# Spectrogram Enhancement Using Multiple Window Savitzky-Golay (MWSG) Filter for Robust Bird Sound Detection

Nithin Rao Koluguri, G. Nisha Meenakshi, *Student Member, IEEE*, and Prasanta Kumar Ghosh, *Senior Member, IEEE*

***Abstract*—Bird sound detection from real-field recordings is essential for identifying bird species in bioacoustic monitoring. Variations in the recording devices, environmental conditions, and the presence of vocalizations from other animals make the bird sound detection very challenging. In order to overcome these challenges, we propose an unsupervised algorithm comprising two main stages. In the first stage, a spectrogram enhancement technique is proposed using a multiple window Savitzky–Golay (MWSG) filter. We show that the spectrogram estimate using MWSG filter is unbiased and has lower variance compared with its single window counterpart. It is known that bird sounds are highly structured in the time–frequency (*T–F*) plane. We exploit these cues of prominence of T-F activity in specific directions from the enhanced spectrogram, in the second stage of the proposed method, for bird sound detection. In this regard, we use a set of four moving average filters that when applied to the enhanced spectrogram, yield directional spectrograms that capture the direction specific information. We propose a thresholding scheme on the time varying energy profile computed from each of these directional spectrograms to obtain frame-level binary decisions of bird sound activity. These individual decisions are then combined to obtain the final decision. Experiments are performed with three different datasets, with varying recording and noise conditions. Frame level F-score is used as the evaluation metric for bird sound detection. We find that the proposed method, on average, achieves higher F-score ($10.24\%$ relative) compared to the best of the six baseline schemes considered in this work.**

***Index Terms*—Bioacoustic monitoring, bird sound detection, directional spectrograms, Savitzky-Golay filter.**

## I. Introduction

ENVIRONMENT monitoring has become essential, as the increase in habitat loss and changes in the global climate are driving various species of flora and fauna to extinction [1], [2]. Bioacoustic monitoring is a popular method to study and aid conservation efforts of endangered animals and birds [3]. Bioacoustic signals typically capture the vocalizations of the different species which could shed some light on their behavior and interaction in their habitat. With the availability of the recording sensors [4], bioacoustic monitoring can furnish data throughout the day providing researchers and conservationists with rich information of the species under study and the environment they live in. Such efforts have been taken to identify and study birds [5], [6]; insects [7], [8]; and other animals such as frogs [9], elephants [10] etc. In monitoring birds, typically, identification or classification of the bird species from a bioacoustic signal is of interest. An essential step that precedes species classification is the bird sound detection from the recorded signal. Bird sound detection is the task of identifying bird acoustics in a given noisy bioacoustic audio and thereby segmenting the recording into bird sounds and noise. Segmentation is a critical step which can influence the performance of a bird call identification system [11]. Segmented data aids further analysis compared to the raw data, in terms of computational time, as it reduces search space for classification algorithms. This work focuses on automatic bird sound detection from a noisy bioacoustic audio signal in an unsupervised setting, i.e., automatically finding segments of time that contain bird sounds in the given audio signal.

Several attempts have been made to detect bird sounds for automatically identifying or classifying the bird species. Applications such as bird call retrieval are also related to the task of bird specifies identification [12]. Dong *et al.* characterized a variety of bird species based on local gradients and entropy based features to retrieve bird vocalizations, corresponding to a query vocalization, from large audio recordings using a $K$ nearest neighborhood approach [13]. Classification of bird calls based on a parametric representation has been addressed by Harma [14]. In his work, the syllables in the bird calls are segmented by employing sinusoidal model. Although robust to noise, this method can only be employed for birds that have a tonal or a near tonal call. A more popular segmentation scheme is based on thresholding the time-domain energy envelope [5], [11]. Somervuo *et al.* [5] reported that the performance of the segmentation scheme is found to be species dependent. Acoustic features computed using species specific front-end parameters have found to provide better results [15], [16]. Such species specific segmentation may not be suitable for a generic bioacoustic audio signal.

Manuscript received July 13, 2016; revised December 21, 2016; accepted March 24, 2017. Date of current version May 23, 2017. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Herve Glotin. *(Corresponding author: G. Nisha Meenakshi.)*

N. R. Koluguri was with the Electronics and Communication Engineering, National Institute of Technology, Surathkal, Karnataka 575025, India (e-mail: nithinrao.koluguri@gmail.com).

G. N. Meenakshi and P. K. Ghosh are with the Indian Institute of Science, Bangalore, Karnataka 560012, India (e-mail: gnisha@ee.iisc.ernet.in; prasantg@ee.iisc.ernet.in).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2017.2690562

Apart from the parametric and time-domain based approaches, an image processing point of view has been considered for bird call segmentation, where the spectrogram of the bioacoustic signal is treated as an image. Neal *et al.* [17] reported that a segmentation scheme exploiting both time and frequency signatures of bird sounds outperforms simple temporal energy based schemes. Here, a probability mask of the spectrogram is obtained via classification using a random forest classifier. Applying a threshold on this probability mask yields the binary mask of bird activity in the time-frequency (T-F) plane. Although designed for segmentation from noisy recordings, the classification scheme requires about 500,000 randomly-sampled training examples and a choice of a hard threshold. Wang *et al.* overcame the limitation of a hard threshold by using a Bayesian change point detection on a T-F patch based entropy profile computed from a de-noised spectrogram [18]. A limitation of this work is that the species specific information is necessary for feature computation. Feature such as the frequency bandwidth required for entropy calculation, is species dependent. Sequential modeling of features extracted from a filtered spectrogram via a Hidden Markov Model (HMM) has been done to segment the sounds based on onset detection [19]. The algorithm is reported to perform well across birds, insects and frogs, but requires manually annotated training data for each category. Perceptual linear prediction coefficients are also employed with HMMs to detect bird sounds from long field recordings [20]. In case of a supervised setting, manual annotation is necessary [21], [22], which would require an audio engineer expert and could be, both, time consuming and expensive. This motivates the need for a segmentation scheme in an unsupervised manner.

Several unsupervised bird sound detection schemes have been proposed in the literature. Morphological operations on the spectrogram have been found to be useful in detecting segments of bird calls based on energy thresholding [23]. Morphological filtering of the spectrogram is found to be an essential step in several works in the context of bird classification competitions such as MLSP 2013 [24], LifeCLEF 2014 [25] and NIPS 2013 [26]. These methods typically smoothen the spectrogram and obtain a T-F segmentation of bird sounds using morphological operations [27]–[29]. Although these algorithms perform well, the parameters in morphological filtering may need to be carefully selected for a given dataset and may not work well across different recording conditions.

In this work, we propose a robust unsupervised bird sound detection algorithm to segment bird sounds from a given noisy recording. Typical challenges include, variability in the recording environment, different noise conditions, intra species and inter species variability in the bird sounds. These challenges motivate a need for an algorithm that is dataset and species independent, which would not require parameter selection in a dataset or species specific manner. Since data available in different noise conditions could be limited, we propose an unsupervised methodology, where training data and their manual annotation are not required. We propose an algorithm that has two main stages designed to handle the aforementioned challenges.

The first stage is the spectrogram enhancement to make it robust to the variability in recording and noise conditions that result in a degradation of the spectrogram [14]. In this work, enhancement is performed by denoising a spectrogram to obtain a better T-F representation. In this regard, we propose a multiple window Savitzky-Golay (MWSG) filter. We formulate the denoising problem as a two dimensional Savitzky-Golay (SG) filtering operation on the T-F representation from the spectrograms computed using multiple windows. We show that for a chosen denoising SG filter, the estimated spectrogram at a time-frequency bin is unbiased and has lower variance when multiple windows are used as compared to when a single window is considered.

