

Project 2: Weakly Supervised Learning Applied to the Detection of Manatee Calls

Connor McCurley

Machine Learning for Time Series, Spring 2018

University of Florida

Gainesville, FL, USA 32611

Email: cmccurley@ufl.edu

Abstract—This paper discusses the use of a Weakly Supervised learning system applied to the objectives of audio segmentation and change detection in signal processing. Mel-Frequency Cepstrum Coefficient (MFCC) features were independently generated on audio signals representing manatee calls and background noise. Multiple Instance (MI) learning was applied to determine representative target signatures for each manatee observation enacted in the training dataset. Confidence scores for each MFCC frame of a noisy test signal were realized through the evaluation of an Adaptive Cosine Estimator (ACE) with the signatures learned through MI. Experiments were conducted to determine the optimal parameters for signature creation. Detector performance was evaluated using Receiver-Operator Characteristic (ROC) curves. The optimal detector in this experimentation achieved approximately 0.85 area under the curve (AUC), demonstrating the potential for MI methods in audio segment classification and change detection. It is believed that performance could be improved by generating multiple signatures for each target class to better represent non-stationary in the manatees' voicings.

Index Terms—Weakly Supervised, Multiple Instance Learning, Adaptive Cosine Estimator, Mel-Frequency Cepstrum Coefficients

I. INTRODUCTION

AUTOMATIC audio segment classification is a difficult problem offering colossal potential for advancement in the fields of biometrics, biology, signal processing, computer science, consumer electronics and more. This paper focuses to determine the presence of manatee calls in a noisy background to be used in gaining an enlightened understanding of manatee behavior and intra-ecosystem interactions.

An expanse of approaches has been taken in attempt of solving the category detection and classification problem in audio signals. [1], [2] used statistical change detection to determine when a new class was represented in the data. [19]–[21] implemented Hidden Markov Models to determine if target classes were present in a given audio segment. [22]–[24] utilized deep neural networks for use in speaker recognition and verification. Both HMMs and ANNs are inherently able to emulate natural audio characteristics, which make them valuable tools in audio classification. Researchers in [13]–[18] employed Support Vector Machines using a slew of audio-based features to discriminate between classes represented in data.

This paper, however, explores the use of Multiple Instance Learning (MI) for detection in audio signals [9]–[12]. MI is a form of Weakly Supervised Learning which assumes algorithms have access to only high-level concepts of labels. In other words, sample-level labels are not provided to the learning processes. Given that a particular category cannot be represented, and thus classified, in audio given only a single sample, this approach seemed appropriate.

The remainder of this paper is organized as follows. Section II describes the methodology used to perform audio segment classification and discusses quantitative evaluation procedures. Detection results are presented in Section III. Practical insights to results are given in Section IV. Finally, Section V reveals concluding remarks and discusses future lines of research.

II. METHODOLOGY

A. Data Analysis

The data was first plotted as shown in figure 1 to gain an understanding of the format. The test signal was hand-labeled using the author's auditory discretion. Next, the Fast-Fourier Transforms (FFT) and spectrograms were computed and visualized for each training segment to gain an intuition about possible discriminative features (figures 2,3,4, and 5). From the spectral information, it was determined that frequency content could potentially be exploited to distinguish manatees from background noise. This lead to the implementation of Cepstrum Coefficients as features.

B. Feature Generation

Mel Frequency Cepstrum Coefficients (MFCC) have been proven as adequate discriminatory features in audio classification tasks [3]–[8], [14]. It was assumed that the signals evaluated in this paper could be appropriately deemed as stationary for short segments in time. MFCCs take advantage of this property by estimating the short-term energy spectrum for small windows of data before scaling the content to levels which are in line with the human auditory system. In this paper, MFCCs were computed for the training and test sets following the flow of figure 6. MFCC frames were taken in 25 ms segments with 10 ms of overlap. The number of coefficients calculated for each frame was set at 13. These

