Bird Species Identification from Audio Data

Ching Seh Wu
Department of Computer Science
San Jose State University
San Jose, CA USA
ching-seh.wu@sjsu.edu

Sasanka Kosuru
Department of Computer Science
San Jose State University
San Jose, CA USA
sasanka.kosuru@sjsu.edu

Samaikya Tippareddy
Department of Computer Science
San Jose State University
San Jose, CA USA
samaikya.tippareddy@sjsu.edu

Bhavya Reddy Kotla
Department of Computer Science
San Jose State University
San Jose, CA USA
bhavyareddy.kotla@sjsu.edu

***Abstract*—In order to identify the population of birds, which provides us insights into the ongoing environmental changes, our research focuses primarily on the identification of bird species using audio. The input data set has long recordings of the bird audio. In pre-processing, for reducing noise in audio, we used noise and outlier elimination static filters. We also extracted 26 acoustic features using Feature Extraction to train machine learning models. After the pre-processing, machine-learning models like decision tree, random forest, Naive Bayes classifier, Support Vector Classifier (SVC), k-nearest neighbor (K-NN), and Stochastic Gradient Descent (SGD) were used. The results showed that the Stochastic Gradient Descent (SGD) performed the best with an F1 score of 0.89 and 90% accuracy for 3 classes, when trained with combined features from original audio and cleaned audio files along with feature selection.**

***Index Terms*—Bird classification, Machine learning, Supervised classification, Noise reduction, Melspectrogram**

I. Introduction

The ever-growing human intervention and the constant climate changes have led to drastic effects on the environment as well as on the natural habitat around us. Understanding these changes and keeping track of them is key to understanding and anticipating the detrimental effects that can occur. In this regard, birds are considered one of the most useful indicators: their sensitivity to habitat change and the ease of measuring their census makes them one of the ecologist's favorite tools. Any change in bird population is considered one of the first and primary indicators of environmental upheaval.

The advantage of using bird audio for classification instead of using images is that it is easier to capture quality audio with less expensive equipment than quality images or video. But the problem is classification becomes very hard as, identifying birds manually not only requires a lot of resources but is also extensively time-consuming owing to the physical efforts needed. This is where the advancements in machine learning algorithms can be used to our advantage. These algorithms not only play a key role in analyzing the data set but also help in identifying, detecting, and tracking bird populations.

In this paper, we used a large audio birds data set which was pre-processed using Librosa, Scipy, and TorchAudio. The data pre-processing included data cleaning to generate mono-channel audio files, noise reduction using a noise reduce python library and a high-pass filter to generate clean audio, and feature extraction for obtaining features like Mel frequency spectral coefficients (MFCCs), amplitude envelope, energy, spectral centroid, spectral flux, and zero-crossing rate for training models. Then, data splitting was performed with an 80:20 ratio for train and test data. Finally, audio classification was performed using machine learning algorithms, namely decision tree, random forest, Naive Bayes classifier, Support Vector Classifier (SVC), k-nearest neighbor (K-NN), and Stochastic Gradient Descent (SGD). Finally, Feature selection was performed using recursive feature elimination (RFE) for improving classification accuracy.

II. Related Works

[1] employs a noise separation and classification filter to obtain the required data from the sound. Mel-frequency cepstral coefficient (MFCC), the most common feature in speech recognition systems, is obtained using algorithms. Different algorithms namely, Naïve Bayes, J4.8, and Multilayer perceptron were then used to classify the species. In these approaches, a model is trained to predict the component sources from synthetic mixtures or aggregations created by adding up ground-truth sources. As the performance of the model depends on the degree of match between the training data and real-world audio, relying on this type of synthetic or aggregated training data is problematic, especially since the accurate simulation of the acoustic conditions and classifying source distribution is very challenging.

To combat the above shortcomings, this [2] focuses on developing a completely unsupervised method called mixture invariant training (MixIT). In MixIT, existing mixtures are combined to construct training examples, and the model separates these examples into a variable number of latent sources, such that original mixtures can be approximated by remixing these separated sources. However, this approach was not particular to bird species classification. [3] on the other hand, focuses on using the MixIT model for birdsong data. Precision improvements, along with a downstream multi-species bird classifier, were depicted across three independent datasets. Taking the maximum model activations across the separated channels and original audio yielded the best classifier performance for these datasets.