The second stage of the proposed methodology is used to handle the challenges introduced by the variability in the bird sounds. Bird sounds could have simple monosyllabic structures or complicated patterns of phrases that contain notes and syllables in a rhythmic pattern, resulting in varying T-F structures [30]. In order to capture different T-F structures of the bird sounds, we propose a set of directional filters motivated by the two-dimensional Gabor filters [31]. This involves computation of the prominence of local T-F activity, in the MWSG filtered spectrogram, in four different directions. Thus, we obtain four new directional T-F representations, where each captures the degree of T-F activity in a specific direction. A thresholding scheme is then applied to each of these directional T-F representations to obtain four segmentation decisions. Since each of the four T-F representations contains information specific to a direction in T-F plane, the four decisions are combined to provide the final segmentation boundaries. We find that, across multiple datasets, the proposed algorithm performs better than six baseline schemes, on average. This exhibits the robustness of the proposed method to the variations in recording conditions.

The paper is organized as follows. In Section II, the spectrogram enhancement using MWSG is explained followed by the proposed bird sound detection algorithm. Next, the datasets used in the study and the dataset specific challenges involved in obtaining manual annotation of the bird sounds are discussed in Section III. Section IV details the baseline schemes, the experiments performed and their results. Conclusions are drawn in Section V.

## II. Proposed Bird Sound Detection

The two stages in the proposed algorithm, i.e., the spectrogram enhancement using MWSG filter and the directional filter based bird sound detection scheme (as shown in Fig. 1), are described in detail, in this section. We begin with a brief theoretical overview of the two dimensional Savitzky-Golay (SG) filter and present its multiple window formulation. The numerical values used in the practical implementation of the proposed method are provided in Section IV-C.

### A. Spectrogram Enhancement Using MWSG Filter

*1) Signal Model:* The bioacoustic recording of duration $T$ seconds, is denoted by $y(t)$, $0 \leq t \leq T$. Let the signal be sampled at a sampling frequency of $F_s$ and be denoted as $y[n] =$

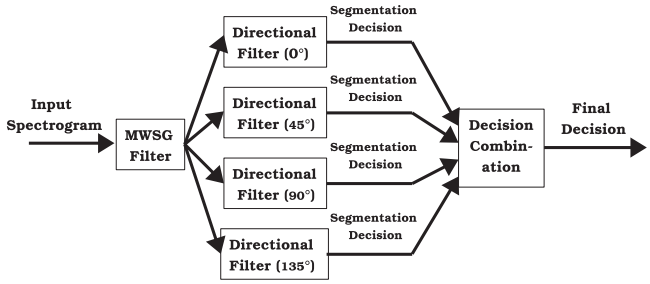



Fig. 1. Block diagram of the proposed bird sound detection scheme.

$y(n/F_s)$, $0 \le n \le N$. Consider a window $w[n]$ of length $N_w$ samples. Since it is known that most bird sounds occur at frequencies higher than 1 kHz [14], we first perform a high pass filtering of the signal at 1 kHz to get $y_f[n]$. The short-time discrete Fourier transform (STDFT) of the signal $y_f[n]$ with an $N_f$ point discrete Fourier transform is denoted by $x_{\mathrm{STDFT}}[n,k]$. We obtain the spectrogram, i.e., the magnitude of the $x_{\mathrm{STDFT}}[n,k]$ and denote it as $x[n,k]$. We now consider the following model for any $(2M+1)\times(2M+1)$ patch of $x[n,k]$.

$$x[n,k] = f[n,k] + z[n,k], -M \le n \le M, -M \le k \le M, \tag{1}$$

where, $z$ is a Gaussian random variable with zero mean and $\sigma^2$ variance, $z \sim \mathcal{N}(0,\sigma^2)$ and $f$ is a two dimensional polynomial of order $p$. We fit the polynomial $f[n,k]$ in a least squares sense to the data $x[n,k]$, giving rise to the two dimensional Savitzky-Golay filter formulation.

*2) Two dimensional Savitzky-Golay filter:* It is known that the least square polynomial smoothing, over a symmetric window, can be viewed as a filtering operation [32]. We now provide a brief derivation of this proposition for a two dimensional polynomial smoothing scenario.

The two dimensional polynomial function $f$, of degree $p$, over the symmetric window of length $2M+1$ is defined as

$$f[n,k] = \sum_{i=0}^{p} \sum_{j \le p-i} a_{ij} n^i k^j. \tag{2}$$

$a_{ij}$ are the coefficients of the polynomial which are to be estimated by a least squares polynomial fit to the data. The number of coefficients in a two dimensional polynomial of order $p$ is denoted by $P = \frac{(p+1)(p+2)}{2}$. Let $\mathbf{a}$ be a $P \times 1$ dimensional vector containing $P$ coefficients $a_{ij}, 0 \le i,j \le p$ and $i+j \le p$, i.e., $\mathbf{a} = [a_{00}, a_{01}, \ldots, a_{p0}]^T$.

The cost function, $\mathcal{C}$, to be minimized to obtain the coefficient vector $\mathbf{a}$ is given by

$$\mathcal{C} = \sum_{n=-M}^{M} \sum_{k=-M}^{M} (x[n,k] - f[n,k])^2. \tag{3}$$

Substituting (2), we obtain,

$$\mathcal{C} = \sum_{n=-M}^{M} \sum_{k=-M}^{M} \left( x[n,k] - \sum_{i=0}^{p} \sum_{j \le p-i} a_{ij} n^i k^j \right)^2. \tag{4}$$

To obtain the coefficient $\mathbf{a}$ we perform partial differentiation of (4), with respect to $a_{rs}$, $\forall r,s, 0 \le r,s \le p$, and $r+s \le p$ and equate it to zero.

$$\begin{aligned} \frac{\partial \mathcal{C}}{\partial a_{rs}} &= \sum_{n=-M}^{M} \sum_{k=-M}^{M} \left( 2 \left( x[n,k] - \sum_{i=0}^{p} \sum_{j \le p-i} a_{ij} n^i k^j \right) \right. \\ &\quad \left. \times \left(-n^r k^s\right) \right) = 0 \\ &\Rightarrow \sum_{n=-M}^{M} \sum_{k=-M}^{M} \sum_{i=0}^{p} \sum_{j \le p-i} a_{ij} n^{i+r} k^{j+s} \\ &= \sum_{n=-M}^{M} \sum_{k=-M}^{M} x[n,k](n^r k^s) \end{aligned} \tag{5}$$

Thus, there are $P$ linear equations which can be written in a matrix-vector notation as,

$$\mathbf{A}^T \mathbf{A}\mathbf{a} = \mathbf{A}^T \mathbf{x},$$

where $A$ is a $(2M+1)^2 \times P$ dimensional matrix. Given that $r$ can vary from $0, \ldots, p$ and $s = p - r$, one row of $A$, corresponding to a specific value of $n$ and $k$, is a $P$ dimensional vector with elements of the form $n^r k^s$. With the given range of $r$ and $s$, we can construct a corresponding vector $\mathbf{a}$ of dimension $P \times 1$, that contains the coefficients $a_{rs}$. $\mathbf{x} = [x_0, x_2, \ldots, x_{(2M+1)^2-1}]^T$ is a $(2M+1)^2 \times 1$ dimensional vector obtained by appending each row of the $(2M+1)\times(2M+1)$ patch $x[n,k]$, where $i$-th element $x_i = x[\xi - M, \nu - M], i = 0, \ldots, (2M+1)^2 - 1$, where $\xi = \lfloor \frac{i}{(2M+1)} \rfloor$ and $\nu = i \bmod (2M+1)$. Here, $\lfloor c \rfloor$ denotes the highest integer less than $c$ and $\bmod$ denotes modulus operator.