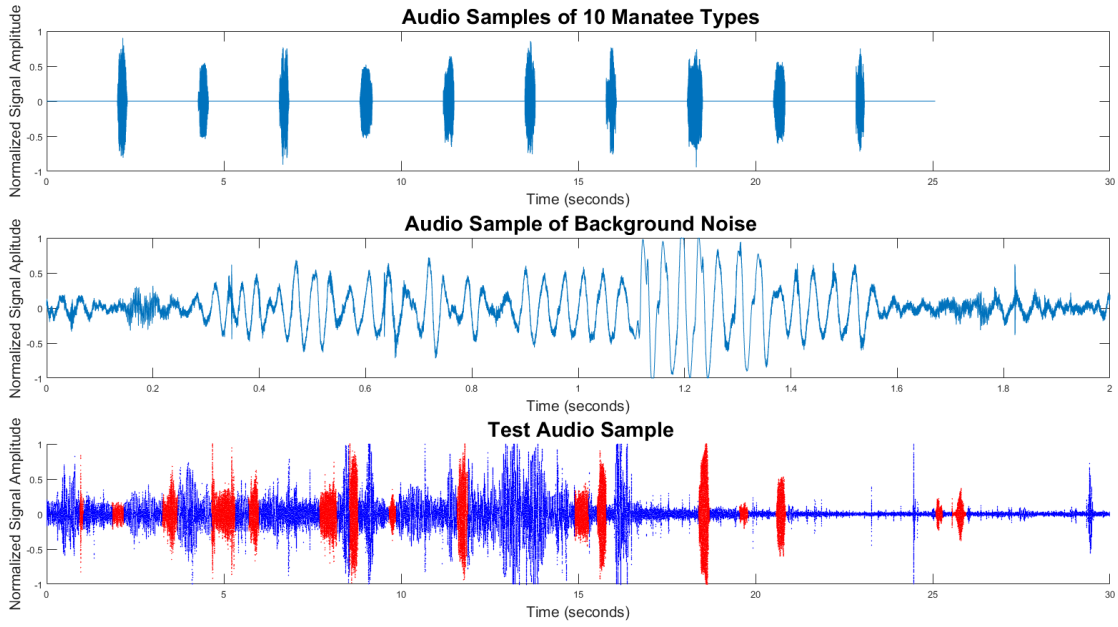


Fig. 1: Raw audio signals. Top: Training segments of 10 unique manatee categories. Middle: Raw training data for background noise. Bottom: Raw test signal colored by hand-created labels. Red portions signify segments from 16 unique manatee observations, while blue represents background noise.

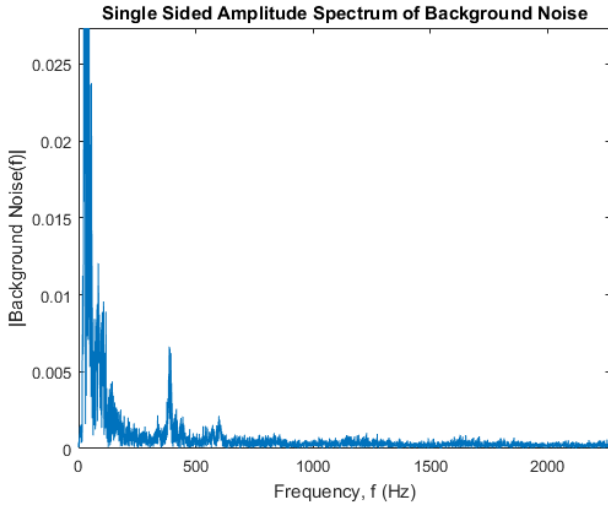


Fig. 2: FFT of background training data. Most of the spectral power exists in frequencies below 600 Hz.

