Bird Audio Detection using Supervised Weighted NMF

Technical Report

Soroush Jamali

Electrical Engineering Department, Hamedan University of Technology
Hamedan 6516913733, Iran
soroushjamali@stu.hut.ac.ir

Juan Ahmadpanah

Electrical Engineering Department, Hamedan University of Technology
Hamedan 6516913733, Iran
juan.ahmadpanah@stu.hut.ac.ir

Ghasem Alipoor

Electrical Engineering Department, Hamedan University of Technology
Hamedan 6516913733, Iran
alipoor@hut.ac.ir

ABSTRACT

This paper reports on the results of our bird audio detection system, developed for Task 3 of the DCACE 2018, challenge that is defined as a binary classification problem. Our proposed method is based on supervised non-negative matrix factorization (NMF) of the constant-Q transform (CQT) spectrogram. Two dictionaries are trained over the training data available for the bird and environment classes. Test samples are then linearly decomposed using a combined dictionary, generated by concatenating these two dictionaries. Classification is performed based on the energy of the activations relevant to each class. However, to further improve the classification performance, we propose to weight each activation coefficient according to the contribution of its corresponding basis in constructing each class. A scheme is proposed to extract this contribution weights from the activation coefficients of the training data. The developed system, evaluated over the development dataset of the challenge, results in up to 80% accuracy.

Index Terms— Bird Audio Detection, DCASE Challenge, Non-negative Matrix Factorization, Constant-Q Transform, Basis Contribution

1. INTRODUCTION

Bird sound can provide valuable and reliable real-time information on wildlife. Bird audio detection (BAD) aims at automatically identifying the presence or absence of bird from audio recordings. It can be considered as a special application of the more general task of audio event detection (AED). BAD has numerous applications in environmental science, e.g. scheduled and continuous data collection, unattended monitoring of biodiversity, assessment of migration and range shift and reliable estimation of the population size and trend. Furthermore, such a system can serve as a preliminary step for some other tasks such as scene identification and species identification and classification. In parallel with the substantial progress in audio signal processing and pattern recognition techniques as well as the ever-growing quantities of bird audio recordings, bird audio detection and classification have attracted a considerable attention in recent years.

This paper reports on the results of our bird audio detection system submitted for Task 3 of the DCACE 2018 challenge that was developed based on the proposed classification method. Our proposed method is based on supervised non-negative matrix factorization (NMF) of the constant-Q transform (CQT) spectrogram. NMF can be used in unsupervised and supervised manners for classification. For supervised classification, one dictionary is usually trained for each class. Test samples are then linearly decomposed using a combined dictionary, generated by concatenating these dictionaries. Classification is performed based on either the energy of the activations relevant to each class or the reconstruction errors provided by each class dictionary. Typically, the activation coefficients associated with all bases of a class dictionary are treated equally, when classification is done based on the activations. However, we think that different bases of a class dictionary have different contribution or significant in constructing that class's data.

The activation coefficients of the training data can provide useful information in this regard. We propose to directly weight each activation coefficient of the given test data according to the contribution of its corresponding basis in constructing each class. A scheme is proposed to extract these contribution weights from the activation coefficients of the training data. This system is described in the next section. Section 3 is dedicated to the simulation results.

2. PROPOSED METHOD

In supervised NMF classification a separate dictionary is trained for each class from its isolated labeled training samples. The task of determining the presence or absence of a bird sound in an audio file can be considered as a two-class decision problem. Hence, we train one dictionary over the CQT spectrograms of the training data available for each class, i.e. the bird and the environment classes. The typical approach for supervised NMF classification and our proposed method to improve the performance of this approach are described in the following sub-sections.

2.1. Typical Supervised NMF Classification

Let $\mathcal{X} = \{x_b^1(t), x_b^2(t), \cdots, x_b^{N_b}(t)\}$ be the collection of N_b training audio samples in which a bird song/call presents. Similarly, $\mathcal{X} = \{x_e^1(t), x_e^2(t), \cdots, x_e^{N_e}(t)\}$ are N_e training bird-free audio samples. The CQT spectrogram of the ith training sample from each class is denoted as $V_c^i \in \mathbb{R}^{S\times L}$, where c can be either b (for bird class) or e (for environment class) and S and L are the number of frequency bins and the number of time frames, respectively. For convenience we consider that all the training (as well as testing) audio samples are of the same number of frames. For each class, the collection of all training spectrograms available for that class is used to train an NMF dictionary with K bases, as follows:

$$\left[V_b^1, V_b^2, \cdots, V_b^{N_b}\right] \approx W_b H_b$$
$$\left[V_e^1, V_e^2, \cdots, V_e^{N_b}\right] \approx W_e H_e$$

$W_b \in \mathbb{R}^{S\times K}$ and $W_e \in \mathbb{R}^{S\times K}$ are the dictionaries of bird and environment classes, respectively, and $H_b \in \mathbb{R}^{K\times LN_b}$ and $H_e \in \mathbb{R}^{K\times LN_e}$ are the corresponding activation matrices. Typically, dictionary and activation matrices are found to minimize a cost function of the reconstruction error.

