INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Robust Detection of Multiple Bioacoustic Events with Repetitive Structures

Frank Kurth[1]

[1]Fraunhofer FKIE, Fraunhoferstr. 20, 53343 Wachtberg, Germany
frank.kurth@fkie.fraunhofer.de

Abstract

In this paper we address the task of robustly detecting multiple bioacoustic events with repetitive structures in outdoor monitoring recordings. For this, we propose to use the shift-autocorrelation (shift-ACF) that was previously successfully applied to F0 estimation in speech processing and has subsequently led to a robust technique for speech activity detection. As a first contribution, we illustrate the potentials of various shift-ACF-based time-frequency representations adapted to repeated signal components in the context of bioacoustic pattern detection. Secondly, we investigate a method for automatically detecting multiple repeated events and present an application to a concrete bioacoustic monitoring scenario. As a third contribution, we provide a systematic evaluation of the shift-ACF-based feature extraction in representing multiple overlapping repeated events.

Index Terms: bioacoustics, multiple repeated events, robust detection

1. Introduction

Methods for automatic pattern recognition have been applied for detecting animal vocalizations in audio recordings for more than a decade now. Only recently, powerful pattern recognition methods have been reported and comprehensively evaluated for large scale acoustic bird detection, segmentation and classification [1, 2, 3]. Despite this enormous progress there is still much room for improvement particularly when analyzing complex field recordings containing mixtures of simultaneous vocalizations and background signals, such as proposed in [4, 5].

In [6] it has been proposed to use the repetitiveness of bird calls as a method for robust call detection. In this paper, we follow this idea and address the task of detecting *multiple overlapping* repetitive calls. To this end we propose to use a recent technique [7] based on generalized autocorrelation functions (ACFs) that has been successfully applied to the tasks of robust F0-estimation in noisy speech [8] and the detection of multiple simultaneous speakers [9].

As a first contribution, this paper illustrates the potentials of various time-frequency representations derived from the generalized ACF in the context of bioacoustical pattern detection. We then investigate a method for automatically detecting multiple repeated events and present an application to a concrete bioacoustic monitoring scenario. Particularly we illustrate that the proposed technique, for the case of repetitive acoustic events, can be used to identify vocalizations of different individuals that are active at the same time, based on a single channel recording only. As a third contribution, we provide a systematic evaluation of the proposed ACF-features capabilities to separate multiple overlapping repeated events in realistic acoustic background environments.

In Section 2 we review the generalized ACF function that is subsequently used for feature extraction. Based on this ACF, several time-frequency transforms for representing repetitive structures are introduced in Section 3 and illustrated in the bioacoustic context. In Section 4 we describe an approach for systematically detecting repeated events in complex audio recordings. Section 5 contains (i) a case-study of applying the algorithm proposed in Section 4 and (ii) a systematic evaluation of the ACF-based features for separating events overlapping in time and frequency.

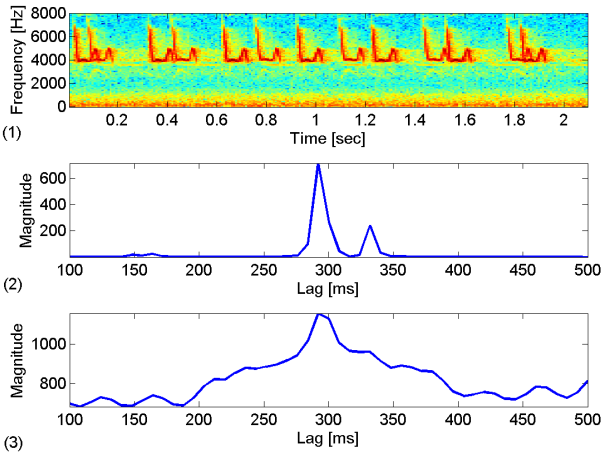

Figure 1: *(1) Spectrogram showing two overlapping call sequences of Phylloscopus collybita. (2) The type 0110 shift-ACF exhibits the IOI of both sequences resp. birds (300 ms and 330 ms). (3) The classical ACF only exibits one of the IOIs.*

2. Generalized Autocorrelation

A classical robust method for detecting repeated components in a discrete time signal x of finite energy is the (sample-based) autocorrelation (ACF) defined as

$$\mathrm{ACF}[x](s) := \sum_{k\in\mathbb{Z}} x(k)\cdot\overline{x(k-s)}. \tag{1}$$