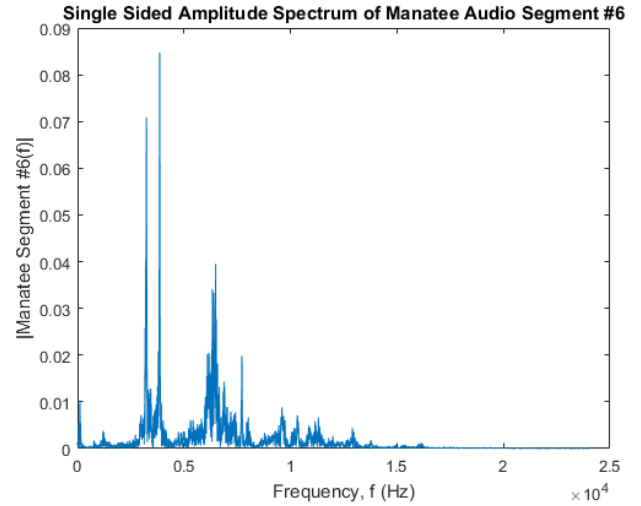


Fig. 3: FFT of a particular manatee training segment. The majority of frequency content is above 600 Hz.

settings are common for audio detection methodologies. A hamming window passing 600 Hz-20 kHz was utilized to cut off low-frequency information in hopes of making the data more discriminable. Examples of MFCC vectors computed on background and manatee training segments are shown in figures 7 and 8. It should be noted that test frames were labeled as containing manatee content if created from at least 1000 positive time series samples.

C. Multiple Instance Learning

Multiple Instance (MI) Learning is a form of Weakly Supervised Learning, meaning that only high level descriptions of labels are accessible in the form of concepts known as "bags". A negative bag is a subset of feature vectors which solely contains instances, or samples, from classes other than the target class. Alternatively, a positive bag must contain *at least* one positive instance, or target sample, but may also contain negative instances. In this way, learning algorithms

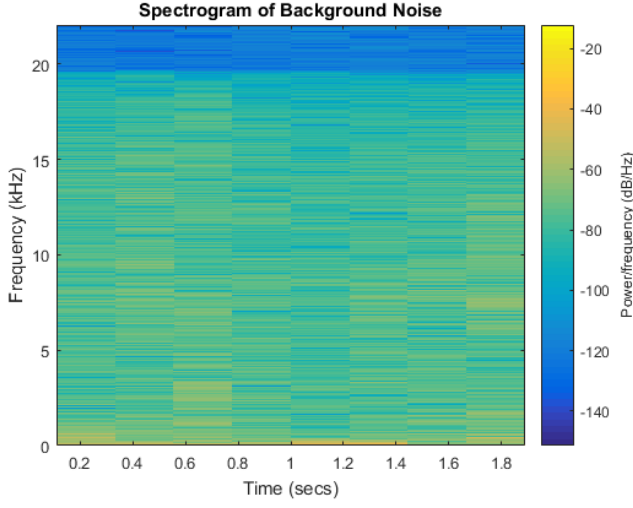


Fig. 4: Spectrogram of background training data.

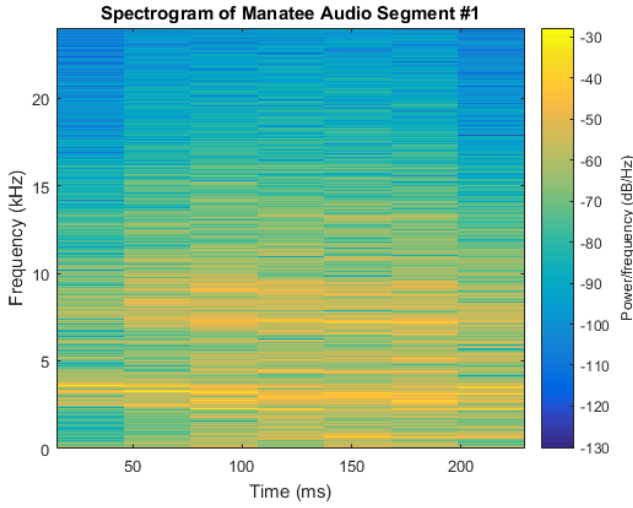


Fig. 5: Spectrogram of a training segment for a particular manatee category.

are provided bag-level labels. This is in contrast to the sample-level labels required by Supervised Learning tasks. Under the assumption that frequency content could discriminate between manatee and background signal segments, this approach to learning was intuitive.

