Methods of EEG Signal Features Extractin Using Linear Analysis in Frequency and Time-Frequency Domains. *ISRN Neuroscience.* 13 February 2014.

Artificial Neural Network Application in the Diagnosis of Disease Conditions with Liver Ultrasound Images. **Kalyan, , et al. 2014.** 2014, Advances in Bioinformatics.

Bechar, and Vigneault, . 2016. Agricultural robots for field operations: Concepts and components. *Biosystems Engineering.* 2016, pp. 94-111.

Bishop, Christopher M. 1995. *Neural Networks for Pattern Recognition.* Oxford : Clarendon Press, 1995.

Bowling, Daniel L., et al. 2010. Major and minor music compared to excited and subdued speech. *The Journal of the Acoustical Society of America.* 05 Jauary 2010, Vol. 127, 1.

Boyce, Meherwan P. 2012. Rotor Dynamics. *Gas Turbine Engineering Handbook (Fourth Edition).* s.l. : Elsevier, 2012.

Dieleman, . 2014. Recommending music on Spotify with deep learning. *Github.* [Online] 5 August 2014. http://benanne.github.io/2014/08/05/spotify-cnns.html.

Fountas, , et al. 2015. Farm machinery managment information system. *Computers and Electronics in Agriculture.* 2015, pp. 131-138.

Gwardys, and Grzywczak, . 2014. Deep Image Features in Music Information Retrieval. *International Journal of Electronics and Telecommunications.* 2014, Vol. 60, 4, pp. 321-326.

Hidalgo, , et al. 2008. Calibration of on-demand irrigation network models. *Journal of Irrigation and Drainage Engineering.* 2008, Vol. 134, 1, pp. 36-42.

Lee, Heow Pueh, Wang, and Lim, Kian Meng. 2017. Assessment of noise from equipment and processes at construction sites. 2017, Vol. 24, 1.

Li, and Chen, . 2017. Driver Vision Based Perception-Response Time Prediction and Assistance Model on Mountain Highway Curve. *International Journal of Environmental Research and Public Health.* 2017, Vol. 14, 2.

Li, , et al. 2009. Review of research on agricultural vehicle autonomous guidance. *International Journal of Agricultural and Biological Engineering.* September 2009, Vol. 2, 3.

Mahana, and Singh, . 2015. Comparative Analysis of Machine Learning Algorithms for Audio Signals Classification. *International Journal of Computer Science and Network Security.* June 2015, pp. 49-55.

Moller, Martin Fodslette. 1993. A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning. *Neural Networks.* 1993, Vol. 6, pp. 525-533.

National Instruments. 2019. Understanding FFTs and Windowing. *National Instruments.* [Online] 19 March 2019.

http://download.ni.com/evaluation/pxi/Understanding%20FFTs%20and%20Windowing.pdf.
Nisar, , Khan, Omar Usman and Tariq, . 2016. An Efficient Adaptive Window Size Selection Method for Improving Spectrogram Visualization. *Computational Intelligence & Neuroscience.* 2016, pp. 1-13.
—. 2016. An Efficient Adaptive Window Size Selection Method for Improving Spectrogram Visualization. *Computational Intelligence and Neuroscience.* 13 July 2016.
Olson, Paul L. and Sivak, . 1986. Perception-Response Time to Unexpected Roadway Hazards. *Human Factors: The Journal of the Human Factors and Ergonomics Society.* 1986.
Peeters, . 2004. *A large set of audio features for sound description (similarity and classification) in the CUIDADO project.* Paris : Ircam, Analysis/Synthesis Team, 2004.
Perceptual effects of spectral modifications on musical timbres. **Grey, and Gordon, . 1978.** 1978, Journal of the Acoustical Society of America, pp. 1493-1500.
Prahallad, . Spectrogram, Cepstrum, and Mel-Frequency Analysis. *Speech Technology: A Practical Introduction.* [Lecture Slides]. s.l., Pensylvania : Carnegie Mellon University.
Priemer, . 1990. Signals and Signal Processing. *Introductory Signal Processing.* Chicago : World Scientific Publishing Company, 1990, pp. 1-10.
Raina, , Chakraborty, and Velankar, . 2014. Automatic Classification of Instrumental Music & Human Voice Using Formant Analysis. *International Journal of Advanced Computer Science & Technology.* April-June 2014, Vol. 2, 2, pp. 242-246.
Reid, John F., et al. 2000. Agricultural automatic guidance research in North America. *Computers and Electronics in Agriculture.* 2000, pp. 155-167.
Reid, John F., Zhang, and Noguchi, . 1999. *Agricultural Vehicle Navigation Using Multiple Guidance Sensors.* 1999.
Ripley, Brian D. 2007. *Pattern Recognition and Neural Networks.* Cambridge : Cambridge University Press, 2007.
Robust Speech Recognition in Noisy Environments Based on Subband Spectral Centroid Histograms. **Gajic, and Paliwal, Kuldip K. 2006.** 2006, IEEE Transactions on Audio, Speech and Language Processing, pp. 600-608.
Ross, , Choi, and Purves, . 2007. Musical intervals in speech. *Proceedings of the National Academy of Sciences of the United States of America.* 5 June 2007, Vol. 104, 23, pp. 9852-9857.
Triggs, Thomas J. and Harris, Walter G. 1982. *Reaction Time of Drivers to Road Stimuli.* Victoria : Monash University, 1982. 0 86746 147 0.
Use of Artificial Neural Networks in Pattern Recognition. **Basu, Jayanta Kumar, Bhattacharyya, and Kim, . 2010.** 2, 2010, International Journal of Software Engineering and Its Applications, Vol. 4.
Zero crossing rate and Energy of the Speech Signal of Devanagari Script. **Shete, and Patil, . 2014.** 1, 2014, IOSR Journal of VLSI and Signal Processing, Vol. 4, pp. 1-5.

10 APPENDIX I – MATLAB DEFAULT TRAINING VALUES

Parameter	Value	Description
net.trainParam.epochs	1000	Maximum number of epochs to train
net.trainParam.goal	0	Performa nce goal
net.trainParam.lr	0.01	Learning rate
net.trainParam.lr_inc	1.05	Ratio to increase learning rate
net.trainParam.lr_dec	0.7	Ratio to decrease learning rate
net.trainParam.max_fail	6	Maximum validation failures
net.trainParam.max_perf_inc	1.04	Maximum performance increase
net.trainParam.mc	0.9	Momentum constant
net.trainParam.min_grad	1e-5	Minimum performance gradient
net.trainParam.show	25	Epochs between displays (NaN for no displays)
net.trainParam.showCommandLine	false	Generate command-line output
net.trainParam.showWindow	true	Show training GUI
net.trainParam.time	inf	Maximum time to train in seconds