Evaluating the performance of the MIRFEE classifier plugin for PAMGuard at differentiating between whale vocalizations and anthropogenic noise in the Salish Sea

Holly T. LeBlond, Lucy S. Quayle, Harald Yurk

Science Branch, Pacific Region
Fisheries and Oceans Canada
Pacific Science Enterprise Centre,
4160 Marine Drive
West Vancouver, BC, V7V 1N6
Canada

2025

Canadian Technical Report of
Fisheries and Aquatic Sciences 3699

Fisheries and Oceans Canada | Pêches et Océans Canada

Canada

Canadian Technical Report of Fisheries and Aquatic Sciences

Technical reports contain scientific and technical information that contributes to existing knowledge but which is not normally appropriate for primary literature. Technical reports are directed primarily toward a worldwide audience and have an international distribution. No restriction is placed on subject matter and the series reflects the broad interests and policies of Fisheries and Oceans Canada, namely, fisheries and aquatic sciences.

Technical reports may be cited as full publications. The correct citation appears above the abstract of each report. Each report is abstracted in the data base *Aquatic Sciences and Fisheries Abstracts*.

Technical reports are produced regionally but are numbered nationally. Requests for individual reports will be filled by the issuing establishment listed on the front cover and title page.

Numbers 1-456 in this series were issued as Technical Reports of the Fisheries Research Board of Canada. Numbers 457-714 were issued as Department of the Environment, Fisheries and Marine Service, Research and Development Directorate Technical Reports. Numbers 715-924 were issued as Department of Fisheries and Environment, Fisheries and Marine Service Technical Reports. The current series name was changed with report number 925.

Rapport technique canadien des sciences halieutiques et aquatiques

Les rapports techniques contiennent des renseignements scientifiques et techniques qui constituent une contribution aux connaissances actuelles, mais qui ne sont pas normalement appropriés pour la publication dans un journal scientifique. Les rapports techniques sont destinés essentiellement à un public international et ils sont distribués à cet échelon. II n'y a aucune restriction quant au sujet; de fait, la série reflète la vaste gamme des intérêts et des politiques de Pêches et Océans Canada, c'est-à-dire les sciences halieutiques et aquatiques.

Les rapports techniques peuvent être cités comme des publications à part entière. Le titre exact figure au-dessus du résumé de chaque rapport. Les rapports techniques sont résumés dans la base de données *Résumés des sciences aquatiques et halieutiques.*

Les rapports techniques sont produits à l'échelon régional, mais numérotés à l'échelon national. Les demandes de rapports seront satisfaites par l'établissement auteur dont le nom figure sur la couverture et la page du titre.

Les numéros 1 à 456 de cette série ont été publiés à titre de Rapports techniques de l'Office des recherches sur les pêcheries du Canada. Les numéros 457 à 714 sont parus à titre de Rapports techniques de la Direction générale de la recherche et du développement, Service des pêches et de la mer, ministère de l'Environnement. Les numéros 715 à 924 ont été publiés à titre de Rapports techniques du Service des pêches et de la mer, ministère des Pêches et de l'Environnement. Le nom actuel de la série a été établi lors de la parution du numéro 925.

Canadian Technical Report of
Fisheries and Aquatic Science 3699

2025

Evaluating the performance of the MIRFEE classifier plugin for PAMGuard at differentiating between whale vocalizations and anthropogenic noise in the Salish Sea

by

Holly T. LeBlond, Lucy S. Quayle, Harald Yurk

Fisheries and Oceans Canada
Science Branch, Pacific Region
Pacific Science Enterprise Centre
4160 Marine Drive
West Vancouver, BC

https://doi.org/10.60825/fmjr-ze05

Cat. No. Fs97-6/3699E-PDF ISBN 978-0-660-77259-2 ISSN 1488-5379

Correct citation for this publication:

LeBlond, H.T., Quayle, L.S., and Yurk, H. 2025. Evaluating the performance of the MIRFEE classifier plugin for PAMGuard at differentiating between whale vocalizations and anthropogenic noise in the Salish Sea. Can. Tech. Rep. Fish. Aquat. Sci. 3699: iv + 28 p. https://doi.org/10.60825/fmjr-ze05

LeBlond, H.T., Quayle, L.S., and Yurk, H. 2025. Evaluating the performance of the MIRFEE classifier plugin for PAMGuard at differentiating between whale vocalizations and anthropogenic noise in the Salish Sea. Can. Tech. Rep. Fish. Aquat. Sci. 3699: iv + 28 p. https://doi.org/10.60825/fmjr-ze05

ABSTRACT

The Music Information Retrieval Feature-Extracting Ensemble (MIRFEE) classifier was developed as a plugin for PAMGuard to provide species classification for the Whistle and Moan Detector (WMD) module. When used on audio recorded by hydrophones deployed in and around the Salish Sea for the purpose of killer whale detection, the WMD is routinely triggered by humpback whale vocalizations and vessel noise. The MIRFEE classifier was thus developed as a means of reducing these false positives. MIRFEE works by extracting features from both detection metadata and audio clips taken from when detections occur. These features are subsequently used as training data for an ensemble learning model.

Pre-recorded hydrophone audio from 12 different deployment locations in the Strait of Juan de Fuca and Southern Gulf Islands across all seasons were run through the WMD. Manually-annotated detections produced by one of three classes—killer whale calls, humpback whale calls, or anthropogenic or environmental noise—were arranged into 18 subsets, and corresponding audio was subsequently run through the MIRFEE Feature Extractor. The resulting feature vectors were used to create two training sets that used different audio clip lengths, and the cross-validation results of the consequent training models are discussed.

LeBlond, H.T., Quayle, L.S., and Yurk, H. 2025. Evaluating the performance of the MIRFEE classifier plugin for PAMGuard at differentiating between whale vocalizations and anthropogenic noise in the Salish Sea. Can. Tech. Rep. Fish. Aquat. Sci. 3699: iv + 28 p. https://doi.org/10.60825/fmjr-ze05

RÉSUMÉ

Le classificateur ensembliste d'extraction des caractéristiques de recherche d'informations musicales (MIRFEE en anglais) a été développé sous la forme d'un plugiciel pour PAMGuard pour fournir la classification des espèces au détecteur de sifflements et gémissements (Whistle and Moan Detector en anglais). Lorsqu'il est utilisé sur les fichiers audio enregistrés par les hydrophones déployés dans et autour de la mer des Salish pour la détection des orques, le détecteur est régulièrement déclenché par les vocalisations des baleines à bosse et le bruit des navires. Le classificateur MIRFEE a donc été développé pour réduire le nombre de ces faux positifs. MIRFEE extrait les caractéristiques des métadonnées des détections ainsi que des extraits audio correspondant à chaque détection. Ces caractéristiques sont ensuite utilisées comme données d'apprentissage pour un modèle ensembliste.

Des fichiers audio d'hydrophones provenant de 12 déploiements différents dans le détroit de Juan de Fuca et du sud des îles Gulf, comprenant toutes les saisons, ont été analysés par le détecteur. Les annotations correspondant à chacune des trois classes de détection—vocalisations des orques, vocalisations des baleines à bosse, ou le bruit anthropique ou environnemental—ont été organisées en 18 sous-ensembles, et les fichiers audio correspondants ont ensuite été traités par l'extracteur de caractéristiques MIRFEE (Feature Extractor en anglais). Les vecteurs de caractéristiques qui en résultent ont été utilisés pour créer deux jeux de données d'apprentissage avec différentes durées d'extraits audio, et les résultats de validation croisée des modèles d'apprentissage qui en découlent sont discutés dans ce rapport.

INTRODUCTION

Passive acoustic monitoring (PAM) is a tool for detecting vocalizing cetaceans and has been used effectively to mitigate disturbances and determine habitat presence, spatial and temporal use of habitats, and animal movement (Zimmer 2011, Sousa-Lima et al. 2013, Verfuss et al. 2018). Its effectiveness has substantially increased in the last few decades due to the development of digital acoustic recording devices that can collect, store, and transmit large amounts of acoustic data (Sousa-Lima et al. 2013). This has created the ability to acoustically monitor underwater environments continuously for long periods of time. The large volume of acoustic data created by this technological advancement has also created the need for automated tools to efficiently analyse these large-scale datasets. This includes the development of methods for detection and classification of sounds based on their respective sources (e.g., whale, vessel, natural ambient sound, etc.) (Gibb et al. 2019).

In an automated detector-classifier (DC) system, a detection algorithm scans through audio data, which is often converted into spectrographic imagery, for possible marine mammal vocalizations, while a classifier model attempts to determine what actually triggered the detector. Some DC systems use the energy of harmonics in specific frequency bands exceeding a set threshold for call detection. The spectrographic image quality of the vocalizations is influenced by the signal-to-noise ratio (SNR) and the distance of the caller from the receiving hydrophone. Higher sound frequencies attenuate faster with distance than lower ones, often resulting in higher harmonics disappearing from the image while lower harmonics remain visible. Higher ambient noise levels also decrease the SNR, resulting in calls being masked. As a result, the classifier performance of a DC system usually decreases with decreasing SNR and increasing distance between caller and receiver (Binder and Hines 2019). Modifications to the input audio data such as adding frequency filters and noise reduction procedures may be useful for increasing the signal strength over the acoustic background.

A variety of automated DC systems exist and those applying machine learning algorithms have become increasingly accurate at detecting cetacean calls. In recent years, deep learning models trained on spectrogram images of sound files have become more widely used (Usman et al., 2020). Alternative DC systems extract relevant acoustic information from audio data to make classifications based off of patterns and differences between classes found in this information. Either method requires the creation of training sets that consist of human-annotated audio data. After training a classifier model to achieve a certain level of precision and recall on either a separate testing set or through cross-validation, the model is deemed effective for use on new acoustic data.

While DC systems can be useful for determining the presence of certain species in an aquatic environment, there are often additional challenges: underwater soundscapes are dynamic, as sound propagation is highly variable and is influenced by environmental factors, such as bathymetry, water temperature, salinity, and pressure, all of which affect the sound speed of signals (Vagle et al. 2021). Furthermore, the target species is often one of multiple simultaneously-present sound sources at any given location. Many cetacean species produce sounds in overlapping frequency ranges, so it cannot always

be assumed that every automated detection within a certain frequency range was produced by the target species. For this reason, the automated detections and species classifications often need to be verified by a human listener (Socheleau et al., 2015), as some DC systems may produce a disproportionately large number of false positives in certain environments. This process is time-consuming and often requires multiple human annotators to manually process large detection data sets. The ability to detect whales quickly is very important when classified detections are used to mitigate disturbances that pose threats to the health of a target species, such as being struck by vessels, getting entangled in fishing gear, or entering an area that is contaminated with oil sheens after a spill.