The objective of MI in this paper was to learn signatures, given manatee and background MFCC feature vectors, which could accurately represent manatee MFCCs. Although there were many ways to approach this task, k-means clustering for example, the author was interested in employing the approach developed by Jiao et. al. This method, named miACE, uses a hybrid Dictionary Learning algorithm to develop representative signatures given bags of feature vectors. Refer to [10] for full details on miACE implementation. Target signatures were learned using data bagging described in section II-E. In most cases, a representative signature was created for each manatee

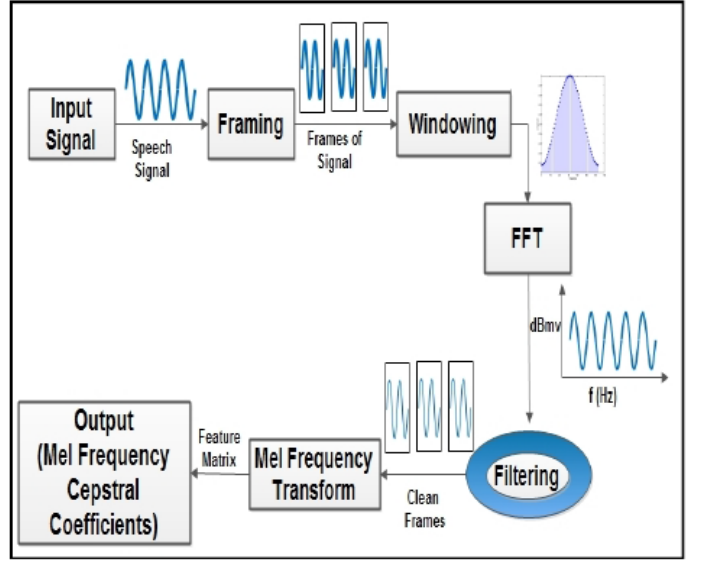


Fig. 6: Block diagram demonstrating the generation of MFCCs from raw audio signals [3].

category in the training data (10 in total) and added to a target signature dictionary to be used in detection scoring.

D. Adaptive Cosine Estimator

The Adaptive Cosine Estimator (ACE) is a measure of the cosine angle between vectors. It was assumed in this paper that ACE would make an appropriate distance metric to distinguish manatee MFCC vectors from background. Equation 1 measures the cosine of the angle between a representative target vector, \mathbf{s} and a test MFCC vector, \mathbf{x} minus the mean of the background, μ_b in a space whitened by the background covariance, Σ_b . The feature whitening adds scale-invariance to the comparison. Therefore, ACE simply measures the similarity of test features by their shapes.

$$D_{ACE}(\mathbf{x}, \mathbf{s}) = \frac{\mathbf{s}^T \Sigma_b^{-1} (\mathbf{x} - \mu_b)}{\sqrt{\mathbf{s}^T \Sigma_b^{-1} \mathbf{s}} \sqrt{(\mathbf{x} - \mu_b)^T \Sigma_b^{-1} (\mathbf{x} - \mu_b)}} \quad (1)$$

ACE produces a score between 0 and 1 signifying the shape similarity of the vectors compared. A score of 1 represents perfect shape matching.