We can now obtain the estimate of $\mathbf{a}$, denoted by $\hat{\mathbf{a}}$ using the following equation

$$\begin{aligned} \hat{\mathbf{a}} &= (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{x}. \\ &= \mathbf{H}\mathbf{x}. \qquad (6) \\ &= [\mathbf{h_1}^T \mathbf{h_2}^T \ldots \mathbf{h_P}^T]^T \mathbf{x}, \qquad (7) \end{aligned}$$

where, $\mathbf{h_i} = [h_i[0], h_i[1], \ldots, h_i[(2M+1)^2 - 1]]$ denotes the $i$th row of $\mathbf{H}$. The output obtained after smoothing using the SG filter is the value of the polynomial $f$ at its central point of the $(2M+1)\times(2M+1)$ patch, i.e., $f[0,0]$ [33]. This corresponds to the 0th coefficient, $a_{00}$ (2) whose estimate $(\hat{a}_{00})$ for a given patch is obtained as the first element of $\hat{\mathbf{a}}$ (7), i.e., $\hat{a}_{00} = \mathbf{h_1}\mathbf{x}$. The smoothed output at the center point of a given patch is thus given by $\hat{x}[0,0] = \hat{a}_{00}$.

For a given $M$ and $p$, we can find the bias and variance of the least square estimate of the polynomial coefficient vector $\hat{\mathbf{a}}$ [34]. We know that $\hat{\mathbf{a}}$ is an unbiased estimate with $\mathcal{E}(\hat{\mathbf{a}}) = \mathbf{a}$, where $\mathcal{E}$ is the expectation operator. The co-variance of $\hat{\mathbf{a}}$ is $\Sigma = [\Sigma_{i,j}] = \sigma^2(\mathbf{A}^T\mathbf{A})^{-1}$. The mean $\mu_{\hat{x}_{00}}$ and the variance $\sigma^2_{\hat{x}_{00}}$ of $\hat{x}[0,0]$ are $a_{00}$ and $\Sigma_{1,1}$, respectively. Since $\mathbf{h_1}$ does not depend on the chosen patch elements, we can construct a fixed 2-dimensional filter with impulse response

$$\begin{aligned} h[n,k] &= h_1[(n+M)*(2M+1) + (k+M) + 1], \\ &- M \le n \le M, -M \le k \le M, \end{aligned} \tag{8}$$

and across all patches in the spectrogram we can obtain the smoothed spectrogram as

$$\hat{x}[n,k] = \sum_{r_1=-M}^{M} \sum_{r_2=-M}^{M} h[r_1,r_2]x[n-r_1,k-r_2]. \quad (9)$$

We now extend the above, to a multiple window SG filter, where multiple spectrograms computed using multiple windows are considered.

*3) Multiple window Savitzky-Golay filter:* Let there be $L$ estimates of the spectrogram obtained by choosing windows $w_l[n]$, $1 \leq l \leq L$ of $L$ different lengths, namely, $N_w^1, \ldots, N_w^L$ respectively[1]. The $l$th estimate of the magnitude of the STDFT is given by $x_l[n,k]$, $l = 1, \ldots, L$. Extending the model for a patch of size $(2M+1) \times (2M+1)$ from $L$ spectrograms we obtain,

$$x_l[n,k] = f[n,k] + z_l[n,k]; \; -M \leq n \leq M, \\ -M \leq k \leq M, l = 1, \ldots, L. \quad (10)$$

The cost function for this multiple window signal model is given by

$$\mathcal{C}_{\mathrm{MW}} = \sum_{l=1}^{L} \left( \sum_{n=-M}^{M} \sum_{k=-M}^{M} (x_l[n,k] - f[n,k])^2 \right). \quad (11)$$

Proceeding in a manner similar to Section II-A2, in place of (5), we obtain,

$$\sum_{l=1}^{L} \sum_{n=-M}^{M} \sum_{k=-M}^{M} \sum_{i=0}^{p} \sum_{j \leq p-i} a_{ij} n^{i+r} k^{j+s} \\ = \sum_{l=1}^{L} \sum_{n=-M}^{M} \sum_{k=-M}^{M} x_l[n,k](n^r k^s). \quad (12)$$

Let $\mathbf{x}_l$ be a column vector obtained from the patch $x_l[n,k]$, in a manner similar to obtaining $\mathbf{x}$ from $x[n,k]$. (12) in matrix-vector notation can be written as,

$$L\mathbf{A}^T\mathbf{A}\mathbf{a} = \mathbf{A}^T \sum_{l=1}^{L} \mathbf{x}_l.$$

Let $\mathbf{X} = \frac{1}{L}\sum_{l=1}^{L} \mathbf{x}_l$. This represents the average of $L$ different $(2M+1) \times (2M+1)$ patches from spectrograms computed using $L$ different windows. We can now obtain the multiple window estimate of $\mathbf{a}$, denoted by $\hat{\mathbf{a}}_{\mathrm{MW}}$, from the following equation

$$\hat{\mathbf{a}}_{\mathrm{MW}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{X}. \\ = \mathbf{H}\mathbf{X}. \quad (13)$$

We can obtain the $\hat{a}_{MW\,00}$, and, hence, $\hat{x}[n,k]$ following the steps outlined in Section II-A2. Thus, we find that the MWSG filtering can be interpreted as performing a two dimensional Savitzky-Golay filtering on a spectrogram obtained by averaging spectrograms computed using multiple windows. This estimate is also an unbiased one with $\mathcal{E}(\hat{\mathbf{a}}_{\mathrm{MW}}) = \mathbf{a}$. The variance of this estimate can easily be shown as $\frac{1}{L}\sigma^2(\mathbf{A}^T\mathbf{A})^{-1}$. Thus, we see that by MWSG filtering, we obtain an estimate of the spectrogram $\hat{x}[n,k]$ which has lower variance compared to that by a single window ($L = 1$) SG filtering. Fig. 2 shows the spectrogram, and its enhanced version using the single window SG filter and MWSG filter for a sample bird sound. From the figure, we can see that while regions of very high energy are enhanced by (single window) SG filtering, the MWSG filter enhances regions in the T-F plane containing bird acoustic activity even with relatively lower intensity. We also observe that the contrast between the enhanced regions and the background is more in MWSG filtered spectrogram, compared to the single window counterpart, indicating the improvement in enhancement due to the use of multiple windows.

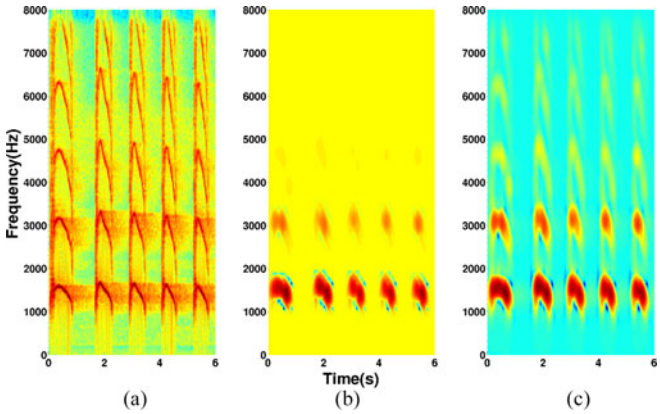



Fig. 2. (a) Spectrogram, (b) Single window SG filtered spectrogram and (c) MWSG filtered spectrogram of an exemplar field recording.

