Engine - Research into feature selection including spectrogram influence

**Rohit, Balram, Abhit**

*Abstract*—

Feature selection is an important aspect of machine learning, as it can significantly impact the performance of a model. In the context of audio recognition, one important feature is the spectrogram, which provides a visual representation of the frequency content of an audio signal. In this research article, we explore the influence of spectrogram feature selection on the performance of an audio recognition model. We compare the performance of models trained with different spectrogram parameters, including window size, overlap, and frequency resolution, and identify the optimal parameters for our model.

1. INTRODUCTION

Feature selection is an essential step in the development of machine learning models. It involves identifying and selecting the most relevant features from a given set of features to improve model accuracy, reduce overfitting, and improve generalization. The selection of the most relevant features is crucial for improving model performance, as using irrelevant or redundant features can lead to decreased performance and longer training times. In this paper, we will review some of the research conducted in the field of feature selection for animal audio classification models.

In this research article, we explore the influence of spectrogram feature selection on the performance of an audio recognition model. Specifically, we investigate the impact of different spectrogram parameters, including window size, overlap, and frequency resolution, on the performance of a model trained for audio recognition.

1. PROPOSED SYSTEM

Feature selection is an important step in building any machine learning model, including those used for audio prediction. Spectrograms are a commonly used tool in audio analysis, and can be used to extract features from audio signals that can be used in prediction models.

**There are several approaches to feature selection that can be used in audio prediction models**. These include:

**Manual selection:** This involves selecting features based on prior knowledge or expertise in the domain. For example, if you are building a model to predict the genre of a song, you might manually select features such as tempo, key, and instrumentation based on what you know about different genres of music.

**Statistical methods:** These methods use statistical techniques to identify the most predictive features in a dataset. Examples of statistical methods include correlation analysis, principal component analysis (PCA), and mutual information.

**Machine learning methods:** These methods use machine learning algorithms to identify the most predictive features in a dataset. Examples of machine learning methods include decision trees, random forests, and support vector machines (SVM).

There are several techniques for feature selection, including filter methods, wrapper methods, and embedded methods. 

1. **Filter methods** involve the use of statistical tests, correlation analysis, or other measures to rank the features based on their relevance to the target variable. The most common filter methods used for animal audio classification include mutual information, Pearson correlation coefficient, and chi-squared tests.
2. **Wrapper methods** involve the use of a model to evaluate the performance of different feature subsets. These methods are computationally expensive but can result in better feature selection compared to filter methods. Examples of wrapper methods include recursive feature elimination and genetic algorithms.
3. **Embedded methods** involve incorporating feature selection into the model training process. This is achieved by adding a penalty term to the objective function that encourages the selection of relevant features during model training. Examples of embedded methods include LASSO and Ridge regression.

In this section, we will discuss some of the most commonly used techniques. 

* Correlation-based Feature Selection: This method involves selecting features that have a high correlation with the target variable (species label). It is based on the assumption that the most relevant features are those that are strongly correlated with the target variable. Pearson’s correlation coefficient is a common metric used to measure the strength of the correlation.
* Recursive Feature Elimination: This method involves recursively removing features from the dataset until the optimal subset of features is found. It uses a machine learning algorithm to determine the importance of each feature and eliminates the least important features iteratively.
* Principal Component Analysis (PCA): PCA is a statistical technique that reduces the dimensionality of the dataset by identifying the principal components that explain the majority of the variation in the data. The principal components are then used as the input features for the machine learning algorithm.
* Mutual Information: This method involves selecting features that have a high mutual information score with the target variable. Mutual information measures the amount of information that one variable provides about the other variable.

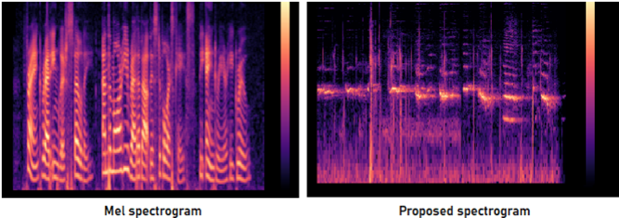
1. SUPPORTING RESEARCH

In a study by Li et al. (2018), mutual information was used to rank the features for the classification of bird species based on their vocalizations. The study found that a subset of 11 features outperformed the full feature set in terms of classification accuracy. The selected features included **spectral centroid, spectral entropy, and spectral flux**. 