In 2013, Ness and colleagues (2013) released the Orchive—a collection of over 20,000 hours of machine learning-annotated hydrophone audio recorded for killer whale research by OrcaLab, located on Hanson Island off of northern Vancouver Island near Port McNeill, British Columbia (Ness et al., 2013). Ness and colleagues used a custom-made machine learning classifier that extracted the Mel-frequency cepstral coefficients (MFCCs), spectral centroid, spectral rolloff, spectral flux, and zero-crossing rate from frames of audio clips and constructed feature vectors out of the means and standard deviations of these features. These features are commonly used in the field of music information retrieval (MIR) (Alías et al., 2016) for tasks such as musical instrument recognition (Benetos et al., 2006) and genre classification (Tzanetakis & Cook, 2002). Therefore, one might infer that these techniques may be useful in bioacoustics for identifying animal vocalizations through the analysis of timbre and other audio qualities. This technique was used to separate killer whale calls from background noise and voice notes from researchers, and then was further used to separate between a selection of common northern resident killer whale (NRKW) call types. When trained on a support vector machine (SVM) classifier, this resulted in accuracies of 96.5% and 98.5% for each respective task. Following this, a deep-learning model produced with the ORCA-SPOT toolkit (Bergler et al., 2019) achieved a time-based precision value of 93.2% when attempting to automatically segment killer whale calls from background noise in the entirety of the Orchive. The acoustic environment surrounding OrcaLab, however, is likely substantially different from those of the areas frequented by Southern Resident Killer Whales (SRKW) between Vancouver and the mouth of the Strait of Juan de Fuca, which experience significantly more commercial and recreational vessel traffic, and the Orchive audio was recorded only when a live monitoring crew noticed the presence of killer whale vocalizations (Bergler et al., 2019). Additionally, it was noted that the ORCA-SPOT deep-learning model had a tendency to produce false positives in certain instances of tonal vessel noise (Bergler et al., 2019).

This report describes initial efficacy testing of the Music Information Retrieval Feature-Extracting Ensemble (MIRFEE) classifier—a publicly-available PAMGuard plugin. Designed to be used in conjunction with the Whistle and Moan Detector (WMD) in a DC system, MIRFEE uses audio-based feature extraction in combination with features calculated from detection metadata to classify WMD detections. The classifier attempts to adapt such methods to an acoustically-dynamic environment with high potential for encounters of different triggering sound sources to occur simultaneously. While this module was developed primarily for the purpose of differentiating between killer whale

calls, humpback whale calls, and tonal vessel noise off of southern Vancouver Island, it was designed to be adaptable for use with any sound source that can trigger the WMD.

Ecological background

Killer whales (*Orcinus orca*) and humpback whales (*Megaptera novaeangliae*) both occur along the coast of the northeast Pacific Ocean. Their distribution in this region coincides with international shipping lanes, along with a heavy presence of commercial and recreational vessel activity (Vagle et al., 2021). This overlap raises concerns about acoustic and physical disturbance, including potential vessel strikes (Thornton et al., 2022). The three populations of killer whales that occur in the study area include two populations of the fish-eating Northern Resident and Southern Resident ecotypes, and a mammal-eating Transient ecotype (Bigg's). The Southern Resident Killer Whales (SRKWs) are listed as Endangered and both Northern Resident Killer Whales (NRKWs) and Bigg's Killer Whales are listed as Threatened under the Species at Risk Act (Ford et al., 2017; Fisheries and Oceans Canada, 2007). Killer whales actively use sound to navigate, socialize and forage (Ford, 1989), and have acute hearing which allows them to maintain contact with group members while spread out over large areas, especially under quiet conditions (Miller, 2006). Killer whales produce echolocation clicks for navigation and foraging (Au et al., 2004), whistles appear to be used by group members to communicate over short distances (Thomsen et al., 2002; Riesch & Deecke, 2011), while mono-phonic or bi-phonic pulsed calls are used in social communication over distances of several kilometers (Filatova et al., 2013).

Humpback whales produce both social calls and songs that can propagate over long distances (Payne & McVay, 1971; Molder et al., 2024). The British Columbia coast is considered to be a summer feeding ground for humpback whales (McSweeney et al., 1989; Ford et al., 2010), so the songs that are produced are often abbreviated and considered to be "training songs" compared to the fully formed songs produced in the breeding ground. Social calls are often lower frequency than sounds produced during songs and do not follow a particular pattern (Dunlop et al., 2008). The distinctiveness of killer whale and humpback sounds and their ability to propagate over several kilometers makes them a good target candidate for automated detection, which has become an essential tool in determining variations in temporal and spatial distribution of whales.

The magnitude of disturbance is often directly related to the behaviours of the target species; for example, most odontocetes (toothed whales) are highly mobile and move relatively fast (1 to 6 meters per second (Rohr et al. 2002)), which makes them less prone to being struck by large vessels, but not enough to substantially reduce the risk of strikes from smaller vessels such as pleasure craft. On the other hand, mysticetes (baleen whales) tend to be less mobile and often remain in the same location for longer periods of time, making them more prone to ship strikes. The time needed to detect and classify whale species and alert nearby vessel operators of their presence is therefore crucial.

PAMGuard

PAMGuard, an open-source acoustic analysis program developed by researchers from the Sea Mammal Research Unit (SMRU) at the University of St Andrews (Gillespie et al., 2008), is commonly used for passive acoustic monitoring (PAM) tasks. One of these tasks, performed by the Whistle and Moan Detector (WMD) module within the program, is to mark harmonic contours that exceed a magnitude threshold in a spectrogram stream, potentially signalling the presence of cetacean vocalizations (Gillespie et al., 2013). Background noise is reduced using a median filter, average subtraction, and, optionally, Gaussian kernel smoothing to reduce background noise (Gillespie et al., 2013).

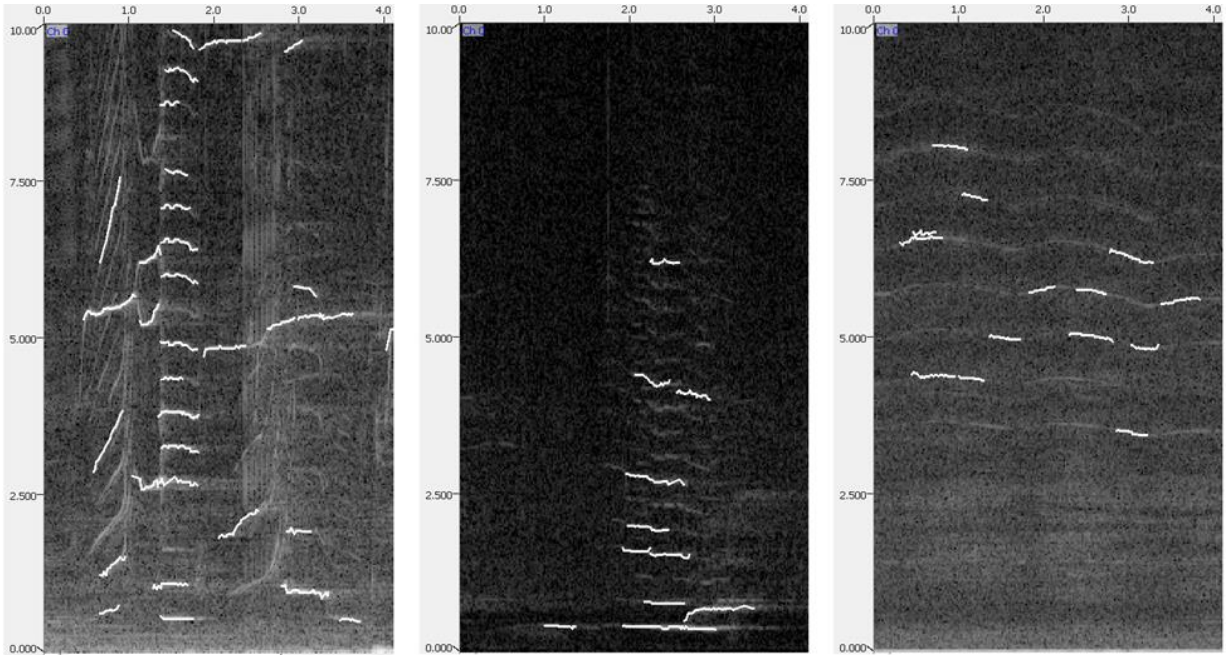

Figure 1. Screenshots from PAMGuard displaying spectrograms with Whistle and Moan Detector contours (bright white lines) triggered by killer whale calls (left), a humpback whale's moan (centre), and, undesirably, tonal propeller noise from a passing motor vessel (right). Each white line (also known as a contour) represents a single independent detection.

The WMD is effective at detecting killer whale and humpback whale calls while filtering out constant tones produced by vessels due to noise reduction. However, it struggles to ignore vessels that produce wavering pitch or intermittent tones (Figure 1). Humpback whale calls can contain harmonics that overlap with the frequency range of killer whale calls, and tonal vessel noise can occur across the frequency spectrum, so detections from only one of these sources may not be completely isolated using only a frequency threshold.

There are two pre-existing machine learning classifiers for WMD detections that come with PAMGuard: ROCCA, which includes models for differentiating between specific odontocete species, but currently does not allow users to produce their own training data (PAMGuard, no date), and the Whistle Classifier, which does allow for the latter.

The Whistle Classifier was found to be capable of differentiating between encounters from four odontocete species in the Polar Atlantic, achieving recall rates from 92.3% to 97.4% across species (Gillespie et al., 2013). However, this classifier was designed to extract features and make predictions over sizeable quantity-defined groupings of detection fragments rather than making predictions for time-defined groups of detections representing individual “calls”. This approach may not be desirable in situations where multiple sound source encounters occur simultaneously, as is common in the Salish Sea.

METHODOLOGY

Plugin overview

The MIRFEE classifier was developed as a PAMGuard plugin to provide the WMD with sound source identification capabilities. The plugin consists of several modules, those relevant to this study being the Feature Extractor and the Test Classifier.