[4] is a working note of Piczak et. al. from BirdCLEF 2016. The focus of this paper was on evaluating single-label classifiers suitable for recognizing the main (foreground) species present in the recording. The audio files were converted to mono-channel format during preprocessing, from which mel-scaled power spectrograms were generated using the Librosa library. An ensemble model with three different network architectures was proposed in this work. The Keras Deep Learning library was used to build the three networks. Each of these networks converts the input into spectrogram segments and predicts which species will be dominant. Averaging the decisions made across all segments of the input file yields the final prediction. This submission had a mean average precision of 41.2 percent for background species and 52.9 percent for foreground species. It did not, however, handle noise reduction during preprocessing, which was addressed in [5].

[5] proposed a solution that uses a visual representation of the audio as an input to the CNN. The audio files are first converted to WAV format, which is then split into chunks and normalized. Only relevant information was extracted from these chunks by discarding any chuck which was not loud enough (below the threshold). Finally, spectrograms are created by converting STFT output to image using a color map. In this paper, CNN is trained in two stages. The first was done with a colored spectrogram and the second with a black-and-white spectrogram. The method described in this paper employs a pre-trained MobileNet network designed for mobile devices. As a result, it has a small architecture and is quick to evaluate. However, when CNN was trained with ten classes, the accuracy dropped by more than 40 percent. A more reliable pre-trained convolutional neural network, such as ResNet, may aid in achieving higher accuracies.

The authors in [6], used the sequence of syllables in bird sounds and compressed the variable length sequence to a fixed-dimensional feature vector. Syllable pairs were used instead of single syllables to understand the temporal structure of the bird sounds and represented using Gaussian syllable prototypes. They used nearest neighbor classifiers on 3 different histogram representations i.e. 3 Gaussian syllable prototypes - 10, 30, and 50 Gaussians with accuracies of 76, 79, and 80 percent respectively. This paper handles the sparseness of histograms and facilitates the comparison of the histograms. However, the problem with using the syllable approach is, it is challenging to get robust segmentation for a low Signal-to-noise ratio which is addressed in [7].

In [7], a probabilistic approach is proposed using a statistical manifold. First, meaningful features are extracted from the audio by considering the audio signals as time series of samples and converted into a spectrogram and the Fourier transform is applied to distinguish between different sounds in a frame based on frequency. The unique approach in this paper from [6] is instead of bird song syllables they used histograms of frequencies, as syllables are heavily dependent on accurate segmentation and not suitable for audio with a low Signal-to-noise ratio. Also, rather than averaging all the frame-level features into a single length vector, they aggregated frame-level features by representing feature distribution in the histogram as they observed that multi-modality is common in bird sounds and by averaging, we lose significant information. As a result, a multidimensional feature vector of d-dimensions is generated, and for classification, a feature vector of frequencies for each histogram bin is given as input. The authors also used a ‘codebook’ approach to generate these histogram features for high-dimension vectors like MFCCs. In this paper, nearest-neighbor classifiers were used with statistical divergence measures, namely L1, Kellinger, and KL. The proposed model provides a simpler approach than using bird syllables and the model provides competitive or better accuracy results(85-90 percent) with the benchmark SVMs.

III. Dataset Description

The dataset is downloaded from [8], collected from Xeno-canto, an open-source website dedicated to sharing wildlife sounds worldwide. The dataset consists of 153 bird species ranging from A to M. For this research, audio files with a rating greater than three (this rating indicates audio quality), a number of samples greater than a hundred, and a duration of fewer than twenty seconds are considered. After performing the above filtering, a total of 30 species were selected.