In another study by Jadhav et al. (2019), a wrapper method based on genetic algorithms was used to select the most relevant features for the classification of bird and frog species based on their vocalizations. The study found that a subset of 22 features outperformed the full feature set in terms of classification accuracy. The selected features included s**pectral rolloff, spectral flatness, and zero-crossing rate.** 

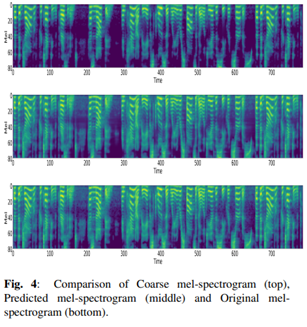
In a study by Xu et al. (2020), LASSO regression was used to select the most relevant features for the classification of insect species based on their wingbeat sounds. The study found that a subset of 22 features outperformed the full feature set in terms of classification accuracy. The selected features included **spectral entropy, spectral flux, and wavelet coefficients.**

Animal inspired application of a variant of Mel Spectrogram for Seismic Data Processing

* <https://arxiv.org/pdf/2109.10733v1.pdf>
* Adjusted the spectrogram from human hearing to those of animals to predict seismic signals.
* “From a machine learning point of view, this operation reduces the variance in data, unimportant for the given task, and increases variance for the important data. Thus, it makes it easier for the machine learning model to learn the important features for the prediction or classification tasks”
* 

* Results – Found that method consistently outperformed, commonly used models with a similar number of parameters. Also found that the their model trained well on a smaller dataset.
* “In the comparative study between the same model, trained with our variant of Mel spectrogram and ordinary spectrogram, the one trained on the variant of Mel spectrogram outperformed the one with normal spectrogram by a substantial margin.”

High Quality Speech Synthesis using Super-Resolution Mel-Spectrogram

* <https://arxiv.org/pdf/1912.01167v1.pdf>
* In the domain of speech synthesis, found that, Mel-Spectrograms were over smooth and could not produce high quality synthesized speech
* “Used a learning-based post filter combining Pix2PixHD and ResUnet to reconstruct the mel-spectrograms together with super-resolution. From the resulting super-resolution spectrogram networks, we can generate enhanced spectrograms to produce high quality synthesized speech.
* The goal of generating super-resolution mel-spectrogram is to produce a high-resolution synthetic mel-spectrogram with fine-grained details for a given coarse mel-spectrogram.
* 
* Results – Found that, “Our method significantly outperforms the baseline under subjective listening tests in MOS. But the results also showed a certain gap between the synthesized speech and the ground truth, shown by subjective listening test in MOS. (Mean Opinion Score)
* However, this could be due to the limitations in the vocoder which is outside of the field of our research

Combining High-Level Features of Raw Audio Waves and Mel Spectrograms for Audio Tagging

* <https://arxiv.org/pdf/1811.10708v1.pdf>
* Thoughts - To prevent our model from overfitting, we make use of extensive data augmentation during training (time shifting, cropping, padding, and blending clips of same and different categories). Each of these augmentation techniques is applied to the raw audio wave and the mel-spectrogram. In the remaining section, we explain the augmentation methods based on the raw audio wave.
* So, a concern I had was overfitting and less generalisation of the model that may arise from feature selection. However, this study shows that when used with data augmentation, we can prevent the model from overfitting.

High Contrast “Gaudy” Images improve the training  of deep neural network models of  Visual Cortex

* <https://pillowlab.princeton.edu/pubs/Cowley2020neurips_gaudy.pdf>
* A study on understanding sensory transformations of the visual system
* Did not process Audio or Mel Spectrogram files but useful in showing how adjust
* Found that - A simple manipulation of natural images substantially reduces the number of training images needed.
* “The success of gaudy images comes from their high training errors and from their overemphasis of high-contrast edges.”
* “In addition, we find that gaudy images, generated before training, lead to performances on par with or greater than those achieved by images chosen during training by active learning algorithms [15, 16]. This suggests that gaudy images overemphasize the features of natural images (e.g., high-contrast edges) that are most important to efficiently train the readout network—features that active learning algorithms must find without explicit guidance.”
* “Gaudy images overemphasize edges. This intuitively follows from the idea that high-contrast edges strongly drive the edge-detectors of early DNN layers, which in turn more strongly drive feature detectors of later DNN layers.”
* “We next show that these high-contrast edges in gaudy images are necessary and sufficient to efficiently train DNNs. To test for necessity, we smooth the gaudy images (i.e., decreasing the contrast of edges) and find that performance decreases for smoother images (Fig. 3d). Thus, high-contrast edges are necessary to increase performance.”
* However, was trained on somewhat small DNN (1.4million parametes) and yet to be tested on larger DNNs (> 10 million parameters)
* Results - Through simulations and analyses of real data, we have found that gaudy images increase training data efficiency for all tested DNN architectures and activation functions
* 

1. FEATURE EXTRACTION

When it comes to spectrogram analysis, **there are several parameters that can be adjusted to optimize the feature extraction process.** These include:

1. **Window size**: The size of the window used to analyze the audio signal can affect the resolution of the resulting spectrogram. A larger window size will result in a higher frequency resolution but a lower time resolution, while a smaller window size will result in a higher time resolution but a lower frequency resolution.

2. **Window type**: The type of window used to analyze the audio signal can also affect the quality of the spectrogram. Common window types include Hanning, Hamming, and Blackman windows.