A local maximum of $|\mathrm{ACF}[x]|$ at position s indicates repetitions in x at a distance, or *lag*, of s samples. The basic principle of the classical ACF is that signal components repeating at a lag of s samples within an analyzed signal x are emphasized by a shift-product $\mathbb{O}_s^0[x](k) := x(k)\cdot\overline{x^s(k)}$, where x is multiplied with the conjugate of its s-shifted version $x^s(k) := x(k-s)$. Fig. 1 (1) shows a spectrogram of two overlapping call sequences of the Common Chiffchaff (Phylloscopus collybita). Each sequence consists of 7 calls. The calls of the first bird have an inter onset interval (IOI) of 300 ms between subsequent calls. The calls of the second bird have IOIs of 330 ms. The classical

http://dx.doi.org/10.21437/Interspeech.2016-83

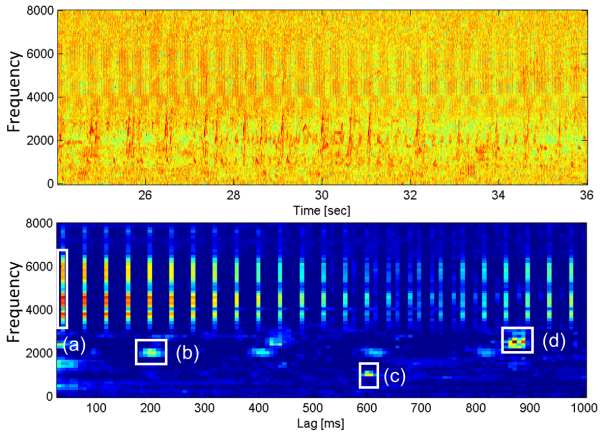

Figure 2: *Top: Spectrogram of a field recording containing a mix of repetitive and non-repetitive vocalizations. Bottom: Type 0110 subband shift-ACF with the automatically detected four most dominant repeating components marked by (a)-(d).*

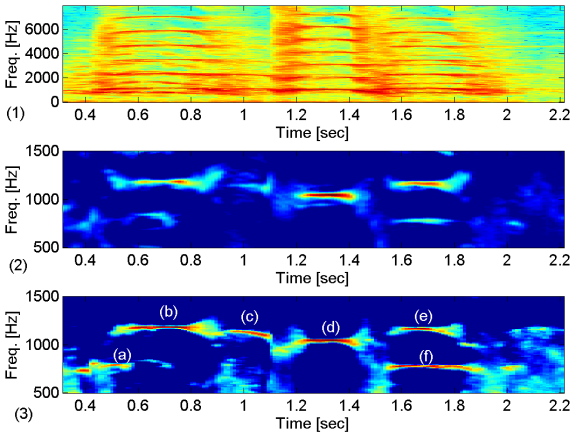

Figure 3: *(1) Spectrogram for a mix of two calls of the Common crane, (2) column-wise classical ACF, (3) type 010 shift-ACF with positions of true F0 trajectories marked as (a)-(f).*

ACF (3) has a peak region at around 300 ms, roughly indicating the true IOIs.

In [7] the shift-ACF was proposed to improve the performance of classical ACF for cases of multiple repetitions, i.e., the case that an event is repeated more than two times at the same IOI. The first principle underlying the shift-ACF is to apply the shift-product, or *type 0*, operator $\mathbb{O}_s^0$ iteratively to amplify repeating components. The second idea is to complement $\mathbb{O}_s^0$ by a, *type 1*, shift-minimum operator $\mathbb{O}_s^1[x](k) := \min(|x(k)|, |x^s(k)|)$ in order to suppress non-repeating components. This can be generalized by arbitrarily composing operators $\mathbb{O}_s^t := \mathbb{O}_s^{t_1} \circ \cdots \circ \mathbb{O}_s^{t_n}$ where $t = (t_1, \ldots, t_n) \in \{0, 1\}^n$ specifies which sequence of operator types is applied. The *shift-ACF of type* t and length n is then defined by

$$\mathrm{ACF}^t[x](s) := \sum_{k \in \mathbb{Z}} \mathbb{O}_s^t[x](k). \qquad (2)$$

The classical ACF is obtained as special case of a type 0 shift-ACF. Fig. 1 (2) shows the type 0110 shift-ACF of the mixed Chiffchaff calls. This representation reveals both of the true IOIs by rather sharp peaks at 300 and 330 ms. As described in [7], the improved representation of multiple repeating events is a mathematical property of the shift-ACF and thus an advantage over classical ACF.