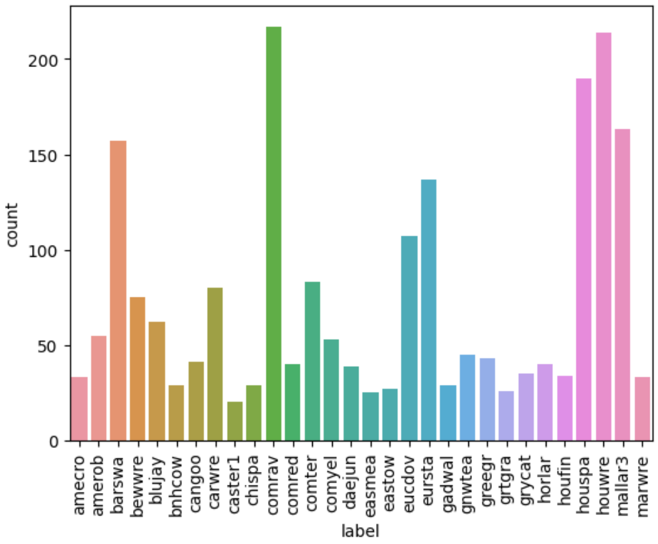

Fig. 1. Plot showing the imbalance in the dataset

IV. Methodology

We propose a bird classification system that focuses on acoustic features like spectral centroid, bandwidth, zero cross rating, MFCC, etc. The system is designed using the following steps:

A. Data Preparation

The audio files in the dataset consisted of both mono and stereo channel. The number of channels used to capture and playback audio signals is the primary distinction between stereo and mono. While stereo signals are recorded and played back on two audio channels (the left and right channels), mono signals are recorded and played back on a single audio channel [9]. Thus, in order to make audio analysis easier, we use mono, so that we would only have one channel to process. Also, all the files were in mp3 format. Librosa library was used to convert mp3 into WAV files. This is because, unlike MP3 file format which is lossy, WAV files are lossless, meaning that WAV audio is a high-quality uncompressed file. These are best suited to our scenario, given that audio files have a lot of noise like traffic, human sounds, or other low-frequency sounds. Librosa library also converts stereo into mono-channel audio files.

B. Noise reduction

We used a python library called Noise Reduce which uses spectral gating to reduce the background noise. In spectral gating, a spectrogram of an audio signal is computed, noise threshold is estimated to compute a mask, which is later used for noise gating. After this, we applied another high-pass filter to the data for further cleaning. The assumption here is, given that most of the bird noises are high-frequency, passing them through this filter, would help us extract them from the low-frequency background noise like traffic and human sounds.

Figure 2 shows different audio signals plotted in blue, orange, green, and red. The original signal contains noise and is represented in blue. After applying a high-pass filter to the original audio, we obtained the audio signal in orange. As the noise was not reduced using only the high-pass filter, we used the Noise Reduce Python library. This reduced the noise significantly, and the audio signal is represented in green. On top of the Noise Reduce library, we used a high-pass filter to further eliminate noise, and the final audio signal is represented in red.

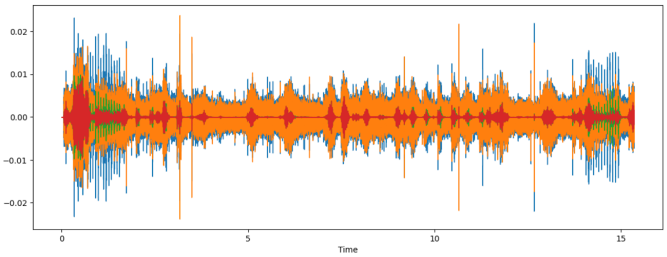

Fig. 2. Comparison of various noise reduction techniques

C. Feature Extraction

For numeric data, a total of twenty-six acoustic features were extracted. These features are as follows: Short Term Fourier Transform (STFT), Mel Frequency Cepstral Coefficients (MFCC), Root Mean Square (RMS), Spectral centroid, Spectral bandwidth, Spectral roll-off, and Zero crossing rate. Each of them can be explained as follows,