### B. Directional filtering based bird sound detection

Bird calls demonstrate a wide variety of textures in the spectrogram [35] comprising different time frequency events in different directions. These events could span a varying range of time and frequency widths. In order to capture these time frequency events in different directions, we propose to use directional filters. Gabor filters [31] are well-known for capturing and representing such events in different directions. Unlike Gabor filters, we define moving average filter of length $(2R+1)$ in four directions, namely $0^\circ$ (along time axis), $45^\circ$, $90^\circ$ (along frequency axis) and $135^\circ$. The impulse responses of these filters, $g_1[n,k]$, $g_2[n,k]$, $g_3[n,k]$, $g_4[n,k]$, are shown in Fig. 3 with $R = 5$. As observed from the figure, coefficients only in a specific direction capture direction-specific information, when convolved with the MWSG filtered spectrogram $\hat{x}[n,k]$. The directional spectrogram $\hat{x}_{\mathrm{Dq}}[n,k]$, $q = 1,2,3,4$ are obtain by convolving $\hat{x}[n,k]$ with $g_q[n,k]$ as follows:

$$\hat{x}_{\mathrm{Dq}}[n,k] = \sum_{r_1=-R}^{R} \sum_{r_2=-R}^{R} g_q[r_1,r_2]\hat{x}[n-r_1,k-r_2]. \quad (14)$$

[1]The windows, while have different length, are of identical type. We observe that varying window type does not change the results further.

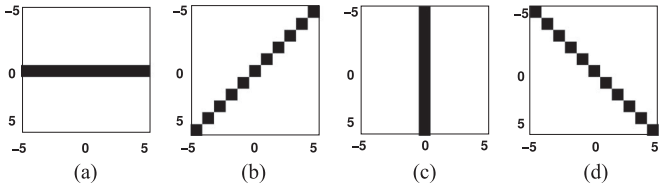



Fig. 3. Impulse responses of directional filters, (a) $g_1[n,k]$, (b) $g_2[n,k]$, (c) $g_3[n,k]$ and (d) $g_4[n,k]$, corresponding to $0^\circ$, $45^\circ$, $90^\circ$ and $135^\circ$, respectively.

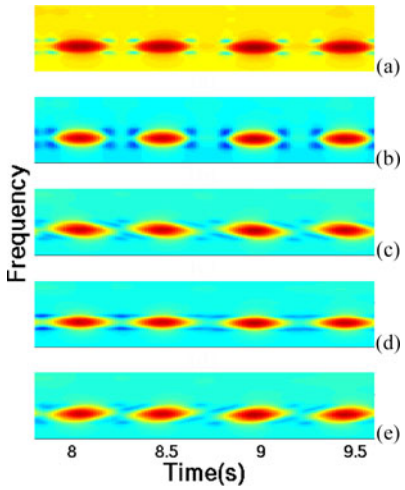



Fig. 4. (a) MWSG filtered spectrogram $\hat{x}[n,k]$; Directional spectrograms (b) $\hat{x}_{D1}[n,k]$ (c) $\hat{x}_{D2}[n,k]$, (d) $\hat{x}_{D3}[n,k]$, (e) $\hat{x}_{D4}[n,k]$ of an exemplar field recording.

Fig. 4 shows the MWSG filtered spectrogram and the directional spectrograms, $\hat{x}_{\mathrm{Dq}}[n,k]$ obtained by convolving $g_q[n,k]$, $q = 1,2,3,4$ for a sample bird sound recording. From the figure, it can be observed that directional spectrograms are effective in capturing direction specific information from the enhanced spectrogram.

From (9) and (14), it is important to note that both MWSG and directional filtering are linear time invariant (LTI) operations on the spectrogram $x[n,k]$. Hence, for implementation they could be combined to obtain a two-dimensional filter $\gamma_q[n,k] = h[n,k] * g_q[n,k]$, where $*$ denotes the convolution operation. Thus, a directional spectrogram $\hat{X}_{\mathrm{Dq}}[n,k]$ can be directly obtained from the spectrogram $x[n,k]$ by convolving with $\gamma_q[n,k]$.

Given a directional spectrogram of an audio recording, we perform the following steps.

1) Compute the frame wise energy, $e_q[n]$, of the spectrogram $x_{\mathrm{Dq}}[n,k]$, $q = 1,2,3,4$.

$$e_q[n] = \sum_{k=0}^{\frac{N_f}{2}} x_{\mathrm{Dq}}[n,k]^2. \tag{15}$$

2) Compute the histogram of $e_q[n]$ using 10 bins [23]. Let the bin centers be denoted by $c_i^q$, $1 \le i \le 10$. Let the bin corresponding to the maximum count be $c_{i*}^q$. We hypothesize that the bin $c_{i*}^q$ corresponds to the noise floor.

3) Select bin $c_{i*+1}^q$ to be the threshold, i.e, consider all frames that have an energy greater than or equal to the threshold to contain bird sound.

Thus, for each of the four directional spectrograms we obtain binary decisions $d_q[n]$ to detect the frames with bird sounds as follows:

$$d_q[n] = \begin{cases} 1 & \text{when } e_q[n] \ge c_{i*+1}^q \\ 0 & \text{Otherwise} \end{cases}. \tag{16}$$

The final decision, $d[n]$ is obtained as $d[n] = \max_q d_q[n]$. This indicates that any T-F activity prominent in any of the four directions is considered in the final decision.

## III. Dataset

To examine the robustness of the proposed bird sound detection scheme, we consider three different datasets. The details of the three datasets are provided in Table I. The first dataset we consider, was released for the IEEE international workshop on machine learning for signal processing competition- MLSP 2013 dataset [36]. The dataset consists of recordings from 19 different species of birds and is divided into a training and a test set of 502 and 143 audio recordings, respectively. Out of the 502 training samples, only 179 contain at least one bird call. We consider only this subset of the MLSP dataset for our experiments. Each audio recording is of a duration of 10 s with a sampling frequency of 16 kHz. It is to be noted that although the MLSP 2013 competition dealt with the task of bird species identification from noisy audio recordings, in the current work, we use the MLSP dataset for bird sound detection and not for bird species identification or classification. The second dataset we consider is 'Bird Songs of Florida' (BSF) [37]. This dataset consists of 99 recordings from 109 species of birds found in Florida. Available at a sampling frequency of 44.1 kHz with an average duration of 44.45 s we downsample the recordings to 16 kHz and consider only the first 10 s of each recording, to be consistent with the MLSP dataset. The third dataset considered for the study is 'Bird Songs of the Lower Rio Grande Valley and Southwestern Texas' (BSR) [38]. The BSR dataset consists of 99 audio recordings with an average duration of 44.37 s from 99 different bird species. Similar to the processing done to BSF, the recordings from BSR are downsampled to 16 kHz from 44.1 kHz and the first 10 s of each recording is considered, to have consistent datasets. BSF and BSR datasets are commercially available in audio CDs. Recordings of duration less than 10 s in BSF and BSR datasets are not considered for experiments. Thus we obtain 98 recordings from each of BSF and BSR corpora as indicated in Table I.

### A. Manual Annotation for Evaluation

To evaluate the performance of the proposed methodology, we require the ground truth bird sound segments. The manual annotation of the segments of bird sounds is done by an audio engineer, by listening to the audio recordings and by careful observations of the spectrograms. BSF and BSR datasets are found to be relatively less noisy compared to the MLSP 2013

TABLE I
DETAILS OF THE THREE DATASETS WITH EXPERIMENT SPECIFIC DETAILS

| Dataset Name | Place of data collection | Sampling frequency (Hz) | No. of species | No. audio files | Total duration (s) |
|---|---|---|---|---|---|
| MLSP [36] | Oregon | 16000 | 19 | 179 | 1790 |
| Bird Songs of Florida (BSF) [37] | Florida | 44100 (16000) | 109 | 98 | 980 |
| Bird Songs of the Lower Rio Grande Valley and Southwestern Texas (BSR) [38] | Lower Rio Grande Valley and Southwestern Texas | 44100 (16000) | 99 | 98 | 980 |

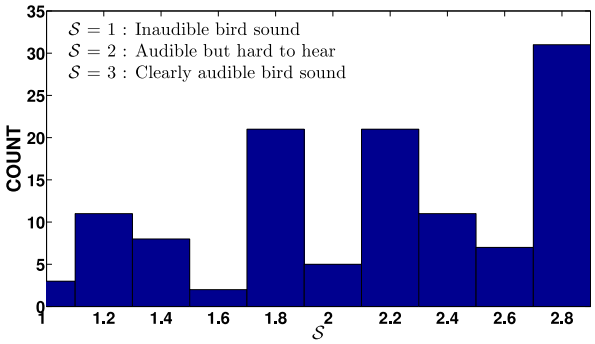



Fig. 5. Histogram of the average score of audibility for all recordings which do not have 'clearly audible bird sounds'.

dataset. Most of the audio recordings of the MLSP dataset are corrupted by different abiotic environmental noises including heavy wind, rain, waterfall and by different biotic noises including bee sounds. In BSF and BSR datasets a few recordings are corrupted by such noises.