The Feature Extractor module produces audio clips where WMD detections occur and sends them to a Python script to extract audio features. The extracted features are incorporated into feature vectors used for training and testing data. “Feature vector” refers to the array of feature values produced by all the selected feature extraction algorithms that represents an individual detection. The Feature Extractor can take a WMD data stream as input, or it can use pre-existing annotated WMD data and extract clips at the corresponding timestamps without the presence of the WMD in the configuration. The Feature Extractor arranges detections into “call clusters”, which are separated by a specified time threshold. While the classifier predictions are performed on each individual detection, the average probability scores across each detection in a cluster provides the overall prediction for the call cluster, which are what are used for calculating the precision, recall, and accuracy percentages in this report. The purpose of this was to provide predictions at the call level rather than on every individual detection produced by, for example, stacked harmonics, although it should be noted that some clusters can contain multiple separate calls that happen to occur closely together.

Most feature extraction algorithms are performed on a short-time Fourier transform (STFT, the magnitude of which produces a spectrogram) of each audio clip, or are performed on successive frames of the audio time-series. These algorithms output an array of values that represent each time frame, so the actual value used in the feature vector is often the mean or standard deviation of this array. Conversely, all header data features are single constants taken directly from detection metadata and are used as-is.

The following features are taken directly from WMD detection metadata and do not change with audio settings:

- Contour duration, in milliseconds
- Contour frequency (minimum or maximum), in hertz
- Contour frequency range, in hertz

- Contour frequency “slope”, in hertz per second – the absolute frequency range divided by the contour duration, and is always a positive number

The following features are calculated from contour slice data (i.e. the frequencies and timestamps of each “pixel” in a contour), and also do not change with audio settings:

- Contour slice frequencies, in hertz
- 1st derivative of contour slice frequencies, in hertz
- 2nd derivative of contour slice frequencies, in hertz
- Contour “start-to-end slope”, in hertz per second – “as-the-crow-flies” slope in terms of time and frequency between the first and last slices of the contour, and can be either positive or negative

The following features were used with the Orchive (Ness et al., 2013) and are extracted from the produced audio clips using the Librosa Python library (McFee et al., 2015):

- Mel-frequency cepstral coefficients (MFCCs)
- Spectral centroid, in hertz
- Spectral rolloff, in hertz
- Spectral flux (onset strength)
- Zero-crossing rate

The following features are also provided by the Librosa library (McFee et al., 2015), and were added experimentally:

- Root mean square (RMS), in decibels
- Spectral bandwidth, in hertz
- Spectral contrast
- Spectral flatness
- “Spectral magnitude”, in decibels – the magnitudes from a specified range of short-time Fourier transform (STFT) frequency bins

The following features were calculated from formants, which are commonly used to acoustically represent vowels in human speech (Kent & Vorperian, 2018), but are adapted here for experimental use with whale calls:

- Frequency of a specific formant, in hertz

- Ratio between the frequency of a selected formant and the first formant

The formants were calculated using Librosa's linear predictive coding (LPC) function (McFee et al., 2015) and adapting a formula from MathWorks.com (MathWorks, no date), which in turn was adapted from formulae by Snell & Milinazzo (1993). The LPC order (number of poles) was derived by dividing the sampling rate by a quarter of the specified maximum-expected fundamental frequency, as per an equation described in the user manual for IRCAM's AudioSculpt 3.0 software (IRCAM, no date).

Lastly, the following features use the Praat pitch-tracking algorithm from the Parselmouth Python Library (Jadoul et al., 2018). Note that these features are only calculated from frames where the tracker found a discernable pitch, and the features involving harmonics use a series of FFTs with a length equal to the sampling rate in place of the aforementioned STFT:

- Praat fundamental frequency, in hertz
- Total harmonic distortion (fundamental; THD_F) – generally used as a metric for determining sound reproduction quality in audio systems (Westerhold, 2022), but is re-appropriated here as a measurement of brightness (how much emphasis there is on the upper harmonics)
- "Harmonics-to-background ratio" – the mean magnitude of FFT bins corresponding to approximated harmonics divided by the median magnitude of the whole FFT frame
- Harmonic centroid – the spectral centroid divided by the found fundamental frequency, effectively the "centroid" as a harmonic number (Hermes et al., 2016)
- "Bin-exclusive harmonic centroid" (BEHC) – an alternative version of the harmonic centroid that only takes the FFT bins corresponding to approximated harmonics into account

When calculating BEHC on a single frame, a histogram is created where the number of the strongest harmonic is appended to a set 100 times, and the number of every other harmonic is appended to said set n times:

$$n = \text{floor}\left(100 \cdot \frac{magnitude\ of\ harmonic}{magnitude\ of\ strongest\ harmonic}\right) \quad (1)$$

The mean of the histogram from each frame fills the resulting array. If the Praat algorithm was unable to find any fundamental frequency, then the array defaults to consisting of a single 1. The purpose of this version of the harmonic centroid was to avoid background noise between harmonics being factored into the equation.

Additionally, the module includes options to specify the length of the produced audio clips (or, alternatively, to have each clip be as long as their corresponding contour), and

to specify the length and hop size of the STFT produced for features that use it as an input. It also provides the option to apply high- or low-pass Butterworth filters to the audio, as well as the option to apply noise reduction on the clips. The noise reduction algorithm works as follows:

$$N[\omega] = \frac{\sum_{\tau=0}^{L-1} |X_N[\tau, \omega]|}{L} \cdot S \tag{2a}$$

$$X_{NR}[\tau, \omega] = \begin{cases} X_O[\tau, \omega] \cdot \dfrac{|X_O[\tau, \omega]| - N[\omega]}{|X_O[\tau, \omega]|}, if\ N[\omega] < |X_O[\tau, \omega]| \\ 0 + 0i, otherwise \end{cases} \tag{2b}$$

Where equation 2a is the average frequency band magnitude at frequency bin ω, and equation 2b is the signal amplitude at time frame τ and frequency bin ω after undergoing noise reduction. X_O is the STFT of the unaltered audio clip where a detection occurs; X_N is the STFT of a "noise profile" clip taken slightly before the first detection in the cluster, with a rectangular window function applied to it; L is the time length of X_N, in frames; S is a specified scalar; and X_{NR} is X_O after noise reduction. In simpler terms, a separate clip is taken just before the first detection in a call cluster, the mean magnitude is calculated from each of its frequency bands, and each STFT bin in the original clip is proportionally scaled down such that the aforementioned mean magnitudes of corresponding frequency bins are subtracted from the original STFT's magnitudes. The STFT is then converted back to an audio time-series via an inverse short-time Fourier transform (ISTFT) for features that use a time-series as input. The purpose of this specific formulation of noise reduction is to remove unwavering tonal sounds that occur both during and before the actual detection, such as when a whale is vocalizing with monotone vessel noise in the background.

After processing, saved feature vector data produced by the Feature Extractor can then be matched with corresponding manually-annotated WMD data to create training sets using the MIRFEE Training Set Builder module. A training set can then be loaded into one of the MIRFEE classifier modules–either the Live Classifier for live processing of data directly from the Feature Extractor, or the Test Classifier for cross-validation of the training set–where it is fitted into an ensemble learning model from the Scikit-Learn Python library (Pedregosa et al., 2011). The classifier predictions are performed on each detection, with the overall prediction for the cluster being the class with the highest average prediction score across all the detections. This is done in order for the classifier prediction to represent the entire call, rather than having a large number of individual predictions for every WMD detection produced by stacked harmonics.

Data collection

Digital audio recordings were retrieved from passive acoustic recorders deployed in various locations off of southern Vancouver Island between May 2018 and March 2023 (Figure 2). These recorders were either SoundTrap high-frequency autonomous recorders (ST600-HF, Ocean Instruments, Auckland, New Zealand, www.oceaninstruments.co.nz), Autonomous Marine Acoustic Recorders (AMAR,

JASCO Applied Sciences, Halifax, NS, Canada, www.jasco.com), or TR-ORCAs (Turbulent Research, Bedford, NS, Canada, www.turbulentresearch.com), and were deployed at water depths of 50 to 100m in areas where whales had previously been observed during visual surveys (Thornton et al. 2022). The collected digital audio data were arranged into 18 subsets for cross-validating the resulting training set, arranged by deployment location and time frame (Table 1).

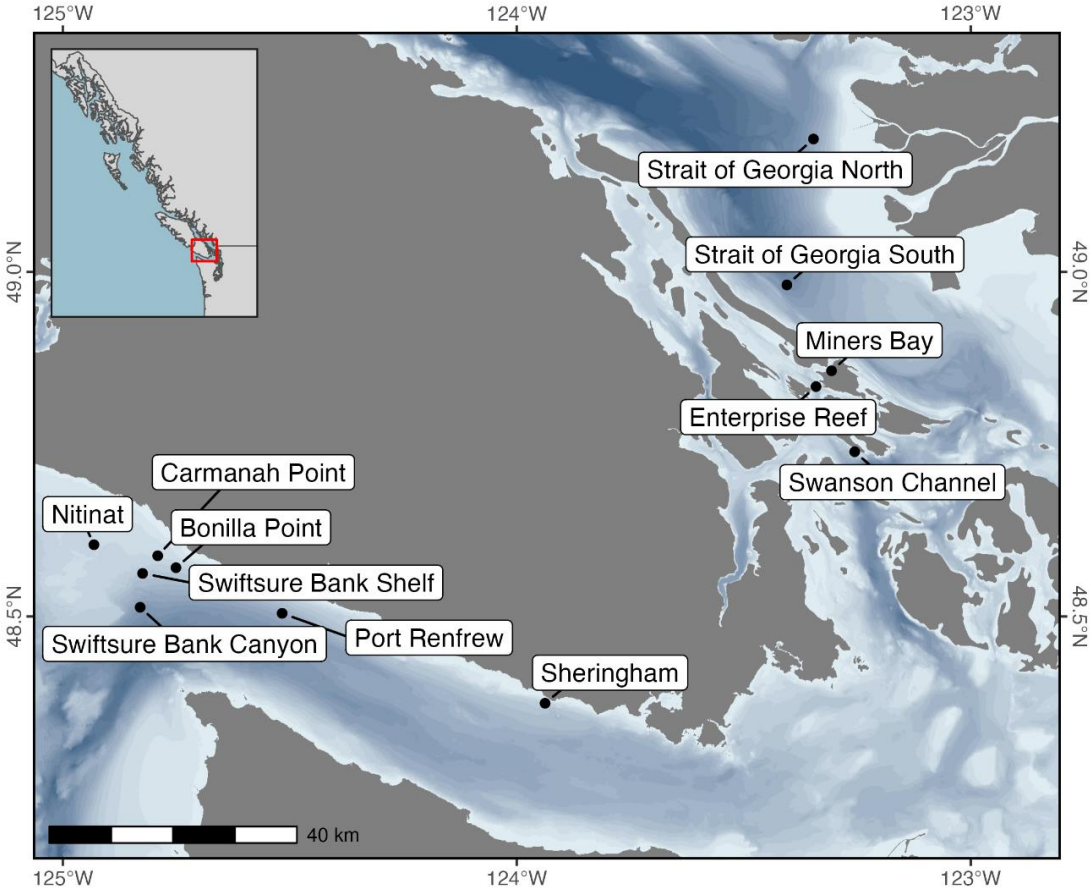

Figure 2. Map of southern Vancouver Island and Gulf Islands, displaying locations where hydrophones were deployed.