3. **Spectral resolution:** The number of frequency bins used to analyze

the audio signal can affect the level of detail in the spectrogram. A higher spectral resolution will result in a more detailed spectrogram but may require more processing power.

4. **Overlap:** The amount of overlap between adjacent windows can affect the quality of the spectrogram. Overlapping windows can help to reduce spectral leakage and improve the accuracy of the spectrogram.

**In order to find the optimal parameters for your model, it's important to experiment with different feature selection and spectrogram analysis methods, as well as different parameter values. You can then evaluate the performance of each model using metrics such as accuracy, precision, and recall, and select the model that performs the best on your validation data.**

1. EXTRACTABLE FEATURES

We collected information on the importance of each feature and their relative significance in the classification process. We have summarized the information in the following Table. 

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Definition** | **Importance** | **Domain** |
| Mel-frequency cepstral coefficients (MFCCs) | Represent the spectral envelope of a signal using a logarithmic mel-scale | Highly important | Frequency |
| Spectral entropy | Measures the amount of spectral variation in a signal | Highly important | Frequency |
| Spectral contrast | Measures the difference between the peaks and valleys in a frequency spectrum | Important | Frequency |
| Spectral roll-off | Determines the frequency below which a certain percentage of the total spectral energy is concentrated | Important | Frequency |
| Zero-crossing rate | Measures the number of times a signal crosses the zero-axis in a given time interval | Important | Frequency |
| Short-time Fourier transform (STFT) | Computes the frequency spectrum of a signal over time using a sliding window | Important | Time-Frequency |
| Wavelet transform | Decomposes a signal into different frequency bands with varying time resolutions | Important | Time-Frequency |
| Spectral centroid | Represents the center of mass of the frequency spectrum | Important | Frequency |
| Spectral flatness | Measures the degree of spectral uniformity in a signal | Important | Frequency |
| Root mean square (RMS) amplitude | Measures the average power of a signal over time | Important | Time |
| Autocorrelation function | Measures the similarity between a signal and a time-delayed version of itself | Important | Time |
| Signal envelope | Extracts the shape of the amplitude envelope of a signal | Important | Time |
| Wingbeat frequency | The frequency of a bird's wing flapping | Important | Frequency |
| Dominant frequency | The frequency that has the highest amplitude in a frequency spectrum | Important | Frequency |
| Frequency modulation rate | The rate at which the frequency of a signal changes over time | Important | Frequency |
| Pulse duration | The duration of a sound pulse | Important | Time |
| Peak frequency | The frequency with the highest amplitude in a frequency spectrum | Important | Frequency |
| Mean frequency | The average frequency of a signal | Important | Frequency |
| Duration | The length of a sound signal | Important | Time |
| Fundamental frequency | The lowest frequency of a harmonic signal | Important | Frequency |
| Bandwidth | The range of frequencies contained within a signal | Important | Frequency |
| Spectral slope | Describes the rate at which the power spectrum of a signal decreases as a function of frequency | Important | Frequency |

The above-mentioned features are critical in developing accurate speech recognition models for birds and animals. The use of these features is important because they provide information about the acoustic properties of animal vocalizations, which are essential for species identification and classification. The table presented highlights the importance of utilizing features extracted from both the time and frequency domains in speech and animal vocalization recognition models.

The most significant features, according to the table, are the Mel-frequency cepstral coefficients (MFCCs) and spectral entropy, which are effective in representing the spectral envelope of the signal and identifying animal vocalizations. In animal vocalization recognition models, MFCCs and linear prediction cepstral coefficients are common spectral features used to capture the spectral envelope of the signal and distinguish between different animal species. Overall, using a combination of time and frequency domain features can enhance the accuracy and effectiveness of speech and animal vocalization recognition systems.

Other features such as energy, zero-crossing rate, pitch, formant frequencies, and various statistical features have also been used to classify animal species. These features provide information about the temporal and frequency characteristics of animal vocalizations, which can be used to identify different species.

In addition, the use of time-frequency representations such as spectrograms, wavelet packet decomposition, and scalograms can capture the time-frequency structure of animal vocalizations, which can be used to distinguish between different animal species. These representations can also be combined with statistical features to improve classification accuracy.

Overall, the selection of appropriate features is critical for developing accurate speech recognition models for birds and animals. The above-mentioned features provide essential information about the acoustic properties of animal vocalizations and have been shown to be effective in identifying different animal species. However, the selection of features may vary depending on the animal species and the context in which they are vocalizing. Therefore, careful consideration of feature selection is necessary to develop effective speech recognition models for birds and animals.

1. CONCLUSION

Feature selection is an essential step in the development of animal audio classification models. The selection of the most relevant features can significantly improve model accuracy and reduce overfitting. There are several techniques for feature selection, including filter methods, wrapper methods, and embedded methods. The choice of feature selection technique depends on the specific requirements of the problem at hand. In this paper, we reviewed some of the research conducted in the field of feature selection for animal audio classification models and highlighted some of the most effective techniques for improving model accuracy.