E. Experiments

ACE was utilized as the detection metric for the test signal. For each method of test signature generation, ACE was computed between the test frames and each of the target signatures in the dictionary. The highest ACE confidence (closest to one) among all signatures was taken as the confidence for an individual frame. Since the signatures in the dictionary were created to represent manatee classes, a low ACE confidence on a test frame would deem that particular frame as belonging to background. Target signatures were computed in the following fashion:

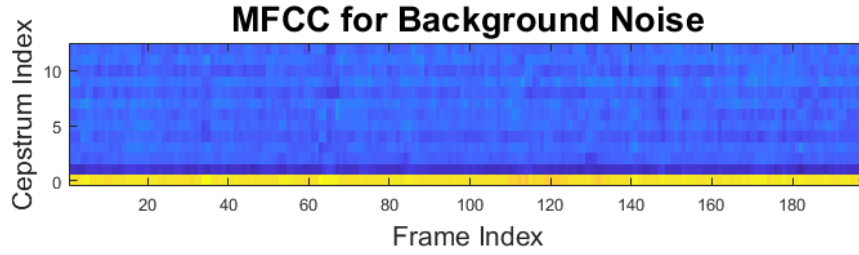


Fig. 7: 200 frames of 13 MFCCs representing background noise.

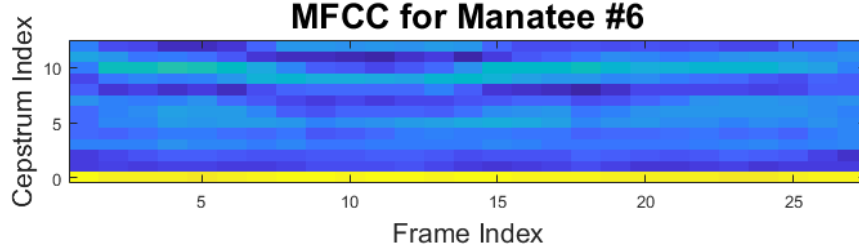


Fig. 8: 30 frames of 13 MFCCs representing a particular manatee category.

- 1) 1 signature for each manatee category was learned by taking the average of the features representing each class (10 signatures)
- 2) 1 signature for each manatee category was learned by taking the median of the features representing each class (10 signatures)
- 3) 1 signature was learned in total using a single negative bag and a single positive bag containing instances of every manatee category (1 signature)
- 4) 1 signature was learned for each manatee category using one negative bag and 1 positive bag containing instances of only a single manatee class (10 signatures)
- 5) 1 signature was learned for each manatee category using 6 negative bags and 6 positive bags, each containing instances of only one manatee class (10 signatures)
- 6) 1 signature was learned for each manatee category using 10 negative bags and 10 positive bags, each containing instances of only one manatee class (10 signatures)

Classifier results were quantified in the form of Receiver-Operator Characteristic (ROC) curves. Classification performance is provided in section III.

III. RESULTS

In this section, classification results for each experimental method tested are presented in the form of ROC curves. A ROC curve measures the probability of correct classification for varying thresholds of class division. Essentially, they define how well a scoring algorithm can predict the correct class while minimizing the number of false positive classifications. A value of 1 for the area under the curve (AUC) signifies zero mis-classifications among all categories.

A. ROC Curves

Figure 9 demonstrates the ROC curves for varying methods of target signature generation: average of manatee features (blue), median of manatee features (red), one signature created from a single positive bag with instances from each manatee category (green), 1 signature for each manatee category using one manatee class per positive bag (magenta), 1 signature for each manatee category where each signature was trained using one manatee class separated into 6 positive bags and 6 negative bags (black), and 1 signature for each manatee category where each signature was trained using one manatee class separated into 10 positive bags and 10 negative bags (cyan). Table I reveals the total detection accuracy. It can be seen that the best performance among the methods compared was provided by creating 10 signatures, one for each manatee category, by taking the median of the representative MFCC features. This result, shown by the red curve, achieved an AUC of 0.8414. The worst performance was exhibited by the single manatee signature representing all 10 manatee classes (green), demonstrating an AUC of 0.4472. Each of the other detectors averaged 0.8 AUC.

IV. DISCUSSION

In this sections, observations are made on results and insight is given to potential influences.