The locations and time periods in Table 1 were chosen to capture a variety of seasonal, environmental and anthropogenic conditions. Note that the audio from each time frame that was included in this study does not necessarily encompass the full time period between start and end dates. As such, the number of detections in each subset are not representative of the presence, or lack thereof, of each species at each location. For example, the recordings from Swiftsure Bank Shelf in November 2018 (Subset 1) took place during a near-constant presence of humpback whales at that site. This resulted in an overwhelming number of detections over only five days, so only a portion of the calls were annotated. Strong calls with visible harmonics were prioritized over quieter calls with little to no visible harmonic content. This was done as strong calls were much less common than weak calls in the rest of the dataset.

When processing the audio through PAMGuard for acquiring WMD detection data, all audio files were downsampled to 48 kHz and streamed through the FFT engine module with an FFT length of 2048 samples, a hop length of 1024 samples, and a Hann window function, with click removal kept off. In most subsets, the downsampled audio files were saved via the PAMGuard Sound Recorder module while it was simultaneously being run through the WMD, in order to reduce the file sizes for easier processing; however, following annotation, it was discovered that the anti-aliasing filter in the Decimator module had been erroneously removed. There was an attempt to fix this by re-downsampling all of the raw audio, but this revealed that PAMGuard version 2.01.04, which was used to process subsets 1, 3, 4, 5, 8, and A, had a bug in the Decimator module that produced random extra samples in each recorded file, which resulted in the contours only matching up time-wise with the actual calls in the files produced by the Sound Recorder and not in the raw audio. This was not replicable, and due to the time constraints of re-processing and re-annotating the data from these sets, they were kept as-is, with the compromise that the audio would be further downsampled to 32 kHz for the Feature Extractor from this point on, and any contours above 12 kHz would be ignored, as counter-measures against the aliased audio. Visual analysis of the audio before the addition of these counter-measures revealed the (albeit rare) presence of signal-produced aliasing, which was generally caused by especially-strong instances of KW or HW harmonicism that were mistaken for secondary call components.

WMD noise and thresholding processes were kept at their default values, with a median filter length of 61, an average subtraction update constant of 0.02, and thresholding at 8 dB. The minimum contour frequency was set to 200 Hz, and the minimum contour length was set to 15 slices (~341 ms).

Manual annotation of the WMD detections was performed using the Whistle and Moan Annotation Tool (WMAT), which comes included with the MIRFEE plugin. Detections were labelled by species or sound source in the “species” column of the WMAT. When combining annotation data with feature vector data in the Training Set Builder, all detections that were unambiguously produced by killer whales regardless of ecotype were all simply labelled “KW” in the training set; likewise, all humpback whale detections were labelled “HW”, and all detections caused by either man-made or non-biological sound sources were labelled “V/E” (for “vessel or environment”). The vast majority of detections could be fit into one of these three categories, with rare exceptions, most notably a Pacific white-sided dolphin encounter that produced a large number of detections occurred in subset 2, but this species rarely occurred outside of this instance. Detections from outside the three main categories were ultimately omitted from the training set in this study due to their small quantities.

Every detection in the final training set came from an audio file that contained at least five detections (the “Full Audio Set”). To save processing time during the Feature Extractor settings testing phase, a smaller audio set (the “Reduced Audio Set”) was created by randomly selecting files from the Full Audio Set, with the additional constraints that each subset could only contribute 2000 detections from each species, and each audio file could not contain more than 200 detections. This was to prevent an audio file with a large number of detections from representing too large of a portion of that set. Any detection that could not be labelled with certainty was omitted from the

training set for this study. Any contours that overlapped temporally with any other contour with a different label were removed as far as could be done. In cases where a call cluster contains contours from more than one label, only the detections where the label had a plurality were kept, using a built-in feature of the Training Set Builder. Lastly, contours were grouped into call clusters with a separation threshold of two seconds.

Feature Extractor settings testing

To determine the optimal Feature Extractor settings, the Reduced Audio Set was repeatedly run through the Feature Extractor using different combinations of settings with a feature vector consisting of audio-based features with generic parameters. The resulting training sets were cross-validated with the Test Classifier twice using the following settings:

- The *HistGradientBoostingClassifier* (Pedregosa et al., 2011) was used as the training model, with a learning rate of 0.1, 100 iterations at maximum, and maximum depth turned off. This model was found to be the most effective in prototype testing for virtually every test compared to Scikit-Learn's other ensemble models, so it was used exclusively.

- Leave-one-out cross-validation, where the training set was partitioned by the first digit of each detection's subset ID (equivalent to the ID column in Tables 1 and 2), thus, no detections were tested against training data that came from the same subset.

- For each instance of the training model for each partition, the number of data entries from each label were reduced to match that of the least-populous label, in order to reduce bias. Additionally, it was set such that each call cluster only contributed, at most, two detections each to the training data, in order to reduce bias towards especially clear calls that produced a lot of detections. Which detections were discarded were chosen randomly by the algorithm for every test run.

All audio was downsampled to 32 kHz, and all contours above 12 kHz were omitted from this study. For the first run, the clip length was set to 350 ms, the STFT length was set to 1024 bins, the STFT hop size was set to 512 samples, and noise reduction and filters were turned off. From there, the settings were iteratively adjusted for each run and the best result for a setting was kept for testing the next setting; in order, these were noise reduction, high-pass filtering without noise reduction, noise reduction and high-pass filtering combined, clip length, and STFT length and hop.

After determining which audio settings produced the best overall recall values, to determine which features to use for the final tests, the Reduced Audio Set was run through the Feature Extractor using a longer feature vector with contour metadata features added and a wider variety of combinations for features with specific parameters. ANOVA-F scores were calculated for each feature using Scikit-Learn's *SelectKBest* function (Pedregosa et al., 2011) to determine which features performed

the best at distinguishing between different species' calls. In some cases where multiple features were found to be too similar to each other (e.g. minimum versus maximum contour frequency), only the one that scored highest was kept. Some features that were implemented for the audio settings testing were not found to be useful in this run, and were thus removed from the final vector and not further mentioned in this report. Additionally, the algorithms of some features (specifically spectral contrast, formants, harmonic centroid, and BEHC) were modified after this run, so the new implementations of these features were run through again and tested similarly.

The final feature vector used consisted of the following:

- Contour duration
- Minimum contour frequency
- Contour frequency range
- Contour frequency slope
- Slice data frequencies: standard deviation
- 1st derivative of slice data frequencies: mean, standard deviation, and maximum (3 features)
- 2nd derivative of slice data frequencies: mean, standard deviation, and maximum (3 features)
- Slice data frequency start-to-end slope
- Formants, using 22 poles (maximum expected fundamental of 6000 Hz), minimum formant frequency of 90 Hz, maximum formant bandwidth of 400 Hz:
    - Frequencies of formants 1 to 4: means and standard deviations (8 features)
    - Frequency ratios of formants 2 to 4 against formant 1: means and standard deviations (6 features)
- MFCCs, 12 coefficients calculated:
    - Coefficients 1 to 12: means and standard deviations (24 features)
    - All coefficients combined: mean and standard deviation (2 features)
- Praat fundamental frequencies, pitch tracking range of 50 to 16000 Hz: median and standard deviation (2 features)
- Praat-tracked harmonic features, pitch tracking range of 50 to 16000 Hz:
    - Total harmonic distortion, calculated on 8 harmonics: median and standard deviation (2 features)
    - Harmonics-to-background ratio, calculated on 8 harmonics: mean and standard deviation (2 features)
    - Harmonic centroid: mean and standard deviation (2 features)
    - BEHC, calculated on 8 harmonics, frame means: median and standard deviation (2 features)
- Root mean square: mean and standard deviation (2 features)
- Spectral bandwidth:
    - Power of 2, normalized: standard deviation
    - Power of 4, normalized: mean
    - Power of 4, not normalized: standard deviation
- Spectral centroid: mean and standard deviation (2 features)

- Spectral contrast, 4 frequency bands, first-bin frequency cutoff of 500 Hz, linear:
    - Bands 1 to 4: means and standard deviations (8 features)
    - All bands combined: mean and standard deviation (2 features)
- Spectral flatness, power of 3: mean and standard deviation (2 features)
- Spectral flux: mean and standard deviation (2 features)
- Spectral magnitude:
    - 0 to 1000 Hz: mean and standard deviation (2 features)
    - 1000 to 2000 Hz: mean and standard deviation (2 features)
    - 2000 to 4000 Hz: mean and standard deviation (2 features)
    - 4000 to 12000 Hz: mean and standard deviation (2 features)
- Spectral rolloff, threshold of 85%: mean and standard deviation (2 features)
- Zero-crossing rate: mean, standard deviation, and maximum (3 features)

This comes to a total of 96 features.

Following settings testing, the Full Audio Set was run through the Feature Extractor using the settings combination with the best results and the aforementioned feature vector, which, as will be explained in the results, used a 2-second clip length. From this point on, this training set will be referred to as “Set A”.

As the vast majority of contours (and harmonics that caused said contours) were much shorter than the 2-second clip length, likely resulting in some clips containing a large chunk of silence or background noise, the Full Audio Set was run through again with the same settings and features as Set A, but with the clip length set to variably match that of each contour, in order to analyze which features are affected by such a discrepancy. This second training set will be referred to as “Set B”.