Severely degraded audio recordings are found to be difficult to annotate even by a human listener. To quantify the degree of degradation, we compute a score of audibility for each audio recording for all three datasets, by a listening test. Each audio recording is evaluated by 7 subjects. The subjects are not reported to have any hearing defects. The audibility score is given on a three point scale of 1 to 3, with 1 being 'Inaudible bird sound' to 3 being 'Clearly audible bird sound'. Each audio recording is then provided with a new score $\mathcal{S}$, obtained by averaging the listeners' response. We find that all the audio recordings of BSF and BSR datasets secured an average score of 3, while many recordings in the MLSP dataset secured scores less than 3, indicating degradation in the recording. A histogram of $\mathcal{S}$ for values less than 3 is shown in Fig. 5. There are a total of 120 among the 179 recordings in MLSP for which at least one listener provided a score less than 3. From the figure, it is clear that among these 120 recordings, 45 recordings have an average score of less than 2. For a fair comparison of different bird sound detection algorithms, the scores $\mathcal{S}$ obtained from listeners' judgment are incorporated in the evaluation.

## IV. EXPERIMENTAL RESULTS AND DISCUSSION

### A. Baseline Schemes

We compare the performance of the proposed algorithm with six baseline schemes. The first scheme (BL1) is the one that secured the first position in the MLSP 2013 competition [27]. In this work, Fodor obtained a binary image by thresholding a Gaussian smoothened spectrogram, on which morphological operations are done to obtain T-F segments of bird activity. We consider the winning algorithm of the LifeCLEF 2014 Bird Task [29], proposed by Lasseck, as the second baseline scheme (BL2). Here, a binary image is obtained by median based thresholding, after which, morphological operations followed by median filtering is performed for noise reduction. Spatially connected segments are then grouped to obtain segments of bird sounds. The third baseline scheme (BL3) is the one proposed by Fagerlund [11]. In his work, an iterative thresholding of the energy of the bioacoustic signal is done to obtain the segments of bird sounds. A simple post-processing step of combining closely separated segments is then done to obtain the final segmentation. The fourth scheme (BL4) considered is the one proposed by Oliveira *et al.* [23]. Here, segmentation is done by thresholding the energy computed from a morphologically opened spectrogram. The fifth and the sixth baseline schemes (BL5 and BL6) are based on the state-of-the-art scattering transform technique employed for audio classification. We consider the scattering representations obtained by the time scattering framework proposed by Balestriero *et al.* [39] for BL5 and that obtained by the joint time-frequency scattering framework proposed by Andén *et al.* [40] for BL6. It is to be noted that for BL5 we use the first layer scattering representation while for BL6 we use the second order time-frequency scattering representation.[2] We perform the steps outlined in Section II-B from (15) to (16) on the scattering representations to obtain the final decision. The proposed algorithm and the baseline schemes are implemented using MATLAB R2014a software.

### B. Evaluation Metric

To evaluate the performance of the baseline schemes and the proposed methodology, we use a frame level $F$-score for each recording. $F$-score is defined as the harmonic mean of precision and recall. While precision accounts for the accuracy among those detected as positives, recall accounts for the true positive rate. For each dataset we report the mean and standard deviation of the individual $F$-scores. In addition to the traditional $F$-score, we also propose a modified $F$-score. From Section III, we observe that the audio recordings have varying degrees of degradation due to different types of noises. To ensure that

[2]We use the MATLAB toolboxes *scatnet-0.2* and *scattering.m* available from http://www.di.ens.fr/data/scattering/

TABLE II
AVERAGED $F$-SCORE, WITH STANDARD DEVIATION IN BRACKETS, ACROSS MULTIPLE DATASETS WITH $M = 21$ AND $p = 3$

| Dataset | $\mathcal{S}$ | BL1 | BL2 | BL3 | BL4 | BL5 | BL6 | Proposed method |
|---|---|---|---|---|---|---|---|---|
| MLSP: group1 | $\leq 1.667$ | 0.135 | 0.039 | 0.101 | 0.036 | 0.112 | 0.143 | **0.369** |
| | | (0.114) | (0.186) | (0.089) | (0.100) | (0.085) | (0.123) | (0.243) |
| MLSP: group2 | $1.667 > \mathcal{S} < 2.266$ | 0.279 | 0.296 | 0.170 | 0.161 | 0.227 | 0.279 | **0.546** |
| | | (0.198) | (0.328) | (0.113) | (0.192) | (0.1608) | (0.195) | (0.269) |
| MLSP: group3 | $\geq 2.333$ | 0.557 | 0.514 | 0.306 | 0.328 | 0.448 | 0.542 | **0.675** |
| | | (0.255) | (0.309) | (0.135) | (0.238) | (0.177) | (0.204) | (0.189) |
| MLSP: overall | | 0.427 | 0.393 | 0.243 | 0.245 | 0.345 | 0.420 | **0.600** |
| | | (0.279) | (0.344) | (0.148) | (0.238) | (0.209) | (0.248) | (0.244) |
| MLSP: overall | modified $F$-score | 0.472 | 0.443 | 0.265 | 0.277 | 0.383 | 0.465 | **0.629** |
| | | (0.277) | (0.335) | (0.146) | (0.242) | (0.202) | (0.238) | (0.229) |
| BSF | $= 3$ | **0.749** | 0.736 | 0.526 | 0.518 | 0.691 | 0.681 | 0.721 |
| | | (0.199) | (0.158) | (0.226) | (0.245) | (0.193) | (0.198) | (0.189) |
| BSR | $= 3$ | 0.750 | **0.805** | 0.505 | 0.454 | 0.579 | 0.574 | 0.695 |
| | | (0.142) | (0.107) | (0.188) | (0.215) | (0.184) | (0.192) | (0.171) |
| All data | | 0.596 | 0.590 | 0.385 | 0.371 | 0.496 | 0.528 | **0.657** |
| | | (0.281) | (0.319) | (0.227) | (0.264) | (0.249) | (0.248) | (0.219) |
| All data | modified $F$-score | 0.632 | 0.632 | 0.409 | 0.397 | 0.526 | 0.558 | **0.675** |
| | | (0.261) | (0.291) | (0.222) | (0.258) | (0.235) | (0.232) | (0.206) |

the algorithms are not penalized for their poor performance on a highly degraded audio recording, we define the modified $F$-score as the weighted average of the individual $F$-scores with the corresponding normalized listeners' average score of audibility (Section III-A) as the weights.

### C. Performance of Proposed Method With Fixed M and P

Similar to the work by Wang *et al.*, in the proposed method, we consider $L = 5$ estimates of the spectrogram, using Hamming window, with $N_w^1 = 32$, $N_w^2 = 64$, $N_w^3 = 128$, $N_w^4 = 256$ and $N_w^5 = 512$ samples. We consider $N_f = N_w^L$. Hence, the $L$ estimates of the spectrogram have the same number of frequency bins, 512, in this case. The spectrograms are computed with a frame shift of 256 samples. Thus, the $L$ multiple window estimates of the spectrogram have the same dimensions in the T-F plane.