A. Results

The best detector exhibited in this paper obtained an AUC score of approximately 0.84. This is in line with state-of-the-art detectors given the un-solved problem of audio classification. From figure 9, it can be determined that the optimal methods for target signature generation did not actually stem

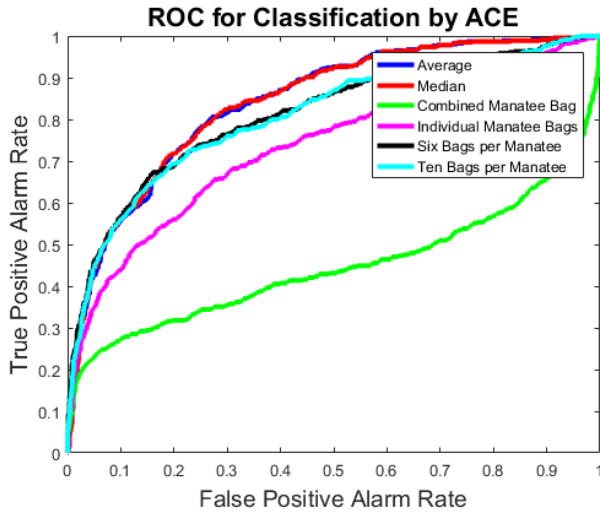


Fig. 9: Receiver-Operator Characteristic curves for ACE detection using various target signature generation methods. Optimal AUC is seen by creating the target signatures using Median (red), followed by Average (blue), six bags per manatee (black), 10 bags per manatee (cyan), individual manatee bags (magenta), and combined manatee bag (green).

TABLE I: AUC for Various Signature Generation Methods

Signature Creation	AUC
Average	0.8406
Median	0.8414
Combined Manatee Bag	0.4472
Individual Manatee Bags	0.7379
Six Bags per Manatee	0.8133
Ten Bags per Manatee	0.8085

from the MI approach, but from simply taking the medians and averages of manatee category feature vectors. The signatures created to represent all manatee classes provided the worst performance among the methods compared, which was expected. Given that the manatee category sets showed large variations in statistical information between each other, it is intuitive that a single signature could not adequately detect all manatee classes. Additionally, signatures trained using multiple positive and negative bags demonstrated detection accuracy which was significantly better than those trained on single positive and negative bags which, again, aligned with prior intuition. Through observation of the miACE objective function in [10], one can deduce that MI methods thrive when given many moderate subsets of data, in contrast to fewer large bags. Therefore, given a greater set of training data, more bags could be created to potentially aid in developing better-representative signatures. Additionally, given that both manatee and background audio segments exhibited non-stationarity, the author believes it would be advantageous to learn multiple signatures for each manatee category (instead of just 1) to

more accurately capture the possible frequency characteristics exhibited by each manatee class. This extension to miACE is currently being explored at the University of Florida.

Additionally, windowing was utilized to generate the MFCC features. While this approach is beneficial in the sense that it created invariance between differing sampling frequencies, it also exhibited a few problems. One, the parameters of windowing, namely window length and percentage of overlap between frames, was immensely difficult to choose and played a large role in detection performance. Also, windowing assumes stationarity among the samples within the window. This is detrimental when the samples in a window lay on a border for statistical variation. Addressing each of these issues would likely increase algorithmic performance.

Finally, the experiments conducted in this paper assumed that ACE would be an adequate distance metric for confidence scoring. However, this does not imply that there does not exist a superior metric. The author would like to explore alternative measures such as a simple Euclidean distance or Information Theoretic approach to test the effects on discrimination between manatee and background instances.

B. Potential Improvements

In addition to the methods mentioned above, two mechanisms which could aid with manatee classification from background include anomaly detection pre-processing and a post-processing method. Anomaly detection procedures such as the approach developed in [25] utilize a linear or nonlinear adaptive filter to predict background noise. When a target presents itself in the data, a large spike in prediction error occurs. This method could be used to localize potential manatee audio segments in the test data so that a smaller subset would require scoring. Using anomaly detection as a pre-processing method would likely decrease the number of false alarms exhibited in detection. Also, score smoothing could be applied to potentially improve classification accuracy. Given that manatee calls generally last longer than the 25 ms window, it is highly unlikely that the true window labels would change with high frequency. Therefore, smoothing could be applied to lower the chance for mis-classifications, thus, improving detection performance.