RESULTS AND DISCUSSION

Data collection

The detection and call cluster counts for each species in each subset are detailed in Table 2. Data was organized into 18 subsets, containing 316,555 WMD detections organized into 111,859 call clusters. The KW, HW, and V/E labels applied to 37.7%, 34.4%, and 27.9% of the detections, respectively, and 28.2%, 39.3%, and 32.4% of the call clusters, respectively. KW call clusters were thus larger and less numerous overall than HW or V/E clusters.

Feature Extractor settings testing

Using the procedure described in the methodology, the best overall recall values were produced using the following audio settings:

- Noise reduction applied with a scalar of 1.5
- A high-pass filter applied with a threshold of 500 Hz and a filter order of 4

- A clip length of 2000 ms (which was the maximum clip length attempted due to the already-extensive processing time)

- An STFT length of 4096 bins and a hop size of 1024 samples

The recall and overall accuracy values of select test runs are displayed in Table 3, in order to demonstrate how each setting change improved the results. Set 05-04, which used the settings above (albeit with a clip length of 1000 ms, to save processing time), performed substantially better than the first test, 00-01, with overall accuracy increases from 71.6% and 71.7% to 81.5% and 81.6%, and KW recall score increases from 63.3% and 64.5% to 80.5% on both runs. While the initial addition of noise reduction (set 01-03) does not appear to significantly improve the results compared to those of set 00-01 and even causes V/E recall to decrease, when using the settings from 05-04 and removing noise removal (set 05-08), overall accuracy decreased to 76.0% and 76.1% and KW recall score decreased to 68.7% and 69.3%, demonstrating that the noise reduction is useful in combination with the updated settings.

Call classification and model analysis

Using the same Test Classifier settings as in the Feature Extractor settings testing, Set A (2-second clip length) and Set B (contour-matched clip length) were cross-validated, with the confusion matrices from these runs shown in Tables 4 and 5. For the results for each subset shown in Tables 1 and 2, see Supplementary Material.

With Set A, recall scores ranged between 86.4% (HW) and 90.1% (V/E), while there was a much wider gap for precision scores, which ranged between 79.6% (KW) and 91.6% (HW), with an overall accuracy of 87.8%. As was observed in settings testing with the smaller audio set, using a 2-second clip length as with Set A produced slightly superior results compared to using the variable clip length as with Set B in every metric, with a difference of 2.4 percentage points in overall accuracy. However, using the variable clip length setting takes roughly 40% as much time to process as using a 2-second clip length, so this may be perceived as a reasonable trade-off.

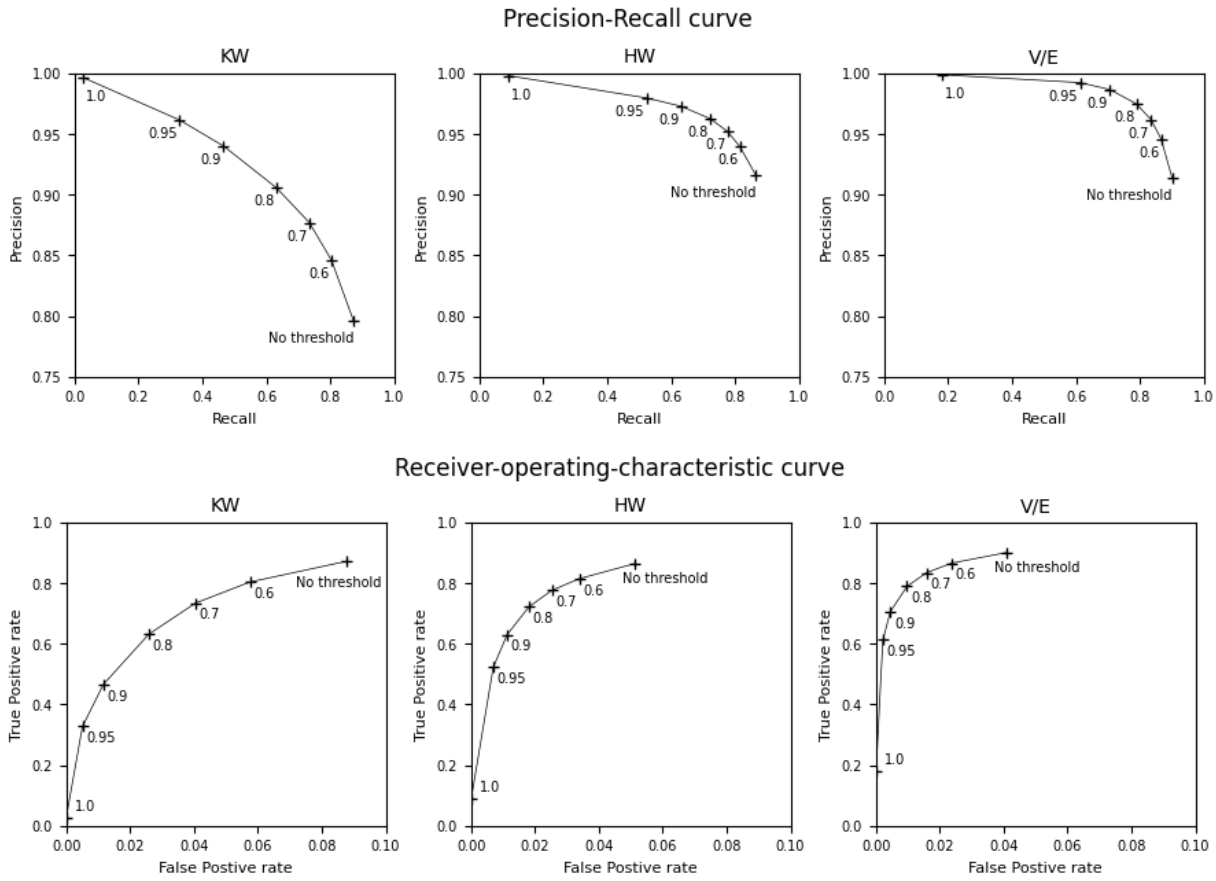

Figure 3. Precision-recall curves and receiver-operating-characteristic curves of Set A's cross-validation run for each class (killer whale, humpback whale, and vessel/environment call clusters, from left to right), with plot points labelled by prediction score threshold value. As there are more than two classes, non-target clusters incorrectly predicted as the other non-target class are counted here as true negatives. Note the axis limits.

The *HistGradientBoostingClassifier* provides class probability scores using the *predict_proba* function (Pedregosa et al., 2011), which in turn provides a certainty metric for each prediction. Using this with a range of specified score thresholds, precision-recall and receiver-operating-characteristic (ROC) curves (Hildebrand et al., 2022) for Set A's cross-validation run (Figure 3) demonstrated that higher score thresholds produce higher precision rates and lower false-positive rates, at the expense of recall and true-positive rates. The KW class notably had a less favourable precision rate at any reasonable threshold.

During the annotation process, it was noted that the majority of HW calls that were annotated were distant moans with little or no harmonicism visible in the spectrogram; this is partially observable in the minimum contour frequency histogram in Figure 4. Therefore, there were concerns about the viability of the classifier when classifying a HW call with stronger harmonicism, such as the one shown in Figure 1, due to those types of calls producing detections within the typical frequency range of KW and V/E detections. These calls were not specially marked when annotated, but they tended to produce a large number of detections per call cluster compared to distant calls with little

to no harmonicism visible. When only analyzing the results from call clusters that contained 10 or more detections each (Table 6), the recall scores for KW and V/E clusters dramatically increase to 98.0% and 98.8%, respectively, whereas the recall score for HW clusters decreases to 77.4%. This implies that the classifier is capable of recognizing HW calls with visible harmonicism, but is worse at it than it is at classifying low frequency HW calls with less harmonic activity. Noting that KW and V/E clusters with large numbers of harmonics were able to be easily identified, simply adding more harmonic/high-frequency HW calls to the training set may alleviate this problem. Inversely, this also implies that the classifier is better at identifying KW calls with a lot of visible harmonics than with KW calls that are either distant or masked by noise.

Performance analysis of select features

In order to analyze the performance of each feature, ANOVA F-scores were calculated for each feature using Scikit-Learn's *SelectKBest* function (Pedregosa et al., 2011). Due to the dependant nature of the training data, p-values and the numerical values of the F-scores are not useful because of the potential for pseudo-replication, but the relative differences in F-scores between features does provide a means of comparing their performance. There was a very large variation in F-scores, which demonstrates that some features are much more useful than others. To demonstrate this, the F-scores for the best-performing features from each algorithm are displayed in Table 7. For the F-scores of all features, see Supplementary Material. It should be noted that when fitting the models for producing these scores, the same random sampling methods used to fit the models for the cross-validation tests are also used, so different scores would be produced with each fit and do not exactly correspond to the same models used in the cross-validation tests; however, the scores produced should provide a close representation nonetheless.

When using a 2-second clip length as per Set A, spectral rolloff, spectral bandwidth, spectral centroid (standard deviation), and spectral flatness all appear to be specifically useful at differentiating V/E detections from the marine mammal vocalizations in this study. Minimum contour frequency, formant frequency, MFCCs, formant frequency ratio, zero-crossing rate, harmonics-to-background ratio, and spectral centroid (mean) all appear to be specifically useful for identifying humpback whale calls. Frequency slope and range, standard deviation of slice data frequencies, Praat fundamental frequency (standard deviation), and spectral flux appear to be specifically useful for identifying killer whale calls. BEHC appears to be useful for all three classes. When using a variable clip length matched to the length of each contour, however, some of the F-scores for the audio-based features drop significantly, especially the ones that are best at identifying V/E detections. One potential reason for this is that vessel noise "encounters" are typically long harmonic drones that produce detections with subtle changes in frequency and magnitude, resulting in most 2-second clips containing vessel noise beyond the length of the contour, whereas KW and HW calls are typically much shorter than 2 seconds, resulting in much of each clip consisting of silence or background noise instead; examples of these phenomena are shown in Figure 1. Conversely, the F-scores for the selected MFCC and formant features increased significantly with the variable clip length, and pitch tracker-based features (barring the standard deviation of the Praat fundamental frequency, for unclear reasons) appear to

have been less affected, possibly explained by those features not taking audio frames where a pitch could not be discerned (e.g. in silence or non-tonal noise) into account. The best features from each "category" in Set B with all three classes are largely the same as those in Set A, with a few exceptions, the most notable by far being the mean of the spectral centroid receiving an F-score nearly three times greater than the standard deviation. Again, this makes sense, as the mean would work better if the clip is limited to an actual vocalization, whereas the standard deviation would work better at telling apart clips that contain a large section of silence from those that do not.