It is known that the typical duration of bird sounds range from a few to a few hundred milliseconds [41]. In this work, we assume that, on average, the variation in bird sounds could be from 150 ms to 200 ms, as also considered by Neal *et al.* [17]. Therefore, we use $R = 5$ corresponding to approximately 175 ms to compute directional spectrograms, in our experiments. As described in Section II-A3, to perform MWSG filtering, we require the patch size $(2M + 1) \times (2M + 1)$ and the polynomial degree $p$. Since we require a symmetric patch, we consider a larger patch size of 21 with $M = 10$, corresponding to 336 ms, about twice that of $R$ to ensure the event in the T-F plane due to bird sound is considered completely along with its neighboring spectro-temporal profile. As it is known that in signal interpolation, cubic polynomials fit the data with the least curvature [42], we choose $p = 3$. It is to be noted that we perform zero padding and symmetric padding (mirror-reflection) of the spectrogram during the MWSG filtering and the directional filtering, respectively, to avoid the boundary effects.

To examine the performance of the proposed algorithm for different audio recording qualities, we divide the MLSP dataset into three groups, namely group1, group2 and group3 consisting of audio recordings with $\mathcal{S} \leq 1.667$, $1.667 > \mathcal{S} < 2.266$ and $\mathcal{S} \geq 2.333$, respectively. Table II provides the F-scores averaged over all recordings in a dataset along with the standard deviation (SD) obtained by the proposed algorithm and the baseline schemes for different datasets including three groups for MLSP dataset. We see that over the entire MLSP dataset as well as its individual groups, the proposed method significantly ($p$-value $< 0.001$ from double sided $t$-test) outperforms the six baseline schemes. From the table it is also observed that, for MLSP dataset, the modified $F$-score increases for all schemes. Even with modified $F$-score, we find that the proposed method, performs significantly ($p$-value $< 0.001$) better than the baseline schemes. Interestingly, we observe that the joint T-F scattering representation (BL6) outperforms the time-scattering representation (BL5) across all the three recording conditions of the noisy MLSP dataset. For the BSF dataset we see that BL1 and our proposed method have comparable ($p$-value $= 0.313$) performance, while for the BSR dataset BL2 significantly ($p$-value $< 0.001$) outperforms all other schemes. Since, we want to examine the robustness of the proposed method across different recording devices and environmental noises during recording, we compute an average $F$-score for all the recordings from all three datasets. As observed from Table II, the best baseline scheme turns out to be BL1 and our proposed algorithm outperforms the best baseline by $10.24\%$ (relative). To examine if the improvement in $F$-score over all recordings from all three datasets from the proposed approach is statistically significant, we perform a statistical pairwise $t$-test for equality of means. We find that the improvement in $F$-score of the proposed method is statistically significant over the six baseline schemes with $p$-value 0.000963, 0.000973, $3.38e-53$, $1.82e-50$, $8.55e-20$ and $1.60e-13$. Even in case of modified $F$-score, across all the recordings (last row in Table II) we find that our proposed method performs better than the baseline schemes.

We find an interesting congruence between the observation made by Dong *et al.* [13] and the proposed multiple window

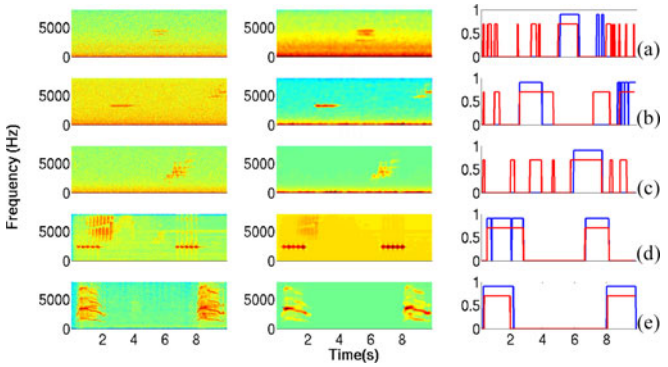



Fig. 6. Illustration of the good performance by proposed method with the recording name, $F$-score and dataset. First, second and third columns show the spectrogram, MWSG filtered spectrogram and the detected bird sounds (in red) and ground truth (in blue), respectively. (a) 'PC15_20090513_070000_0020.wav', 0.850, MLSP group1; (b) 'PC13_20100513_043015_0740.wav', 0.916, MLSP group2; (c) 'PC5_20090705_070000_0030.wav', 0.969, MLSP group3; (d) 'Track 63.wav', 0.985, BSF; (e) 'Track 4.wav', 0.951, BSR.

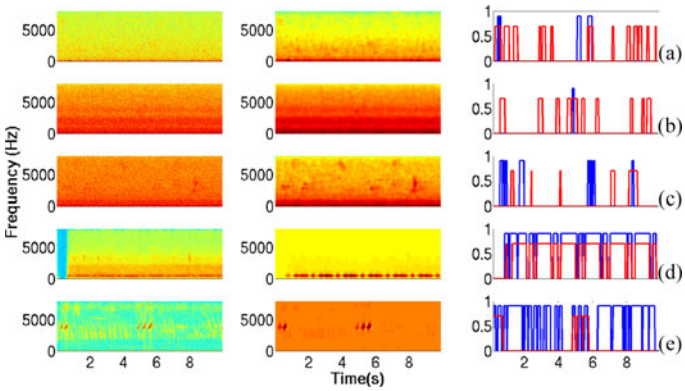



Fig. 7. Illustration of poor performance by proposed method with the recording name, $F$-score and dataset. First, second and third columns show the spectrogram, MWSG filtered spectrogram and the detected bird sounds (in red) and ground truth (in blue), respectively. (a) 'PC15_20090606_070011_0010.wav', 0.083, MLSP group1; (b) 'PC4_20100606_050000_0040.wav', 0.000, MLSP group2; (c) 'PC4_20090705_050000_0040.wav', 0.141, MLSP group3; (d) 'Track 32.wav', 0.113, BSF; (e) 'Track 57.wav', 0.206, BSR.

spectrogram estimation. The authors find that bird calls that are localized in time but characterized by a wideband spectrum (for example, the Eastern Whipbird) could be better detected from a time-compressed spectrogram, compared to the original spectrogram [13]. In the proposed method, among the five multiple window spectrogram estimates, the estimates corresponding to $l = 1, 2$ and 3, each with a frame shift of 256 samples, and window sizes of 32, 64 and 128 samples respectively, could also be viewed as time-compressed spectrograms. The improvement in the performance of the proposed method could be because we combine such estimates of the spectrogram, that are known to be better T-F representations for bird calls that are otherwise difficult to detect.

We illustrate the performance of the proposed method in Figs. 6 and 7 using one audio file from each of the three groups of MLSP dataset and one audio file from each of BSF and BSR datasets. Fig. 6 shows the spectrogram, the MWSG filtered spectrogram and the detected bird sounds (in red) along with the ground truth (in blue) in the audio recordings where the proposed algorithm performs well. Fig. 7 depicts those where the proposed approach performs poorly. Interestingly, we observe that the proposed algorithm performs well when the bird sound has a more prominent T-F structure compared to the background, as in the case of Fig. 6(d) and (e). Although there are a few false alarms in Fig. 6(a)–(c), we see that the bird sounds are detected well. We observe that the performance of the proposed methodology is poor when the bird sounds are less prominent, as in cases of Fig. 7(c) and (e). The proposed method also fails when the noise and the bird sounds are equally prominent in the T-F plane, as in the cases of Fig. 7(a), (b), and (d).