Many of the potential improvements discussed will be explored by the author in future lines of research.

V. CONCLUSIONS

An audio segment classification process employing Weakly Supervised learning was presented in this paper. Target signatures representing 10 separate manatee classes were learned through Multiple Instance methods and used to discriminate test segments from background noise with an ACE scoring algorithm. While the signatures learned from MI provided AUCs above 0.8, they did not contribute to the optimal discriminator among the methods compared.

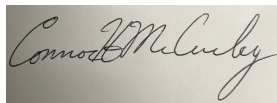
The author believes that target/noise separation and classification could be improved through the employment of several methods. First, anomaly detection could be applied on the test

signal to localize potential manatee class samples, thereby diminishing the total number of points tested and providing less room for false alarms. Next, the number of positive and negative bags utilized by MI could be increased. This would inherently increase the information provided in representative signature generation. Additionally, multiple target signatures could be learned for each manatee class to more accurately represent any non-stationary components in the manatees' calls. Finally, it was assumed that frames were adequately discriminable by the cosine angle of feature vectors. However, detection performance could potentially be improved through the use of an alternative cost function such as Euclidean Distance or an Information Theoretic method.

Future research endeavors towards this topic include, among those listed previously, comparison to common alternative audio classification methodologies such as: Vector Quantization, Artificial Neural Networks, Hidden Markov Models, traditional and Multiple Instance Support Vector Machines and Adaptive Filtering.

HONOR STATEMENT

* I confirm that this assignment is my own work, it is not copied from any other person's work (published or unpublished), and has not been previously submitted for assessment either at University of Florida or elsewhere.