3. Time-Frequency Transforms for Detecting Repetitive Structures

3.1. Subband Shift-ACF

In applications it is frequently more suitable to use a frequency-selective version of the shift-ACF that is defined on certain subbands [10]. Based on the length N discrete Fourier transform matrix $D_N := (e^{\frac{-2\pi i k \ell}{N}})_{0 \le k, \ell < N}$, a window function $w \in \mathbb{R}^N$ and an analysis step size $S \in \mathbb{N}$, an input signal x is split in time frames $x_m^w := (x_{mS} \cdot w_0, \ldots, x_{mS+N-1} \cdot w_{N-1})^\top$. Then, $\mathrm{WFT}_w[x](m) := D_N \cdot x_m^w$ is the m-th column vector of the WFT (spectrogram) of x and subband signals are obtained as row sequences $\mathrm{WFT}_w[x]_{j,:}$, $j \in [0 : N-1]$. The *subband shift-ACF* of type t is defined as the row-by-row shift-ACF:

$$\mathrm{SACF}^t[x](j, s) := \mathrm{ACF}^t[\mathrm{WFT}_w[x]_{j,:}](s). \qquad (3)$$

Fig. 2 (bottom) shows the type 0110 subband shift-ACF based on a spectrogram (top) of a field recording. In the subband shift-ACF various repeating components can be observed as energy-rich components. The four strongest components are indicated as (a)-(d). Component (d) corresponds to repeated calls of a Spotted Crake (Porzana porzana).

3.2. Spectral Shift-ACF

Harmonic sounds, which are very frequent in human and animal vocalizations can also be detected by a variant of the shift-ACF. Whereas a temporal repetition of an acoustic event is simply a time shifted version, a harmonic sound can be modeled as the sum of a fundamental frequency (F0-) trajectory $\sin(f(m))$, i.e., a frequency modulated signal and weighted harmonic components $w_k \cdot \sin(kf(m))$ for $k = 2, 3, \ldots$.

By assuming that $f(m)$ is only slowly varying over time, harmonic components may be locally modeled as repetitions in frequency. Hence, the shift-ACF of the local signal spectrum can be used for analysis as proposed in [8]. More precisely, the *spectral shift-ACF* of type t is defined by

$$\mathrm{SpACF}^t[x](s, m) := \mathrm{ACF}^t[\mathrm{WFT}_w[x]_{:,m}](s), \qquad (4)$$

i.e., by independently computing the shift-ACF for each spectrogram column. For illustration we mixed two field recordings, each containing three call components of the Common crane (Grus Grus). Each of the resulting six call components is characterized by a harmonic spectrum (F0 trajectory plus overtones). Fig. 3 in (1) shows the spectrogram, for illustration restricted to a maximum frequency of 7 kHz. In (2), the column-wise classical ACF computed on the spectrogram is shown, whereas (3) shows the type 010 spectral shift-ACF. The trajectories marked as (a)-(f) indeed correspond to the true six F0-trajectories ((a), (c), (f): first individuum; (b), (d), (e): second individuum). Although component (a) has weaker energy and a somewhat scattered appearance, the true trajectories of both birds are well represented by the shift-ACF. Our experiments show that by using the methods proposed in [8], those F0-trajectories can be reliably extracted.

We conclude this section by only mentioning that time-varying temporal repetitions can be analyzed by using a short-time shift-ACF inspired by the tempogram that is widely used

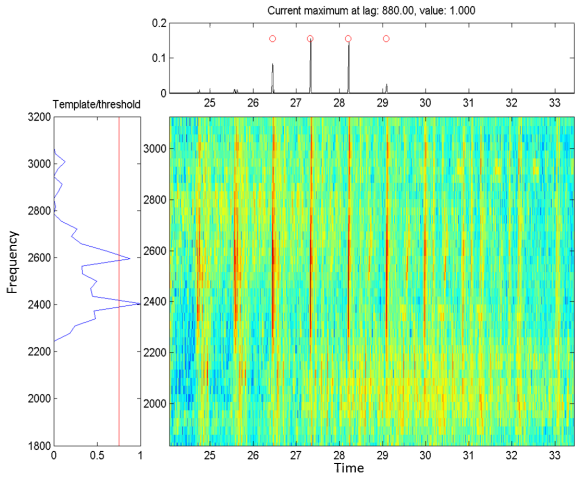

Figure 4: *Center: Shift-spectrogram for a lag of 880 ms applied to field recording shown in Fig. 2. Left: Row sums of shift-spectrogram. Top: Columns sums of shift-spectrogram regions above significance threshold.*