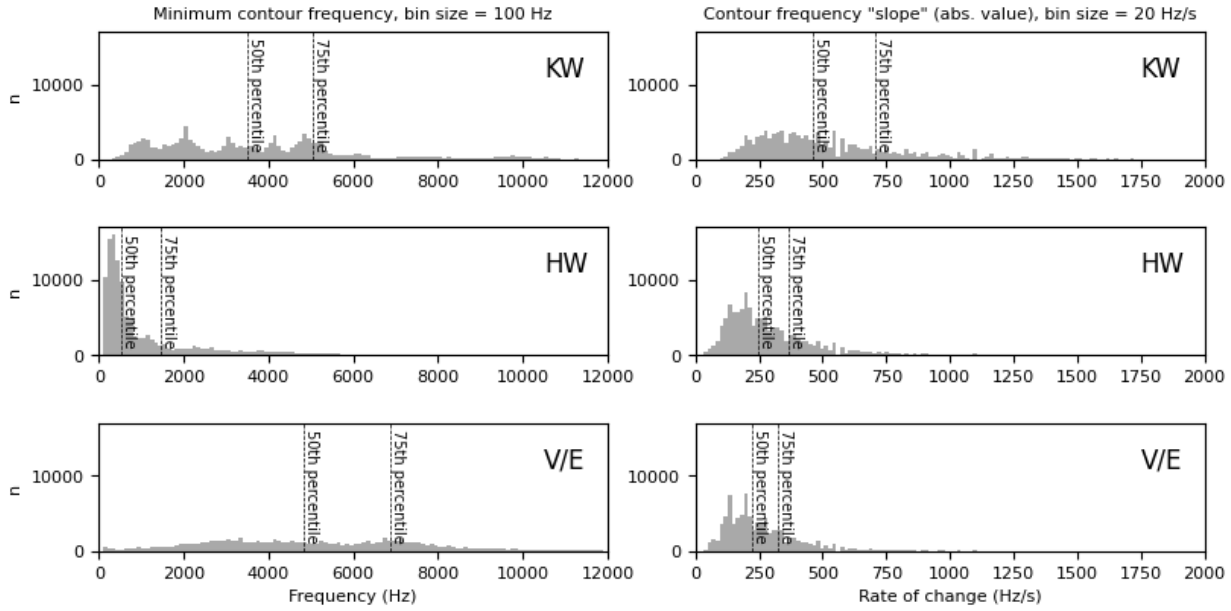

Figure 4. Histograms counting the number of detections by minimum contour frequency (left column) and header data frequency "slope" (right column) for killer whales (top row), humpback whales (centre row) and vessel and environmental noise (bottom row). Note that many if not most detections were caused by harmonics as opposed to the fundamental frequency.

KWs were almost equally as likely to be mistakenly labelled as V/E as HW (Table 4), yet HWs were almost five times more likely to be mistakenly labelled as KW than V/E. A potential contributor to this is that the majority of HW calls were distant low-frequency moans with little or no harmonic activity visible in the spectrogram, and most HW detections had a minimum contour frequency below 1000 Hz, whereas the bulk of KW detections were spread out between 500 and 6000 Hz, while V/E detections were spread out across much of the whole spectrum (Figure 4). Thus, perhaps unsurprisingly, minimum contour frequency produced the highest ANOVA-F score of all features between the three classes (Table 7). This score slightly increases when only KW and HW detections were in the set and doubles when only HW and V/E detections were in the set, but decreases to only a sixth as much when only KW and V/E detections were in the set.

Another tendency that was frequently observed during the annotation process was KW contours being more likely to change in frequency compared to V/E contours, due to the general nature of killer whale calls versus detections caused by monotone singing

propellers. While this has plenty of exceptions (e.g. engine acceleration, monotone KW calls), this pattern generally holds true, as half of the KW detections have a higher header data contour frequency “slope” than the vast majority of HW and V/E detections (Figure 4). The ANOVA-F score for this feature was on the higher end in comparison to other features, and this held true when only KW and V/E detections were in the set and when only KW and HW detections were in the set, but the score for this feature decreases to only a tenth as much with only HW and V/E detections in the set.

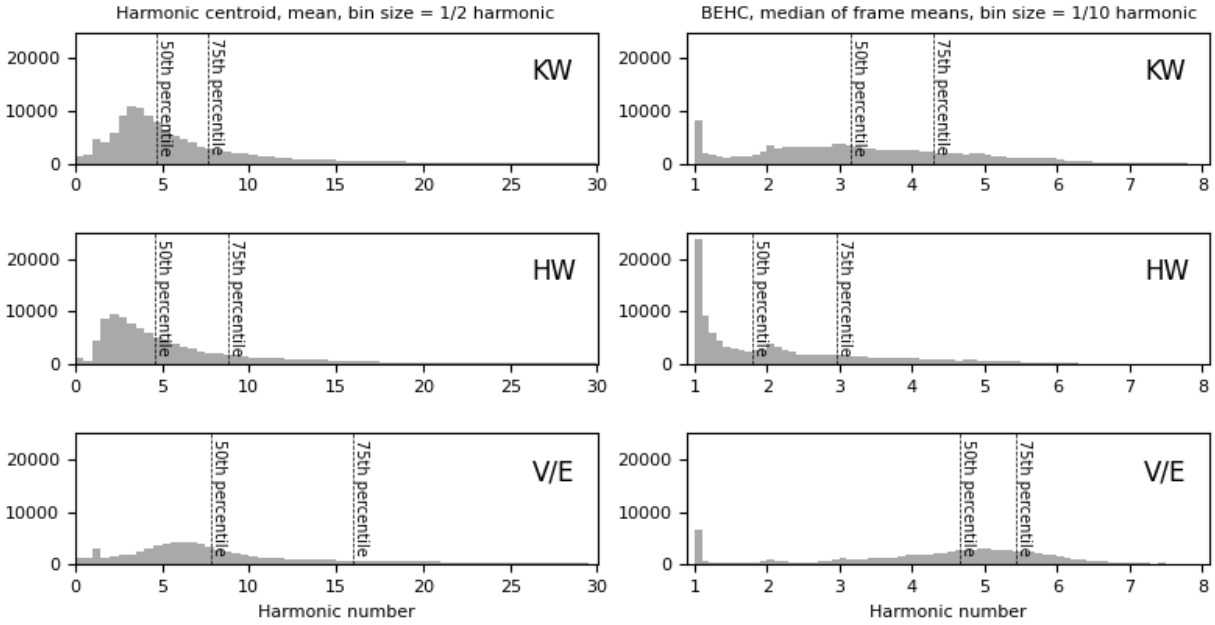

Figure 5. Histograms counting the number of detections in Set B by the mean of the harmonic centroid (left column) and the median of frame means of the bin-exclusive harmonic centroid by (right column) for killer whales (top row), humpback whales (centre row) and vessel and environmental noise (bottom row).

The median of the frame means of the “bin-exclusive harmonic centroid” feature had the highest score of all features for its worst species combination, with its lowest F-score between sets being for KW vs. V/E in Set B. Comparatively, the mean of the original harmonic centroid received a slightly higher score in this particular combination, but received substantially lower scores in every other case. As shown in Figure 5, there are visible differences in BEHC distribution between the classes, with KW detections spread out between the bottom six harmonics and centered around the third harmonic, with HW detections largely clustered around the fundamental and second harmonic, and V/E detections centered around the fifth harmonic, with little activity around the second and third harmonics. With harmonic centroid, the KW and HW distributions overlap much more closely. The key difference between these two features is that the BEHC only takes the bins corresponding to approximate harmonics of the found fundamental into account, whereas the harmonic centroid uses the centroid of the whole spectrum. While the noise reduction algorithm was found to improve the results, it is certainly not perfect at removing background noise that occurs between harmonics, and it does not attempt to remove clicks or work around instances where multiple individuals of the same species are vocalizing; the harmonic centroid would likely be more negatively affected by these factors.

CONCLUSION

This report describes the effectiveness of MIRFEE, an ensemble-type machine learning classifier, used in conjunction with the PAMGuard Whistle and Moan Detector, at differentiating between killer whale and humpback whale vocalizations, and non-biological sounds at various locations in and around the Salish Sea using features derived from detection metadata and corresponding audio data. The training set used to test the efficacy of the classifier covered a wide range of locations, seasons, and environmental conditions. The final training sets contained a total of 316555 detections that were organized into 111859 "call clusters" defined by grouping all detections within two seconds of each other together, with a relatively even distribution between the three classification categories.

When performing leave-one-out cross-validation on a training set consisting of features extracted from a large set of hydrophone audio and sorted into subsets by location and timeframe, the resulting precision and recall rates were 79.6% and 87.2%, respectively, for killer whale calls, 91.6% and 86.4% for humpback whale calls, and 91.4% and 90.1% for vessel and environmental noise, with an overall accuracy of 87.8%. The use of a noise reduction algorithm that proportionally reduces the magnitudes of STFT bins using average frequency band magnitudes from a clip preceding a cluster of detections was found to improve the results, as was the simultaneous use of a 4-order high-pass filter with a threshold of 500 Hz. To improve the classification of humpback whale calls that overlap in frequency with killer whale calls and vessel noise, adding more humpback song and/or highly-harmonic social calls to the training set is recommended.

The vast range of ANOVA F-scores implies that some features are substantially more useful than others. Features taken from the detector's metadata involving frequency and changes of frequency were found to be useful, which is highly beneficial due to their low computational complexity. In terms of audio-based features, while MFCCs, spectral centroid, spectral rolloff, and zero-crossing rate—features used with the Orchive (Ness et al., 2013)—were all found to be useful in this scenario, spectral bandwidth, BEHC, formants, spectral flatness, and harmonics-to-background ratio—all experimental additions—were found to be useful as well, and further investigation into the use of these features for similar purposes is encouraged.