For the classification of a given test sample $x(t)$, its CQT spectrogram, i.e. $V \in \mathbb{R}^{S\times L}$, is first decomposed using a combined dictionary, generated by concatenating the two trained class dictionaries. In other words, the activation matrix of the test spectrogram V over the fix dictionary $W = [W_b, W_e]$ is obtained by solving the following problem:

$$V \approx WH^{ts} = [W_b, W_e]\begin{bmatrix} H_b{}^{ts} \\ H_e{}^{ts} \end{bmatrix}$$

Classification can be then performed based on either the energy of the activation coefficients relevant to each class or the reconstruction errors provided by each class dictionary. In particular, for the activation-based decision, it is assumed that for the test data, from either of classes, the estimated activations associated with the corresponding dictionary are of greater values, as compared to the remaining activations. Hence, the classification can be simply done by assigning the given test sample to the class whose associated activations dominate.

2.2. Contribution Weighting

As described in the previous sub-section, the activations associated with all bases of a class dictionary are treated equally, when classification is done based on the activations. However, we think that different bases of a class dictionary have different contribution or significant in constructing that class's data. Furthermore, as a result of the redundancy of the regions spanned by the bases of different class dictionaries, referring to the decomposition (2), each class's bases can also intervene in representation of the other class's data. For instance, when a test sample from the bird class is decomposed over the concatenated dictionary W, the activations $H_e{}^{ts}$, associating to the environment dictionary bases, are not necessarily zero. Similarly, the bird dictionary $W_b$intervenes in representation of data from the environment class, via its associated activations $H_e{}^{ts}$.

To account for these phenomena, we propose to first weight each activation coefficient of the given test data, i.e. the elements of the matrix H^{ts}, according to the contribution of the dictionary bases in constructing each class. For this end, let $\boldsymbol{g} \in \mathbb{R}^{2K}$ be such a contribution weight vector. The activation matrix H^{ts} is first summed over all time frames to form the aggregative activation vector $\boldsymbol{h} \in \mathbb{R}^{2K}$, i.e.

$$\boldsymbol{h} = \sum_{i=1}^{L} H^{ts}(:, i)$$

This aggregative activation vector is then weighted with the contribution weight vector of the bases as $\boldsymbol{f} = \boldsymbol{g}\odot\boldsymbol{h}$, where $\odot$ denotes the Hadamard (element-wise) product. We define the total weighted activation energy of to the class bird and environment classes as $E_b = \boldsymbol{f}(1\!:\!K)^T\boldsymbol{f}(1\!:\!K)$ and $E_e = \boldsymbol{f}(K+1\!:\!2K)^T\boldsymbol{f}(K+1\!:\!2K)$, respectively. With these modifications, the classification can be now done by assigning the given test sample to the class with the greatest total weighted activation energy, i.e.:

$$c = arg\min_a E_a$$

where a is either b or e.

We suggest to extract the contribution weight vector based on the activation coefficients of the corresponding class's data. For this purpose we decompose the training data of each class using the combined dictionary W as follows:

$$\left[V_b^1, V_b^2, \cdots, V_b^{N_b}\right] \approx W{H_b}^{tr} = [W_b, W_e]\begin{bmatrix} {H_{bb}}^{tr} \\ {H_{be}}^{tr} \end{bmatrix}$$
$$\left[V_e^1, V_e^2, \cdots, V_e^{N_e}\right] \approx W{H_e}^{tr} = [W_b, W_e]\begin{bmatrix} {H_{eb}}^{tr} \\ {H_{ee}}^{tr} \end{bmatrix}$$

${H_{ad}}^{tr}$ represents the contribution of the bases of the class d in representing the data of the class a. Using the definitions $G_b \triangleq {H_{bb}}^{tr} - {H_{eb}}^{tr}$ and $G_e \triangleq {H_{ee}}^{tr} - {H_{be}}^{tr}$, the contribution weight vector is calculated as:

$$\mathbf{g} = \begin{bmatrix} \sum_{i=1}^{LN_b} G_b(:,i) \\ \sum_{i=1}^{LN_b} G_e(:,i) \end{bmatrix}$$

It worth noting that if we use two different contribution weight vectors $\boldsymbol{g}_b = [\mathbf{1} \quad \mathbf{0}]^T$ and $\boldsymbol{g}_e = [\mathbf{0} \quad \mathbf{1}]^T$ (where $\mathbf{1}$ and $\mathbf{0}$ are respectively all-one and all-zero row vectors of length K) for calculating the total weighted activation energy of two different classes, this approach reverts to the typical approach described in the previous sub-section.

3. EXPERIMENTAL RESULTS

The performance of the submitted system is evaluated using the development datasets of the DCASE 2018 challenge. Three datasets were provided for development, each from a separate bird sound monitoring project. All of the datasets contain 10-second-long mono wav files with sampling frequency of 44.1 kHz. Each 10-second audio clip have been manually labelled with a 0 or 1 to indicate the absence/presence of any birds within that file. We used 500 randomly selected files from one dataset for training each class dictionary, i.e. $N_b = N_b = 500$, and files of another dataset for testing. Results of the averaged performance of the proposed contribution weighting approach are compared against that of the typical supervised NMF-based classification approach in Table 1. In this test, the frame length as well as the number of frequency bins were set to S=463. With this choice, each 10-second audio sample contains L=6472 time frames. Furthermore, each class dictionary contains K=70 bases.

	Contribution Weighting	Typical Approach
Accuracy	78% - 83%	%69.50 – %70.50

Table 1. Performance evaluation of the proposed and typical approaches