1. Short-Term Fourier Transform: This is a Fourier-related transform that uses sinusoidal frequency and phase distribution, to determine the frequency and phase content of local sections of a signal as they change over time.
2. Mel Frequency Cepstral Coefficients (MFCC): These are the majority of the features having a count of twenty. It can be defined as a small group of features that summarize in a concise way the overall shape of a spectral envelope in a given frequency band. Figure 3 shows the melspectogram plot for one of the classes, American Crow. Similarly, multiple plots for various classes can be constructed, which then can be further used as inputs for classification as well.
3. Root Mean Square Error: Audio signals are analyzed by square rooting the signal value (amplitude), averaging it over a period of time, and then taking the square root of the average value.
4. Spectral centroid: This is the measure of the amplitude at the center of the spectra of the signal distribution over a window, derived from the information contained in the Fourier transform of frequency and amplitude.
5. Spectral Bandwidth: The bandwidth of light at one-half the peak maximum.
6. Spectral roll-off: An indicator of the frequency below which a specified percentage of the total amount of spectral energy (e.g. 85 percent) resides.
7. Zero crossing rate: The speed at which a signal shifts from being positive to being negative to being zero, or from being negative to being positive. As a crucial feature to distinguish percussive sounds, its value has been widely applied in both speech recognition and music information retrieval.

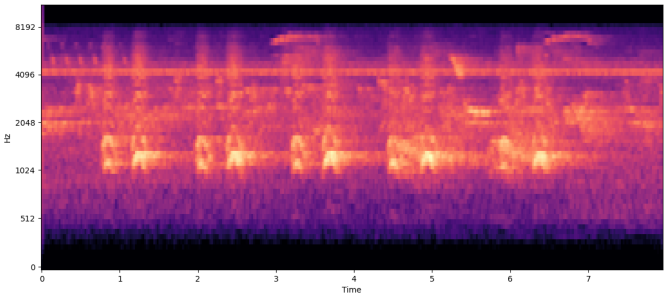

Fig. 3. Melspectrogram of American Crow

D. Dataset shuffle and split

The dataset has been split into the train, and test sets in a ratio of 80:20 with each of these sets having an equal ratio of target classes.

E. Classification models

In order to perform classification on the extracted dataset, the following supervised models were used,

1. K-nearest neighbor (KNN): An algorithm for categorizing data that calculates the likelihood that a data point will belong to one group or another based on the group to which the data points nearest to it belong.
2. Stochastic Gradient Descent (SGD): A probabilistic version of gradient descent where instead of computing the gradient for the entire dataset at each step, the method only computes the gradient for one observation selected at random.
3. Support Vector Classifier (SVC): a supervised linear algorithm that predicts or classifies data using margins.
4. Decision Trees: A type of supervised machine learning algorithm that forecasts or categorizes data using the responses to a previous set of questions. A branching method is used in this graph to show every possible result for a given input.
5. Random Forest: A categorization approach utilizing several decision trees. It uses bagging and feature randomization to create each individual tree in an effort to create an uncorrelated forest of trees whose forecast by committee is more accurate than that of any one tree..
6. Naive Bayes Classifier: A fundamental probabilistic classifier based on the Bayes theorem and strict feature independence hypotheses.

F. Feature Selection

During the analysis of the feature extraction, we were interested in finding which set of features would result in the highest level of classification accuracy. This is because feature selection helps in achieving higher classification by selecting the most suitable feature set, and also helps in reducing classification time.

We used recursive feature elimination (RFE), a feature selection algorithm that reduces model complexity by eliminating lesser significant features. It removes the weakest feature(s) until the specified number of features is reached. To obtain the optimum number of features needed, we performed cross-validation along with RFE for scoring different feature subsets to select the subset with the best scoring. We reduced features from 52 features after performing RFE, to 41 for 3 classes, 37 for 5 classes.

V. Evaluation Metrics

A. Accuracy

The accuracy of a model is calculated using the given formula below.

$$Accuracy = \frac{TN+TP}{TN+FP+FN+TP}$$

Accuracy can be misleading if used with imbalanced datasets, and therefore there are other metrics based on the confusion matrix which can be useful for evaluating performance

B. F1 Score

The F1 score, F score, or F measure is the harmonic mean of precision and sensitivity it gives importance to both factors:

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

where

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

C. Macro F1

The Macro F1 score is the unweighted mean of the F1 scores calculated per class. It is the simplest aggregation for the F1 score.

$$Macro\ F1 = \frac{Sum(F1\ scores)}{no\ of\ scores}$$

D. 5-Fold Cross Validation

Dataset is split into 5 sections and each iteration of these sections is considered a test set while the other sections are used for training. The resulting scores are calculated by calculating the average of all 5 sections' F1 or accuracy values. The formula is given below:

$$S = \frac{1}{5}(\Sigma_{i=1}^{5} S_i)$$

Where S_i is the F1 or accuracy of i-th fold.