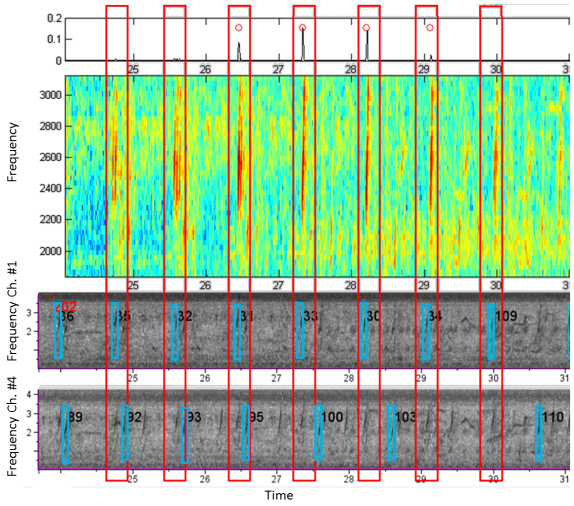

Figure 5: *Shift-spectrogram for field recording shown in Fig. 2 and shift corresponding to first detection result. The detected peaks shown on the top (red circles) essentially match the manual transcription (blue boxes) of channel #1.*

in music retrieval [11]. In [12] we successfully applied the short time shift-ACF (called *repgram*) for analyzing time-varying click-sounds produced by marine mammals.

4. Detection of Mixed Repetitive Sources

In the remainder of this paper, we investigate the application of subband Shift-ACF and spectral shift-ACF, respectively, for detecting temporally repeating and mixed harmonic components, respectively. We adapt a method that was used in [10] for detecting and extracting digital overlapping multi-tone signals. The method is based on first detecting all significant peaks in the subband shift-ACF of a target signal, as illustrated in Fig.2. In this example, the four most significant components are labeled as (a)-(d). For each of those candidate components centered at lag L and frequency F, the following steps are performed: using the identified shift L, each row j of the original spectrogram is processed by a shift-operation $\mathbb{O}_L^t[\mathrm{WFT}_w[x]_{j,:}]$ to emphasize the lag-L repetitions. The resulting *shift spectrogram* is illustrated in Fig.4 (center). Calculating row-sums (Fig.4, left), an energy profile is computed. Using a suitable threshold (red line) further analysis is restricted to a subset of dominant frequency bands around the candidate frequency F. Then, column-sums are calculated to compute a temporal profile (Fig.4, top) from which then onset positions are obtained by peak picking. Based on the detected frequency band and onset positions, furthermore 2D patterns of the detected events may be obtained. We refer to [10] for a detailed description.

5. Evaluation

5.1. Experiment: Bird Vocalizations

The algorithm proposed in Sect. 4 was applied to detect vocalizations of Spotted crakes within the field recordings that were already used as examples in the preceeding discussion. The recordings were made using a four channel microphone array in a cross setup. The presence of five spotted crakes was manually verified and all calls within all channels were annotated individually. In order to reduce false alarms due to strong high-frequency sounds (mainly crickets) it was suitable to exploit the prior knowledge on the expected frequency range of the Spotted crakes' calls to restrict the analyzed frequency range accordingly. Fig. 5 (center) shows the shift spectrogram of the first match candidate (having an $L = 800$ ms repetition lag) obtained from analyzing a 7 second segment of the recording. Red circles (top) show the automatically detected call positions. Below the shift spectrogram, two recording channels (1 and 4) are shown along with manual annotations (blue boxes) of the true call positions. Red boxed are used to compare automatic detections and true call positions. In this case, the automatic detections essentially match the true positions in channel 1. We verified that among the first four candidates output by our algorithm – for a single channel analysis – there were indeed three occurrences of (two different) Spotted crakes. This illustrates a particular advantage of our proposed approach because, while pattern matching methods such as [2] might be able to detect the single calls of a particular species present in a given recording, such methods are not able to *distinguish*, or separate, calls of different individuals.