SUPPLEMENTARY MATERIAL

See supplementary material at https://github.com/htleblond/PAMGuardMIRFEE/tree/main/supplementary%20material for full cross-validation results and ANOVA F-Scores for the full feature vectors of each test run.

The MIRFEE plugin can be downloaded at: https://github.com/htleblond/PAMGuardMIRFEE.

ACKNOWLEDGEMENTS

Special thanks to George Tzanetakis for brief but invaluable insight into the use of MIR features and training models in the early stages of the plugin's development, to Ruth Joy and Kaitlin Palmer for assistance with statistical analyses and for thorough reviews of the manuscript, to Jess Gibbard for assistance with the use of formants, and, lastly, to the fine folks at SMRU for developing PAMGuard in the first place.

REFERENCES

Alías, F., Socorro, J.C. and Sevillano, X., 2016. A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds. *Applied Sciences*, *6*(5), p.143.

Au, W.W., Ford, J.K., Horne, J.K. and Allman, K.A.N., 2004. Echolocation signals of free-ranging killer whales (Orcinus orca) and modeling of foraging for chinook salmon (Oncorhynchus tshawytscha). *The Journal of the Acoustical Society of America*, *115*(2), pp.901-909.

Benetos, E., Kotti, M. and Kotropoulos, C., 2006, May. Musical instrument classification using non-negative matrix factorization algorithms and subset feature selection. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings* (Vol. 5, pp. V-V). IEEE.

Bergler, C., Schröter, H., Cheng, R.X., Barth, V., Weber, M., Nöth, E., Hofer, H. and Maier, A., 2019. ORCA-SPOT: An automatic killer whale sound detection toolkit using deep learning. *Scientific reports*, *9*(1), p.10997.

Binder, C.M. and Hines, P.C., 2019. Range-dependent impacts of ocean acoustic propagation on automated classification of transmitted bowhead and humpback whale vocalizations. *The Journal of the Acoustical Society of America*, *145*(4), pp.2480-2497.

Dunlop, R.A., Cato, D.H. and Noad, M.J., 2008. Non-song acoustic communication in migrating humpback whales (Megaptera novaeangliae). *Marine Mammal Science*, *24*(3), pp.613-629.

Filatova, O.A., Guzeev, M.A., Fedutin, I.D., Burdin, A.M. and Hoyt, E., 2013. Dependence of killer whale (Orcinus orca) acoustic signals on the type of activity and social context. *Biology bulletin*, *40*, pp.790-796.

Fisheries and Oceans Canada. 2007. Recovery Strategy for the Transient Killer Whale (Orcinus orca) in Canada. Species at Risk Act Recovery Strategy Series. Fisheries and Oceans Canada, Vancouver, vi + 46 pp.

Ford, J.K., 1989. Acoustic behaviour of resident killer whales (Orcinus orca) off Vancouver Island, British Columbia. *Canadian Journal of Zoology*, *67*(3), pp.727-745.

Ford, J.K.B., Koot, B., Vagle, S., Hall-Patch, N. and Kamitakahara, G., 2010. Passive acoustic monitoring of large whales in offshore waters of British Columbia. *Canadian Technical Report of Fisheries and Aquatic Sciences*, *2898*.

Ford, J.K.B., Pilkington, J.F., Reira, A., Otsuki, M., Gisborne, B., Abernethy, R.M., Stredulinsky, E.H., Towers, J.R., and Ellis, G.M. 2017. Habitats of Special Importance to Resident Killer Whales (Orcinus orca) off the West Coast of Canada. DFO Can. Sci. Advis. Sec. Res. Doc. 2017/035. viii + 57 p.

Gibb, R., Browning, E., Glover-Kapfer, P. and Jones, K.E., 2019. Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. *Methods in Ecology and Evolution*, *10*(2), pp.169-185.

Gillespie, D., Mellinger, D.K., Gordon, J., Mclaren, D., Redmond, P., McHugh, R., Trinder, P.W., Deng, X.Y. and Thode, A., 2008. PAMGUARD: Semiautomated, open source software for real-time acoustic detection and localisation of cetaceans. *Journal of the Acoustical Society of America*, *30*(5), pp.54-62.

Gillespie, D., Caillat, M., Gordon, J. and White, P., 2013. Automatic detection and classification of odontocete whistles. *The Journal of the Acoustical Society of America*, *134*(3), pp.2427-2437.

Hermes, K., Brookes, T. and Hummersone, C., 2016. The harmonic centroid as a predictor of string instrument timbral clarity. *Audio Engineering Society proceedings*.

Hildebrand, J.A., Frasier, K.E., Helble, T.A. and Roch, M.A., 2022. Performance metrics for marine mammal signal detection and classification. *The Journal of the Acoustical Society of America*, *151*(1), pp.414-427.

IRCAM (no date) *Introduction - Linear Predictive Coding*. https://support.ircam.fr/docs/AudioSculpt/3.0/co/LPC_1.html (Accessed: July 23, 2024).

Jadoul, Y., Thompson, B. and De Boer, B., 2018. Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, *71*, pp.1-15.

Kent, R.D. and Vorperian, H.K., 2018. Static measurements of vowel formant frequencies and bandwidths: A review. *Journal of communication disorders*, *74*, pp.74-97.

Leroy, E.C., Thomisch, K., Royer, J.Y., Boebel, O. and Van Opzeeland, I., 2018. On the reliability of acoustic annotations and automatic detections of Antarctic blue whale calls under different acoustic conditions. *The Journal of the Acoustical Society of America*, *144*(2), pp.740-754.

MathWorks (no date) *Formant Estimation with LPC Coefficients*. https://www.mathworks.com/help/signal/ug/formant-estimation-with-lpc-coefficients.html (Accessed: April 19, 2024).

McFee, B., Raffel, C., Liang, D., Ellis, D.P., McVicar, M., Battenberg, E. and Nieto, O., 2015. librosa: Audio and music signal analysis in python. *SciPy*, *2015*, pp.18-24.

McSweeney, D.J., Chu, K.C., Dolphin, W.F. and Guinee, L.N., 1989. North Pacific humpback whale songs: A comparison of southeast Alaskan feeding ground songs with Hawaiian wintering ground songs. *Marine Mammal Science*, *5*(2), pp.139-148.

Miller, P.J., 2006. Diversity in sound pressure levels and estimated active space of resident killer whale vocalizations. *Journal of Comparative Physiology A*, *192*, pp.449-459.

Molder, Z.A., Halliday, W.D., Reidy, R., Kraemer, C.N. and Juanes, F., 2024. Humpback whale (Megaptera novaeangliae) social calls in southern British Columbia. *Marine Mammal Science*, *40*(4), p.e13138.

Ness, S., Symonds, H., Spong, P. and Tzanetakis, G., 2013. The Orchive: Data mining a massive bioacoustic archive. *arXiv preprint arXiv:1307.0589*.

PAMGuard (no date) *Configuration – PAMGuard Help*. https://www.pamguard.org/olhelp/classifiers/roccaHelp/docs/rocca_Configure.html (Accessed: April 9, 2025).

Payne, R.S. and McVay, S., 1971. Songs of Humpback Whales: Humpbacks emit sounds in long, predictable patterns ranging over frequencies audible to humans. *Science*, *173*(3997), pp.585-597.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., 2011. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, pp.2825-2830.

Riesch, R. and Deecke, V.B., 2011. Whistle communication in mammal-eating killer whales (Orcinus orca): further evidence for acoustic divergence between ecotypes. *Behavioral Ecology and Sociobiology*, *65*, pp.1377-1387.

Rohr, J.J., Fish, F.E. and Gilpatrick Jr, J.W., 2002. Maximum swim speeds of captive and free-ranging delphinids: Critical analysis of extraordinary performance. *Marine Mammal Science*, *18*(1), pp.1-19.

Snell, R.C. and Milinazzo, F., 1993. Formant location from LPC analysis data. *IEEE transactions on Speech and Audio Processing*, *1*(2), pp.129-134.

Socheleau, F.X., Leroy, E., Carvallo Pecci, A., Samaran, F., Bonnel, J. and Royer, J.Y., 2015. Automated detection of Antarctic blue whale calls. *The Journal of the Acoustical Society of America*, *138*(5), pp.3105-3117.

Sousa-Lima, R.S., Norris, T.F., Oswald, J.N. and Fernandes, D.P., 2013. A review and inventory of fixed autonomous recorders for passive acoustic monitoring of marine mammals. *Aquatic Mammals*, *39*(1), pp.23-53.

Thomsen, F., Franck, D. and Ford, J.K., 2002. On the communicative significance of whistles in wild killer whales (Orcinus orca). *Naturwissenschaften*, *89*, pp.404-407.

Thornton, S.J., Toews, S., Burnham, R., Konrad, C.M., Stredulinsky, E., Gavrilchuk, K., Thupaki, P. and Vagle, S., 2022. *Areas of elevated risk for vessel-related physical and acoustic impacts in Southern Resident killer whale (Orcinus orca) critical habitat*. Canadian Science Advisory Secretariat (CSAS).

Tzanetakis, G. and Cook, P., 2002. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, *10*(5), pp.293-302.

Usman, A.M., Ogundile, O.O. and Versfeld, D.J., 2020. Review of automatic detection and classification techniques for cetacean vocalization. *IEEE Access*, *8*, pp.105181-105206.

Vagle, S., Burnham, R., Thupaki, P., Konrad, C., Toews, S. and Thornton, S.J., 2021. *Vessel presence and acoustic environment within Southern Resident killer whale (Orcinus orca) critical habitat in the Salish Sea and Swiftsure Bank area*. Canadian Science Advisory Secretariat (CSAS).

Verfuss, U.K., Gillespie, D., Gordon, J., Marques, T.A., Miller, B., Plunkett, R., Theriault, J.A., Tollit, D.J., Zitterbart, D.P., Hubert, P. and Thomas, L., 2018. Comparing methods suitable for monitoring marine mammals in low visibility conditions during seismic surveys. *Marine Pollution Bulletin*, *126*, pp.1-18.

Westerhold, S., 2022. Total Harmonic Distortion (THD) analysis utilizing the FFT capabilities of modern digital storage oscilloscopes.

Zimmer, W.M., 2011. *Passive acoustic monitoring of cetaceans*. Cambridge University Press.

TABLES

Table 1. Audio subset information including hydrophone deployment location and model, recording timeframes, and amount of audio used in the Full Audio Set.