REFERENCES

- [1] Liu, Weifeng., Principe, Jose C., Haykin, Simon. "Chapter 1 - Background and Preview", "Chapter 2 - Kernel Least-Mean-Square Algorithm", in *Kernel Adaptive Filtering: A Comprehensive Introduction*, New Jersey: John Wiley & Sons, Inc., Publication, 2010, pp. 1-26, 27-68.
- [2] Fancourt, Craig L., Principe, Jose C. "Exploiting Multi-Modality for Segmentation and Modeling of Non-Stationary Time Series" in *Nonlinear dynamical systems : feedforward neural network perspectives*. New York : John Wiley, c2001., 2001. (Adaptive and learning systems for signal processing, communications, and control). ISBN: 0471349119.
- [3] Chauhan, P. M., Desai, N.P. "Mel Frequency Cepstral Coefficients (MFCC) Based Speaker Identification in Noisy Environment Using Wiener Filter" in *Proceeding of the IEEE International Conference on Green Computing, Communication and Electrical Engineering*. 2014.
- [4] Grama, L, Rusu, C. "Choosing an accurate number of mel frequency cepstral coefficients for audio classification purpose" in *International Symposium on Image and Signal Processing and Analysis*. 2017. pp. 225-230.
- [5] Tiwari, V. "MFCC and its applications in speaker recognition" in *International Journal on Emerging Technologies*. 2010. pp. 19-22
- [6] Nijhawan, G., Soni, M. K. "Speaker Recognition Using MFCC and Vector Quantisation" in *International Journal on Recent Trends in Engineering and Technology*. 2014. vol. 11. pp. 211-218.
- [7] Bharti, R., Bansal, P. "Real Time Speaker Recognition System using MFCC and Vector Quantization Technique" in *International Journal of Computer Applications*. 2015. vol. 117. pp. 25-31.
- [8] Prahallad, K. "Topic:Spectrogram, Cepstrum and Mel-Frequency Analysis" in *Speech Technology: A Practical Introduction*.
- [9] Carbonneau, M., Cheplygina, V., Granger, E., Gagnon, G. "Multiple Instance Learning: A Survey of Problem Characteristics and Applications" 2016. Available: <http://arxiv.org/abs/1612.03365>
- [10] Jiao, C., Zare, A., Mcgarvey, R. "Multiple Instance Hybrid Estimator for Hyperspectral Target Characterization and Sub-pixel Target Detection". Under Review. 04/19/2019. Available: <https://arxiv.org/pdf/1710.11599.pdf>
- [11] Kumar, A., Raj, B. "Weakly Supervised Scalable Audio Content Analysis" 2016. Available: <https://arxiv.org/pdf/1606.03664.pdf>
- [12] Carbonneau, M. "Introduction to Multiple Instance Learning". 10/19/2016. Online. Available: <https://www.etsmtl.ca/Unites-de.../Introduction-to-Multiple-Instance-Learning.pdf>
- [13] Andrews, S. Tsochantaridis, I., Hofmann, T. "Support Vector Machines for Multiple-Instance Learning". Online. Available: <https://pdfs.semanticscholar.org/3447/fe054f6af70403cfc39b4d21076337a71128.pdf> Accessed: April, 2018.
- [14] Fagerlund, S. "Bird species recognition using support vector machines" in *Eurasip Journal on Advances in Signal Processing*. 2007.
- [15] Rong, F. "Audio classification method based on machine learning" in *Intelligent Transportation, Big Data & Smart City*. 2016.
- [16] Lin, C., Chen, S., Truong, T., Chang, Y. "Audio Classification and Categorization Based on Wavelets and Support Vector Machine" in *IEEE Transactions on Speech and Audio Processing*. 2005. vol. 5. pp. 644-651.
- [17] Guo, G., Li, S.Z. "Content-Based Audio Classification and Retrieval by Support Vector Machines" in *IEEE Transactions on Neural Networks*. 2003. vol. 14 no. 1, pp. 209-215.
- [18] Andrews, S., Tsochantaridis, I. Hofmann, T. "Support Vector Machines for Multiple-Instance Learning". Online. Available: http://www.robots.ox.ac.uk/~vgg/rg/slides/07_06_11_Chai.pdf. Accessed: April, 2018.
- [19] Jancovic, P., Kokuer, M. "Automatic Detection of Bird Species from Audio Field Recordings using HMM-based Modeling of Frequency Tracks" in *25th European Signal Processing Conference*. 2017.
- [20] Brognaux, S., Drugman, T. "HMM-Based Speech Segmentation: Improvements of Fully Automatic Approaches" in *IEEE/ ACM Transactions on Audio, Speech, and Language Processing*. 2016. vol. 24, no. 1, pp. 5-15.
- [21] Zhu, B., Gan, J., Cai, J., Wang, Y., Wang, H. "Adaptive onset detection based on instrument recognition" in *ICSP2014, Proceedings of*. 2014.
- [22] Chauhan, N., Chandra, M. "Speaker Recognition and Verification Using Artificial Neural Network" in *IEEE WiSPNET Conference*. 2017,
- [23] Fredes, J., Novoa, J., King, S., Stern, R.M., Yoma, N.B. "Locally Normalized Filter Banks Applied to Deep Neural-Network-Based Robust Speech Recognition" in *IEEE Signal Processing Letters*. 2017. vol. 24, no. 4, pp. 377-381.
- [24] Richardson, F., Reynolds, D., Dehak, N. "Deep Neural Network Approaches to Speaker and Language Recognition" in *IEEE Signal Processing Letters*. 2015. vol. 22, no. 10, pp. 1671 - 1675.
- [25] Ho, K.C., Gader, P.D. "A Linear Prediction Land Mine Detection Algorithm for Hand Held Ground Penetrating Radar" in *IEEE Transactions on Geoscience and Remote Sensing*. 2002. vol. 40, no. 6, pp. 1374-1384.