5.2. Multiple Harmonics Detection

We present an evaluation of how good the spectral shift-ACF performs in representing multiple harmonics. This gives an indication on how good a detection algorithm based those features can be expected to be. For our analysis we generate synthetic harmonic bird calls. Each bird call consists of a temporal sequence of short frequency modulated components, each with a predefined number of harmonics. We generate those by randomly sampling start, duration and frequency modulation of F0-trajectories and adding, in this case 5, harmonics. In our experiments, three of such calls are overlapped and additionally mixed with different types of background audio at a particular signal-to-background ratio (for simplicity called SNR).

Fig. 6 (1) shows a Spectrogram of three synthetic calls added at 0 dB SNR to a field recording containing various bioacoustic components. In (4), the true (ground truth) F0-trajectories used to generate the three synthetic calls are shown.

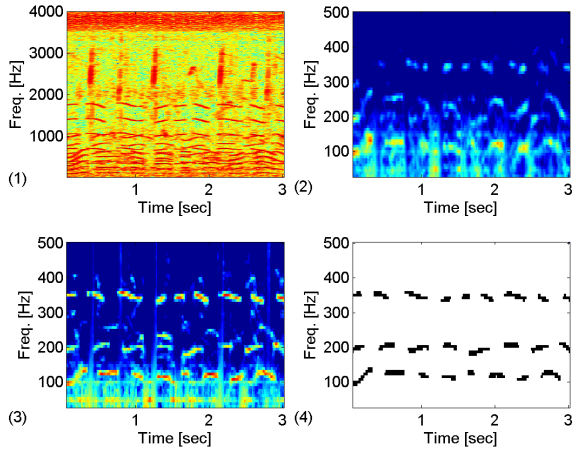

Figure 6: *(1) Spectrogram of three overlapping synthetic vocalizations added to a field recording (at 0 dB). (2) Spectral shift-ACF obtained by classical ACF. (3) Type 010 spectral shift-ACF. (4) Ground truth F0 trajectories of the three vocalizations.*

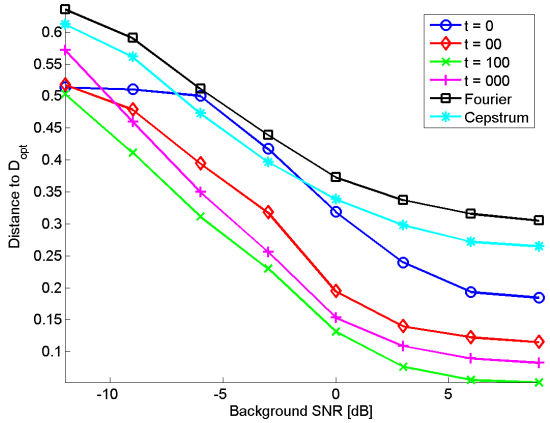

Figure 7: *Detection performance of different features for three overlapping synthetic vocalizations added to random excerpts of field recording using different SNRs.*

Here, the F0-trajectories of the synthetic calls modulate in the frequency bands around 100, 200, and 350 Hz, respectively. In the column-wise classical ACF (2) the trajectries are only hardly visible, while the type 010 spectral shift-ACF (3) much better represents the ground truth data. By thresholding the feature representations (2) and (3) we obtain a binary (detection/non-detection) version for comparison with the (binary) ground truth data, where correct detections and false positives can be computed straightforwardly. By varying the detection threshold, we compute ROC-based performance curves of true positive (i.e., correct) detections (TP) versus false positives (FP). Fig. 7 shows the detection performance of different feature types for the three overlapping synthetic vocalizations mixed with random excerpts of field recording using different SNRs. We use various types of shift-ACF (where type 0 is the classical ACF) as well as the column-wise Fourier transform and the column-wise cepstrum as baseline features. Each value on each curve is obtained from a ROC-evaluation and represents the minimum distance of the respective ROC-curve from $D_{\text{opt}} = (1, 0) =$(TP,FP), i.e., small values are better than larger

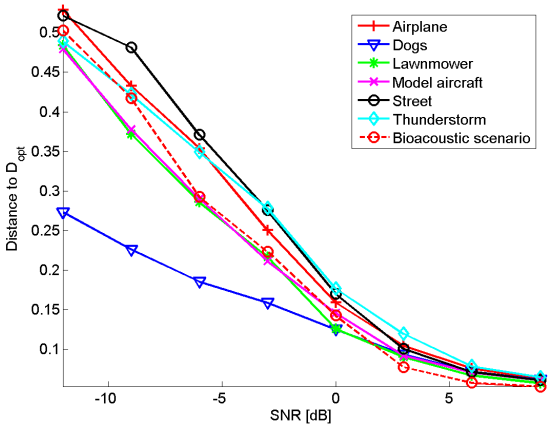

Figure 8: *Detection performance for optimum performing type 100 feature and different types of backgrounds added to three overlapping synthetic vocalizations.*