TABLE III
AVERAGE $F$-SCORE FOR EACH DATASET USING THE $M^*$ AND $p^*$ PAIR SEPARATELY FROM ALL DATASETS

| Database | $M^*$ | $p^*$ | MLSP | BSF | BSR |
|---|---|---|---|---|---|
| MLSP | 11 | 4 | 0.612 | 0.726 | 0.688 |
| | | | (0.242) | (0.185) | (0.183) |
| BSF | 15 | 4 | 0.607 | 0.734 | 0.691) |
| | | | (0.249) | (0.180) | (0.182) |
| BSR | 21 | 3 | 0.600 | 0.721 | 0.695 |
| | | | (0.244) | (0.189) | (0.171) |

### D. Performance of the Proposed Method With Varying M and P

We examine the performance of the proposed method with varying patch size and polynomial order by considering $M = 2, 5, 7, 10, 12,$ and 15 and $p$ varying from 3 to $2M$ in steps of 1. A graphical representation of the average $F$-scores for every $M$ and $p$ combination for each of the three datasets is shown in Fig. 8. From the figure, we observe that for any choice of $M$ and $p$, the highest average $F$-score happens for BSF dataset, followed by that for BSR which is followed by that for MLSP dataset. The optimal $M$ and $p$ combination, corresponding to the highest average F-score separately for each dataset, denoted by $M^*$ and $p^*$, is indicated in the figure along with the best $F$-score. It can be observed that within each dataset, the $F$-scores do not vary drastically for different choices of $M$ and $p$. Specifically, we see that the minimum and maximum $F$-score values are 0.579 and 0.612 for MLSP, 0.696 and 0.734 for BSF and 0.672 and 0.695 for BSR, respectively. Table III summarizes the performance of the proposed algorithm when dataset specific $M^*$ and $p^*$ are applied to other datasets. We notice that the performance is not affected significantly when non-optimal $M$ and $p$ combinations are employed. This indicates that the proposed methodology is less sensitive to the parameters $M$ and $p$ and hence, robust across different recording and noise conditions.

### E. Effect of MWSG Filtering and Directional Filtering on Spectrogram

To understand the contribution of the two main stages of the proposed algorithm, we compare the performance of four different schemes for each of the three datasets. In the first

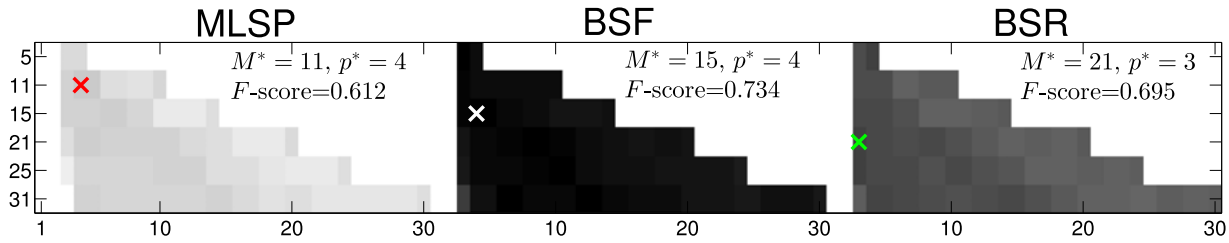



Fig. 8. Graphical representation $F$-scores obtained for different combinations of $M$ (y axis) and $p$ (x axis). The optimal $M^*$ and $p^*$ are indicated in red, white and green '$\times$' symbol for MLSP, BSF and BSR datasets respectively with the corresponding $F$-scores.

TABLE IV
EFFECT OF THE DIFFERENT STAGES IN THE PROPOSED ALGORITHM

| Database | Scheme1 | Scheme2 | Scheme3 | Proposed methodology |
|---|---|---|---|---|
| MLSP | 0.480 | 0.565 | 0.556 | 0.600 |
| | (0.245) | (0.275) | (0.253) | (0.244) |
| BSF | 0.436 | 0.718 | 0.625 | 0.721 |
| | (0.205) | (0.199) | (0.194) | (0.189) |
| BSR | 0.325 | 0.686 | 0.543 | 0.695 |
| | (0.159) | (0.179) | (0.189) | (0.171) |

scheme (Scheme1) we use the spectrogram directly without any enhancement to detect bird sounds. We use the MWSG filtered spectrogram without directional filtering, in the second scheme (Scheme2) to understand the benefit of MWSG filtering based enhancement only. In the third scheme (Scheme3) we use the directional filtering on the original spectrogram to evaluate the contribution of the directional filtering only. The final scheme is the proposed methodology which incorporates, both, the MWSG as well as directional filtering for bird sound detection. Table IV provides the average $F$-scores for each of these schemes for the three datasets. From the table, we observe that the $F$-scores increase from Scheme1 to Scheme2 and Scheme1 to Scheme3, indicating the improvements obtained, individually, by every stage of the proposed algorithm. We find that incorporating both the stages, we obtain the highest average $F$-score indicating merits of both stages of the proposed approach.

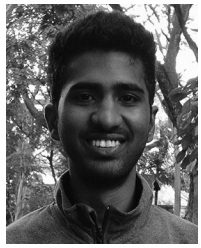
## V. CONCLUSION

In this work, we propose a robust unsupervised bird sound detection scheme. To alleviate the different challenges of recording and noise conditions, we propose a novel spectrogram enhancement scheme based on MWSG filtering of the original spectrogram. We then capture the T-F prominence of bird sounds in specific directions by computing directional spectrograms. We show that the MWSG filtering yields a low variance estimate of the spectrogram. We find that, both, MWSG and directional filtering improve the detection of bird sounds from a noisy field recording. Interestingly, the proposed method is observed to be less sensitive to MWSG filter parameters. The proposed algorithm is found to be effective to detect bird sounds in noisy recordings, as it outperforms the baseline schemes for the degraded MLSP dataset. On average, across all the three datasets we see that the proposed method performs significantly better than the baseline schemes. Future works include, employing the results of the proposed bird sound detection scheme to perform bird species identification from real field recordings.

## REFERENCES

[1] C. D. Thomas *et al.*, "Extinction risk from climate change," *Nature*, vol. 427, no. 6970, pp. 145–148, 2004.
[2] J. M. Hoekstra, T. M. Boucher, T. H. Ricketts, and C. Roberts, "Confronting a biome crisis: Global disparities of habitat loss and protection," *Ecology Lett.*, vol. 8, no. 1, pp. 23–29, 2005.
[3] T. M. Aide, C. Corrada-Bravo, M. Campos-Cerqueira, C. Milan, G. Vega, and R. Alvarez, "Real-time bioacoustics monitoring and automated species identification," *PeerJ*, vol. 1, 2013, Art. no. e103.
[4] M. A. Acevedo and L. J. Villanueva-Rivera, "Using automated digital recording systems as effective tools for the monitoring of birds and amphibians," *Wildlife Soc. Bull.*, vol. 34, no. 1, pp. 211–214, 2006.
[5] P. Somervuo, A. Harma, and S. Fagerlund, "Parametric representations of bird sounds for automatic species recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 2252–2263, Nov. 2006.
[6] C. H. Lee, C. C. Han, and C. C. Chuang, "Automatic classification of bird species from their sounds using two-dimensional cepstral coefficients," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1541–1550, Nov. 2008.
[7] E. Chesmore and C. Nellenbach, "Acoustic methods for the automated detection and identification of insects," in *Proc. 3rd Int. Symp. Sensors Horticulture*, 1997, pp. 223–231.
[8] T. Ganchev, I. Potamitis, and N. Fakotakis, "Acoustic monitoring of singing insects," in *Proc. 2007 IEEE Int. Conf. Acoust., Speech, Signal Process*, Apr. 2007, vol. 4, pp. IV–721–IV–724.
[9] C.-H. Lee, C.-H. Chou, C.-C. Han, and R.-Z. Huang, "Automatic recognition of animal vocalizations using averaged MFCC and linear discriminant analysis," *Pattern Recognit. Lett.*, vol. 27, no. 2, pp. 93–101, 2006.
[10] P. J. Clemins, M. T. Johnson, K. M. Leong, and A. Savage, "Automatic classification and speaker identification of african elephant (loxodonta africana) vocalizations," *J. Acoust. Soc. Amer.*, vol. 117, no. 2, pp. 527–534, 2005.
[11] S. Fagerlund, "Automatic recognition of bird species by their sounds," Ph.D. dissertation, Helsinki University of Technology, Espoo, Finland, 2004.
[12] X. Dong, M. Towsey, A. Truskinger, M. Cottman-Fields, J. Zhang, and P. Roe, "Similarity-based birdcall retrieval from environmental audio," *Ecological Inf.*, vol. 29, no. 1, pp. 66–76, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1574954115001168
[13] X. Dong, J. Xie, M. Towsey, J. Zhang, and P. Roe, "Generalised features for bird vocalisation retrieval in acoustic recordings," in *IEEE 17th Int. Workshop Multimedia Signal Process.*, Oct. 2015, pp. 1–6.
[14] A. Harma, "Automatic identification of bird species based on sinusoidal modeling of syllables," in *Proc. 2003 IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2003, vol. 5, pp. 545–548.
[15] M. Graciarena, M. Delplanche, E. Shriberg, A. Stolcke, and L. Ferrer, "Acoustic front-end optimization for bird species recognition," in *Proc. 2010 IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2010, pp. 293–296.
[16] R. Bardeli, D. Wolff, F. Kurth, M. Koch, K.-H. Tauchert, and K.-H. Frommolt, "Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring," *Pattern Recognit. Lett.*, vol. 31, no. 12, pp. 1524–1534, 2010.