VI. Results, Analysis and Comparison

There were three experiments conducted in this research. The first experiment was to use basic machine-learning models to classify the birds. In this, we have considered three different sets of bird classes: 3, 5, and 30 classes. The 30 classes were selected based on certain constraints mentioned in Section III, and from these 30 classes, we have randomly selected 3 and 5 classes. The results of this experiment are given in Table I. We have observed that the accuracy and F1 scores have decreased with the increase in the number of bird classes. This is because different audio species data had different audio quality samples and a limited number of audio files, resulting in an imbalance of data. By nature, in machine learning algorithms, the accuracy of predictions decreases as the degree of imbalance increases. The best results in this experiment were obtained for Stochastic Gradient descent (SGD) when trained with 3 bird species. Figure 5 shows a plot comparing the F1 scores of models trained with 3, 5, and 30 classes where Stochastic Gradient Descent (SGD) performed the best for 3 classes and 30 classes.

In the second experiment, we compared our models which were generated using the features of both original audio and noise-reduced audio, with the other models that did not use the combinations. We have observed that the models trained with combined data had better results than the models trained using

TABLE I
F1, ACCURACY AND AVERAGE FIT TIME FOR 3, 5, AND 30 CLASSES WITHOUT FEATURE SELECTION

	Metric	SGD	SVM	KNN	Decision Tree	Random Forest	Naive Bayes
3 Class	Macro F1	**0.87**	0.86	0.65	0.64	0.78	0.72
	Accuracy	**0.88**	0.86	0.71	0.66	0.80	0.74
	Avg Fit time	0.008	0.008	0.001	0.01	0.17	0.002
5 Class	Macro F1	0.70	**0.73**	0.58	0.49	0.65	0.59
	Accuracy	0.70	**0.74**	0.61	0.51	0.68	0.60
	Avg Fit time	0.02	0.02	0.001	0.02	0.28	0.005
30 Class	Macro F1	**0.21**	0.18	0.13	0.11	0.19	0.17
	Accuracy	**0.30**	0.36	0.28	0.17	0.34	0.22
	Avg Fit time	0.45	0.30	0.003	0.11	1.10	0.009

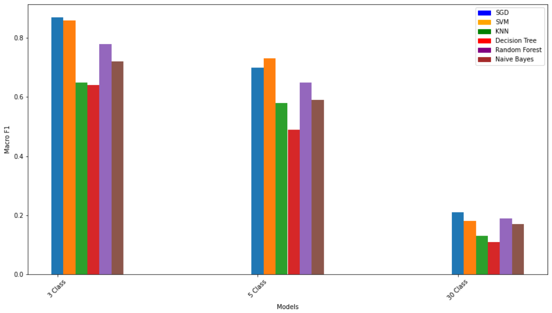

Fig. 4. Comparison of F1 scores for 3, 5, and 30 classes

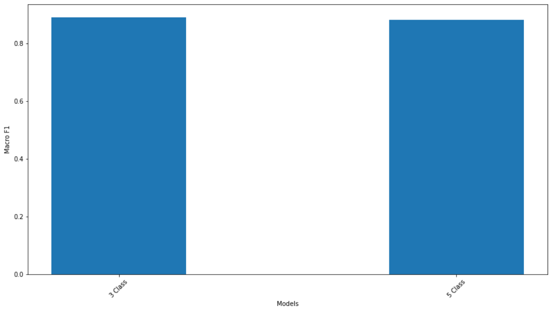

Fig. 5. Comparison of F1 scores for 3, and 5 classes of SGD after feature selection

just the original features or just the noise-reduced features. Table II gives detailed metrics for this experiment. The F1 scores when we considered combined features were 0.87 while the models trained with noise audio had 0.79 and noise-reduced models had 0.72 F1 scores.

The third experiment was conducted by training a stochastic gradient descent model with feature selection on 3 and 5 bird classes. We have observed a significant increase in accuracy and F1 scores after feature selection using the recursive feature elimination technique. Using RFE, a total of 41 features for 3 classes and 37 features for 5 classes were selected. Table III shows the accuracies and F1 scores obtained for stochastic gradient descent using 3 and 5 bird species for training.