ones. We chose distance to D_{opt} as an evaluation measure as it provides a measure of the minimum joint error of false positives and missed detections. The shift-ACF features clearly outperform the classical feature types. For the evaluation shown in Fig. 8, we fixed the optimum performing shift type 100 of the previous experiment and now add seven different types of typical outdoor background sounds to the mix of synthetic calls. The "bioacoustic scenario" is reproduced from the previous experiment to serve as a reference. It can be observed that the shift-ACF features for almost all background types show qualitatively similar results as obtained for the bioacoustic scenario. Concluding we remark that although we have focused on evaluating the spectral shift-ACF, a dual analysis can be performed for the subband shift-ACF.

6. Conclusions

The shift-ACF previously used in speech processing has been shown to be useful for detecting multiple overlapping, repetitive bioacoustic events. We described time-frequency representations for analyzing (i) temporally repeating events (i.e., subband shift-ACF) and (ii) events with harmonic structure (i.e., spectral shift-ACF) and illustrated their application to bioacoustics. The discussed algorithm for multiple repeating event detection has been applied to detect simultaneous vocalizations of the same species in a single channel of a field recording. The features were systematically evaluated in a controlled scenario of harmonic mixture signals with added realistic background noise. A particular potential of the proposed approach is the separation of mixed vocalization for cases where only monophonic recordings are available. For future work it is very promising to apply the proposed methods to more field recordings and to perform an intensive evaluation to an audio corpus containing annotated repetitive events.

7. Acknowledgements

Parts of this work were initiated at Leibniz-Center Schloss Dagstuhl, in Seminar 16092. The author would like to thank Dr. Karl-Heinz Frommolt of Museum für Naturkunde Berlin for providing audio recordings and annotation from the Reference System of Animal Vocalisations, available at http://www.animalsoundarchive.org/RefSys/.

8. References

[1] C. H. Lee, C. C. Han, and C. C. Chuang, "Automatic classification of bird species from their sounds using two-dimensional cepstral coefficients," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1541–1550, Nov 2008.

[2] M. Lasseck, "Towards automatic large-scale identification of birds in audio recordings," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 6th International Conference of the CLEF Association, CLEF 2015, Toulouse, France, September 8-11, 2015, Proceedings*, 2015, pp. 364–375.

[3] T. V. Tjahja, X. Z. Fern, R. Raich, and A. T. Pham, "Supervised hierarchical segmentation for bird song recording," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, April 2015, pp. 763–767.

[4] F. Briggs, B. Lakshminarayanan, L. Neal, X. Z. Fern, R. Raich, S. J. K. Hadley, A. S. Hadley, and M. G. Betts, "Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach," *The Journal of the Acoustical Society of America*, vol. 131, no. 6, pp. 4640–4650, 2012.

[5] P. Jančovič and M. Köküer, "Acoustic recognition of multiple bird species based on penalized maximum likelihood," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1585–1589, Oct 2015.

[6] 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 Recognition Letters*, vol. 31, no. 12, pp. 1524 – 1534, 2010.

[7] F. Kurth, "The shift-ACF: Detecting multiply repeated signal components," in *Proc. IEEE WASPAA*, 2013.

[8] F. Kurth, A. Cornaggia-Urrigshardt, and S. Urrigshardt, "Robust F0 estimation in noisy speech signals using shift autocorrelation," in *Proc. IEEE ICASSP*, 2014.

[9] A. Cornaggia-Urrigshardt and F. Kurth, "Using enhanced F0-trajectories for multiple speaker detection in audio monitoring scenarios," in *Proc. EUSIPCO*, 2015.

[10] F. Kurth, "Robust Detection and Pattern Extraction of Repeated Signal Components Using Subband Shift-ACF," in *Proc. IEEE IWCCSP*, 2014.

[11] P. Grosche and M. Müller, "Extracting predominant local pulse information from music recordings," *IEEE Trans. on ASLP*, vol. 19, no. 6, pp. 1688–1701, 2011.

[12] P. M. Baggenstoss and F. Kurth, "Comparing Shift-ACF with Cepstrum for Detection of Burst Pulses in Impulsive Noise," *Journal of the Acoustical Society of America*, vol. 136, no. 4, pp. 1574–82, 2014.