ID	Location	Hydrophone	Start date	End date	Audio used (mins.) *
1	Swiftsure Bank Shelf	AMAR	2018-11-26	2018-11-30	1474
2	Swiftsure Bank Canyon	ORCA	2021-02-27	2021-05-17	6931
3	Swiftsure Bank Shelf	ORCA	2021-06-03	2021-08-01	5931
4	Bonilla Point	SoundTrap	2021-08-11	2022-01-20	3775
5	Carmanah Point	SoundTrap	2022-03-07	2022-06-08	2296
6	Enterprise Reef	SoundTrap	2022-11-02	2023-03-13	1335

7	Swanson Channel	AMAR	2021-07-31	2021-10-06	4944
8	Port Renfrew	AMAR	2018-05-01	2018-10-11	597
9	Swanson Channel	AMAR	2023-01-19	2023-03-08	1129
A	Swiftsure Bank Shelf	AMAR	2019-03-03	2019-04-01	1721
B	Carmanah Point	SoundTrap	2022-06-15	2022-07-04	1743
C	Nitinat	SoundTrap	2022-06-15	2022-06-24	435
D	Sheringham Point	SoundTrap	2022-07-25	2022-08-23	921
E	Strait of Georgia North	AMAR	2021-09-01	2021-11-30	2324
F	Strait of Georgia South	AMAR	2021-09-01	2021-12-31	5364
G	Swiftsure Bank Canyon	AMAR	2022-07-01	2022-08-12	6086
H	Carmanah Point	SoundTrap	2022-12-01	2023-02-10	1880
I	Miners Bay	SoundTrap	2021-09-17	2022-01-30	1794

* Combined length of all files that contained at least five detections. Note that file lengths were between 3 to 10 minutes in length depending on the subset, and actual calls consisted of a small fraction of these numbers.

Table 2. Audio subset detection and call cluster counts for the Full Audio Set. A "detection" refers to an individual spectrogram contour marked by the Whistle and Moan Detector, whereas a "call cluster" refers to a grouping of detections where each occurs within two seconds of another.

ID	Detection count				Call cluster count			
	KW	HW	V/E	Total	KW	HW	V/E	Total
1	110	22143	34	22287	68	4533	27	4628
2	4326	2273	5033	11632	1600	1324	2867	5791
3	19532	2014	5029	26575	5217	1169	2453	8839
4	7373	33556	3562	44491	1965	11054	1542	14561
5	3696	2765	5296	11757	1464	1879	2855	6198
6	6884	0	2702	9586	1542	0	1117	2659
7	19514	16	22407	41937	2572	10	7075	9657

8	137	661	849	1647	83	416	432	931
9	4882	11	1492	6385	1347	10	604	1961
A	9296	1237	335	10868	3185	1017	277	4479
B	2925	505	3870	7300	1263	264	2521	4048
C	364	8	1105	1477	211	4	753	968
D	1616	0	4003	5619	544	0	1631	2175
E	6126	132	7888	14146	1642	122	3283	5047
F	2641	16822	3463	22926	977	10418	1801	13196
G	28424	6080	11237	45741	7437	2758	4989	15184
H	902	20383	150	21435	364	8834	135	9333
I	672	320	9754	10746	86	193	1925	2204
Total	119420	108926	88209	316555	31567	44005	36287	111859

Table 3. Recall scores for select audio settings-testing runs, demonstrating the improvement between models when new settings were applied. Note that two runs were performed for each set. The “set IDs” were designations used to arrange runs into which category of settings was being tested.

Set ID	Description	Recall (in %)			Overall accuracy (in %)
		KW	HW	V/E	
00-01	Clip length: 350 ms, STFT length: 1024, STFT hop: 512, noise reduction (NR) and filters off	63.3 64.5	71.9 71.6	78.2 77.8	71.6 71.7
01-03	00-01 with NR: scalar 1.5	67.4 67.1	76.3 76.4	74.1 73.9	72.6 72.5
02-06	00-01 with high-pass filter (HPF): 500 Hz, order 4	66.0 65.8	73.5 73.6	81.2 81.2	74.1 74.0
03-01	01-03 NR combined with 02-06 filter	72.5 72.2	79.8 79.1	76.5 75.8	76.2 75.6

04-04	03-01, clip length: 1000 ms	75.6 76.1	78.3 78.5	79.4 79.1	77.8 77.9
04-06	03-01, clip length: 2000 ms	77.2 77.5	78.9 78.5	82.3 82.6	79.7 79.8
05-04	03-01, clip length: 1000 ms, STFT length: 4096, STFT hop: 1024	**80.5** **80.5**	**80.6** **80.6**	83.1 83.3	81.5 **81.6**
05-08	05-04 without NR	68.7 69.3	73.5 73.5	**83.9** 83.7	76.0 76.1

Table 4. The confusion matrix for leave-one-out cross-validation on Set A, which used a clip length of two seconds.

	KW	HW	V/E	Recall
KW	27535	1979	2052	87.2%
HW	4940	38029	1036	86.4%
V/E	2109	1493	32685	90.1%
Precision	79.6%	91.6%	91.4%	87.8%

Table 5. The confusion matrix for leave-one-out cross-validation on Set B, which used a variable clip length that matches that of each contour.

	KW	HW	V/E	Recall
KW	26955	1761	2851	85.4%
HW	5243	36894	1868	83.8%
V/E	2952	1684	31651	87.2%
Precision	76.7%	91.5%	87.0%	85.4%

Table 6. The confusion matrix for leave-one-out cross-validation on Set A, but only including clusters containing 10 or more detections.

	KW	HW	V/E	Recall
KW	2169	19	26	98.0%
HW	270	1088	47	77.4%
V/E	9	5	1177	98.8%
Precision	88.6%	97.8%	94.2%	92.2%

Table 7. ANOVA F-scores of select features with the best-performing parameters from their respective category between all three classes in both sets. Due to the dependent nature of the training data, the numerical values of the F-scores are not useful on their own due to potential for pseudo-replication and are only intended for comparing performance between features.

Feature name	Set A				Set B			
	All three classes	KW vs. HW	KW vs. V/E	HW vs. V/E	All three classes	KW vs. HW	KW vs. V/E	HW vs. V/E
Minimum contour frequency *	46912.4	50821.7	7765.4	100981.6	47159.3	51063.6	7838.0	101146.9
Spectral rolloff, 85% threshold, standard deviation	29758.6	273.1	50371.7	52635.3	11270.8	39.9	17043.9	22963.9
BEHC, 8 harmonics, median of frame means	28736.1	10352.3	18860.0	54270.9	22953.5	15014.5	9073.1	45152.5
Spectral bandwidth, power of 2, normalized, standard deviation	25758.4	418.9	52106.6	37619.8	7952.4	136.2	13611.4	12505.7
Spectral centroid, standard deviation **	21339.0	785.8	38400.0	37321.9	6904.0	282.3	11430.7	13870.3
Frequency of 1st formant, mean	19179.0	26945.5	134.8	37241.3	27007.1	41700.7	18.1	48538.1
MFCCs, 2nd coefficient of 12, mean	18834.2	21426.3	2701.1	35980.8	27428.8	32601.4	2127.9	50932.3
Ratio between 4th and 1st formants, mean	17687.7	16243.0	1464.5	30590.6	28371.6	33822.4	2.8	36083.3
Frequency slope (absolute value) *	17634.5	15864.4	25516.6	1895.9	17741.0	16102.0	25849.3	1921.4
Zero-crossing rate, mean **	16837.2	14481.4	5510.0	35179.0	15061.1	23238.4	650.3	31380.3
Frequency range *	16440.7	11144.0	26523.2	6163.3	16426.0	11357.2	26698.6	6326.1
Zero-crossing rate, maximum ***	16097.0	25743.5	0.4	33542.7	16436.9	27027.6	104.1	32691.9
Slice data frequencies, standard deviation *	14792.7	10686.2	23664.9	4386.0	14807.0	10838.1	23976.1	4507.9
Harmonics-to-background ratio, mean	13399.6	9979.1	3709.8	25078.2	14026.1	6157.1	7851.0	28206.3

Spectral flatness, power of 3, standard deviation **	11060.7	186.5	16895.4	22318.8	1633.7	716.8	1038.3	3287.6
Praat fundamental frequency, standard deviation **	8287.8	7979.3	13121.5	382.4	638.5	506.2	1408.5	98.1
Spectral centroid, mean ***	7976.3	11913.0	564.4	10911.8	18709.0	20353.6	745.7	34359.3
1st derivative of slice data frequencies, maximum *	6239.0	3116.8	11493.5	3347.6	6200.8	3175.9	11782.0	3292.0
Spectral flatness, power of 3, mean ***	6080.8	189.3	9038.6	11805.1	2455.9	2216.9	342.4	3823.6
Harmonic centroid, mean	5747.2	1386.8	10369.4	4282.0	5039.7	1665.6	9296.7	3263.0
Spectral magnitude, 0 to 1000 Hz, standard deviation	5562.9	6785.8	0.5	6088.5	6357.2	6981.6	14.9	7239.5
Slice data frequency start-to-end slope *	4937.4	3539.0	7285.8	1669.4	4836.4	3546.8	7379.1	1634.3
Spectral flux, standard deviation	4572.6	6073.2	5368.7	135.5	2328.5	3850.9	316.0	2850.6
Spectral contrast, 1st band of 4, linear, standard deviation	4340.6	4258.2	228.2	5267.1	3898.6	3767.1	221.5	4999.9
Contour duration *	4083.4	3515.6	783.3	6307.4	4064.2	3526.6	768.6	6490.6
Praat fundamental frequency, median ***	2937.0	3942.6	3238.5	203.6	2665.5	4826.4	1729.3	1101.9
2nd derivative of slice data frequencies, maximum *	2289.0	190.4	4382.9	2917.7	2318.5	184.8	4539.9	2890.9
Root mean square, standard deviation	1867.8	736.5	1317.9	3670.1	1808.4	1266.5	594.4	2976.4
Total harmonic distortion, standard deviation	307.7	5110.9	19.0	339.3	5247.2	4654.0	778.3	7280.1

* Feature is not affected by audio clip length. The difference in scores between Set A and Set B for these features is due to the use of random sampling each time the training model is fitted.

** Feature was the best in its category in Set A, but not Set B.

*** Feature was the best in its category in Set B, but not Set A.