All the models were evaluated using a cross-validation method with 5 folds.

TABLE II
COMPARISON OF STOCHASTIC GRADIENT DECENT MODELED WITH FEATURES WITH, WITHOUT NOISE, AND BOTH

Metrics	Features		
	with & without noise	with noise	without noise
Macro F1	0.87	0.79	0.72
Accuracy	0.88	0.79	0.76
Avg Fit time	0.008	0.006	0.006

VII. Conclusion

Out of all the proposed models, Stochastic Gradient Descent (SGD) had the best accuracy of 90% and F1 score of 0.89 for 3 classes with recursive feature elimination (RFE). For 5 classes, the highest accuracy was 89%, and an F1 score of 0.88 for SGD with RFE. For 30 classes, the best accuracy of 30% and F1 was 0.21 for SGD. Accuracies decreased as the number of classes increased, as different audio species data had varying audio qualities and limited audio files, creating data imbalance.

With the use of the noise reduction library along with a high-pass filter in the preprocessing stage, choosing unique features like spectral centroid, spectral roll-off, zero crossing rate, etc., in the feature extraction stage, and performing recursive feature elimination (RFE) in the feature selection

TABLE III
F1, ACCURACY AND AVERAGE FIT TIME FOR 3 AND 5 CLASSES USING RFE ON STOCHASTIC GRADIENT DECENT

	3 Class	**5 Class**
Features selected	41	37
Macro F1	0.89	0.88
Accuracy	0.90	0.89
Avg Fit time	0.004	0.004

stage, we achieve better accuracy in classification than reported in other literature on similar datasets.

VIII. Future Work

For our experiments, only numeric data was considered. In future work, the Mel spectrogram features can be plotted as images and used for image classification using Convolution Neural Networks (CNNs) or other deep learning models. Furthermore, for classification using numeric data, among the four important features of bird audio i.e., notes, syllables, phrases, and songs, individual or a combination of these features can be experimented with for classification, to see if we obtain better results.

For handling the issue of data imbalance for different species, we need to apply data resampling techniques like oversampling or undersampling, for improving accuracy when using a higher number of classes.

Given that we have audio data, this data can be preprocessed into time series data i.e., each audio recording of a particular length can be arranged into frames, which then can be read in a sequential manner. This time series conversion would then help us to apply machine learning models like the Long Short Term memory model. As an obvious extension, the number of classes can be increased by implementing deep learning models, whose accuracy would not majorly decrease with the increase in classes.

References

[1] A. E. Mehyadin, A. M. Abdulazeez, D. A. Hasan, and J. N. Saeed, “Birds sound classification based on machine learning algorithms,” *Asian Journal of Research in Computer Science*, p. 1–11, 2021.

[2] S. Wisdom, E. Tzinis, H. Erdogan, R. J. Weiss, K. Wilson, and J. R. Hershey, “Unsupervised sound separation using mixture invariant training,” Oct 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2006.12701

[3] T. Denton, S. Wisdom, and J. R. Hershey, “Improving bird classification with unsupervised sound separation,” Oct 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2110.03209

[4] K. J. Piczak, “Recognizing bird species in audio recordings using deep convolutional neural networks - ceur-ws.org,” 2016. [Online]. Available: http://ceur-ws.org/Vol-1609/16090534.pdf

[5] A. Incze, H.-B. Jancso, Z. Szilagyi, A. Farkas, and C. Sulyok, “Bird sound recognition using a convolutional neural network,” *2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY)*, 2018.

[6] P. Somervuo and A. Harma, “Bird song recognition based on syllable pair histograms,” *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*.

[7] F. Briggs, R. Raich, and X. Z. Fern, “Audio classification of bird species: A statistical manifold approach,” *2009 Ninth IEEE International Conference on Data Mining*, 2009.

[8] Vopani, “Xeno-canto bird recordings extended (a-m),” Sep 2020. [Online]. Available: https://www.kaggle.com/datasets/rohanrao/xeno-canto-bird-recordings-extended-a-m

[9] Arthur, “Is stereo or mono audio better? (applications for both),” Jan 2022. [Online]. Available: https://mynewmicrophone.com/is-stereo-or-mono-audio-better-applications-for-both/