[17] L. Neal, F. Briggs, R. Raich, and X. Z. Fern, "Time-frequency segmentation of bird song in noisy acoustic environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2011, pp. 2012–2015.
[18] N. C. Wang, R. E. Hudson, L. N. Tan, C. E. Taylor, A. Alwan, and K. Yao, "Bird phrase segmentation by entropy-driven change point detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2013, pp. 773–777.
[19] T. S. Brandes, "Feature vector selection and use with hidden Markov models to identify frequency-modulated bioacoustic signals amidst noise," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 6, pp. 1173–1180, Aug. 2008.
[20] I. Potamitis, S. Ntalampiras, O. Jahn, and K. Riede, "Automatic bird sound detection in long real-field recordings: Applications and tools," *Appl. Acoust.*, vol. 80, pp. 1–9, 2014.
[21] K. Kaewtip, L. N. Tan, A. Alwan, and C. E. Taylor, "A robust automatic bird phrase classifier using dynamic time-warping with prominent region identification," in *Proc. 2013 IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2013, pp. 768–772.
[22] L. N. Tan, G. Kossan, M. L. Cody, C. E. Taylor, and A. Alwan, "A sparse representation-based classifier for in-set bird phrase verification and classification with limited training data," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2013, pp. 763–767.
[23] A. G. de Oliveira *et al.*, "Bird acoustic activity detection based on morphological filtering of the spectrogram," *Appl. Acoust.*, vol. 98, pp. 34–42, 2015.
[24] Y. Huang, F. Briggs, R. Raich, K. Eftaxias, and Z. Lei, "The ninth annual MLSP data competition," in *IEEE Int. Workshop Mach. Learn. Signal Process.*, Sep. 2013, pp. 1–4.
[25] H. Goëau, H. Glotin, W.-P. Vellinga, R. Planqué, A. Rauber, and A. Joly, "LifeCLEF bird identification task 2014," in *Proc. CLEF 2014*, Sheffield, France, Sep. 2014. [Online]. Available: https://hal.inria.fr/hal-01088829
[26] H. Glotin, Y. LeCun, T. Artieres, S. Mallat, O. Tchernichovski, and X. Halkias, "Neural information processing scaled for bioacoustics, from neurons to Big Data." 2013. [Online]. Available: http://sabiod.org
[27] G. Fodor, "The ninth annual MLSP competition: First place," in *IEEE Int. Workshop Mach. Learn. Signal Process.*, Sep. 2013, pp. 1–2.
[28] I. Potamitis, "Automatic classification of a taxon-rich community recorded in the wild," *PloS one*, vol. 9, no. 5, 2014, Art. no. e96936.
[29] M. Lasseck, "Large-scale identification of birds in audio recordings." in *Proc. CLEF (Working Notes)*, 2014, pp. 643–653.
[30] P. Marler, "Bird calls: Their potential for behavioral neurobiology," *Ann. New York Acad. Sci.*, vol. 1016, no. 1, pp. 31–44, 2004.
[31] T. C. Bau, "Using two-dimensional gabor filters for handwritten digit recognition," Ph.D. dissertation, M.Sc. thesis, University of California, Irvine, CA, USA, 2008.
[32] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Anal. Chem.*, vol. 36, pp. 1627–1639, 1964.
[33] R. W. Schafer, "What is a Savitzky-Golay filter? [lecture notes]," *IEEE Signal Process. Mag.*, vol. 28, no. 4, pp. 111–117, Jul. 2011.
[34] O. Burke, "More notes for least squares," *Dep. Stat., Univ. Oxford*, Oxford, U.K., 2010.
[35] S. S. Chen and Y. Li, "Automatic recognition of bird songs using time-frequency texture," in *Proc. 5th Int. Conf. Comput. Intell. Commun. Netw.*, Sep. 2013, pp. 262–266.
[36] Y. Huang, F. Briggs, R. Raich, K. Eftaxias, and Z. Lei, "MLSP 2013 bird classification challenge," Kaggle, 2013. [Online]. Available: https://www.kaggle.com/c/mlsp-2013-birds/data
[37] G. A. Keller, "Bird songs of florida," Audio CD, Oct. 1, 1997.
[38] G. A. Keller, "Bird songs of the Lower Rio Grande Valley and southwestern Texas," Audio CD, Jan. 1, 2000.
[39] R. Balestriero and H. Glotin, "Scattering decomposition for massive signal classification: From theory to fast algorithm and implementation with validation on international bioacoustic benchmark," in *Proc. IEEE Int. Conf. Data Mining Workshop*, Nov. 2015, pp. 753–761.
[40] J. Andén, V. Lostanlen, and S. Mallat, "Joint time-frequency scattering for audio classification," in *IEEE 25th Int. Workshop Mach. Learn. Signal Process.*, Sep. 2015, pp. 1–6.
[41] A. H. Bass and C. W. Clark, "The physical acoustics of underwater sound communication," in *Acoust. Commun.*, 2003, pp. 15–64.
[42] P. Brigger, J. Hoeg, and M. Unser, "B-spline snakes: A flexible tool for parametric contour detection," *IEEE Trans. Image Process.*, vol. 9, no. 9, pp. 1484–1496, Sep. 2000.



**Nithin Rao Koluguri** was born in Warangal, India, in 1994. He recently completed the under graduation in electronics and communication engineering from the National Institute of Technology Karnataka, Surathkal, India, in May 2016. His research interests include signal processing, automation, and embedded systems.

**G. Nisha Meenakshi** (S'04–M'12–SM'15) received the B.E. degree in electrical and electronics engineering from R.M.D engineering college, Tamil Nadu, India, in 2009, the M.Tech degree in remote sensing from the Department of Civil Engineering, College of Engineering, Guindy, Anna University, Chennai, India, in 2012, and is currently working toward the Ph.D. degree from the Department of Electrical Engineering, Indian Institute of Science (IISc), Bangalore. Her interests include signal processing for speech and audio.

**Prasanta Kumar Ghosh** (S'16) received the B.E.(ETCE) in electronics from Jadavpur University, Kolkata, in 2003, M.Sc.(engineering) in electrical communication engineering from Indian Institute of Science (IISc), Bangalore, in 2006, and the Ph.D. in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 2011. He is an Assistant Professor in the department of Electrical Engineering, IISc. His research interests include nonlinear signal processing methods with applications to speech and audio, audio–visual signal processing, and biomedical signal processing.