ROBUST DETECTION OF JITTERED MULTIPLY REPEATING AUDIO EVENTS USING ITERATED TIME-WARPED ACF

Frank Kurth and Kevin Wilkinghoff

Fraunhofer FKIE, 53343 Wachtberg, Germany
{frank.kurth,kevin.wilkinghoff}@fkie.fraunhofer.de

ABSTRACT

This paper proposes a novel approach for robustly detecting multiply repeating audio events in monitoring recordings. We consider the practically important case that the sequence of inter onset intervals between subsequent events is not constant but differs by some jitter. In such cases classical approaches based on autocorrelation (ACF) are of limited use. To overcome this problem we propose to use ACF together with a variant of dynamic time warping. Combining both techniques in an iterative algorithm, we obtain a method for significantly improved detection of jittered multiply repeating events. In this paper we describe the new iterated time-warped ACF algorithm and evaluate its performance on the bioacoustic application of detecting repeating bird calls in monitoring recordings.

***Index Terms*—** Repeated event detection, dynamic time warping, shift-ACF

1. INTRODUCTION

Robust detection of short-time, repeating audio events is a fundamental task in acoustic monitoring. An important application is the detection of repeated animal vocalizations like for example bird calls. In [1] it was shown that a generalized autocorrelation (ACF), the shift-ACF [2], can be used to exploit the presence of *multiple* repetitions of an audio event to increase the detection performance. Here, ACF-based approaches assume repetitive temporal events with onset times $(t_0, t_0 + \lambda, t_0 + 2\lambda, \dots, t_0 + K\lambda) =: o$. In the ACF-based approach the inter-onset-interval (IOI), or lag, λ is then estimated by comparing a signal x containing the events, with shifted versions of x. In real application scenarios, the IOI λ usually is not perfectly constant but may vary by some jitter $\delta := (\delta_0, \dots, \delta_K)$, leading to distorted onset times $o + \delta$. Although ACF-based approaches are robust to a small amount of jitter, those methods naturally fail when the jitter increases. Motivated by the above application scenario of detecting repetitive bird vocalizations, this paper investigates scenarios where the event jitter is bounded by a maximum individual event deviation of $\max_i |\delta_i| < \lambda/2$. To overcome the shortcomings of ACF-based repetition detection, we propose to combine ACF-techniques with a variant of dynamic time warping (DTW) [3]. Using DTW, we align a signal x with a shifted version x^s prior to the correlation step, which allows us to compensate event jitter. By iterating this approach, we obtain the novel method of iterated time-warped ACF (ITW-ACF) that is advantageous for multiple repeating events. The use of iterated autocorrelation for improved detection of multiply repeated signal events has been initially proposed in [2]. DTW has a long history in the alignment of texts, biological sequences and time series. The variant of subsequence DTW has been successfully applied to the task of audio matching in music retrieval, see [4] for a summary of DTW techniques. The variant of restricted DTW we use in this paper has been initially proposed in [3]. The principle of *frequency* warping has been used in acoustics, particularly in order to adapt signal processing to human perception [5]. To avoid confusion, we note that the concept of warped autocorrelation used in the latter paper is different from the concepts in our paper.

The paper is organized as follows. In Section 2 we summarize the existing shift-ACF- and DTW-approaches. Based on those, we introduce the novel approach of ITW-ACF in Section 3. In Section 4 we describe the results of a comprehensive evaluation in the context of bioacoustic signal detection, an application domain that has recently attracted significant attention [6, 7, 8, 9, 10].

2. BASIC APPROACHES

2.1. ACF and Shift-ACF

The (sample-based) autocorrelation (ACF) of a discrete time signal x of finite energy is defined as

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

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 $\mathcal{O}_s^1[x](k) := x(k) \cdot \overline{x^s(k)}$, where x is multiplied by the conjugate of its s-shifted version $x^s(k) := x(k+s)$. An ACF shows repeating events with an IOI of s by a local maximum of $|\mathrm{ACF}[x]|$ at lag s.

In [2] 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 1*, operator $\mathcal{O}_s^1$ iteratively to amplify repeating components. Secondly, $\mathcal{O}_s^1$ is complemented by a, *type 0*, shift-minimum operator $\mathcal{O}_s^0[x](k) := \min(|x(k)|, |x^s(k)|)$ in order to suppress non-repeating components. This can be generalized by arbitrarily iterating n operators $\mathcal{O}_s^t := \mathcal{O}_s^{t_1} \circ \dots \circ \mathcal{O}_s^{t_n}$ where the *type* $t = (t_1, \dots, t_n) \in \{0, 1\}^n$ specifies which sequence of operators is applied. The *shift-ACF of type* t and *length* n is then defined by

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

with shift-operations (ShOps) $\mathcal{O}_s^t[x]$ depending on s. It can be shown [2] that shift ACF techniques outperform classical ACF when analyzing multiple repeating events. Note that the classical ACF coincides with the type 1 shift-ACF.

For a finite signal x of length N we define the *self-similarity matrix* $S_x := (x(k) \cdot \overline{x(\ell)})_{0 \le k, \ell < N}$. Then, $\mathrm{ACF}[x](s)$ is the sum

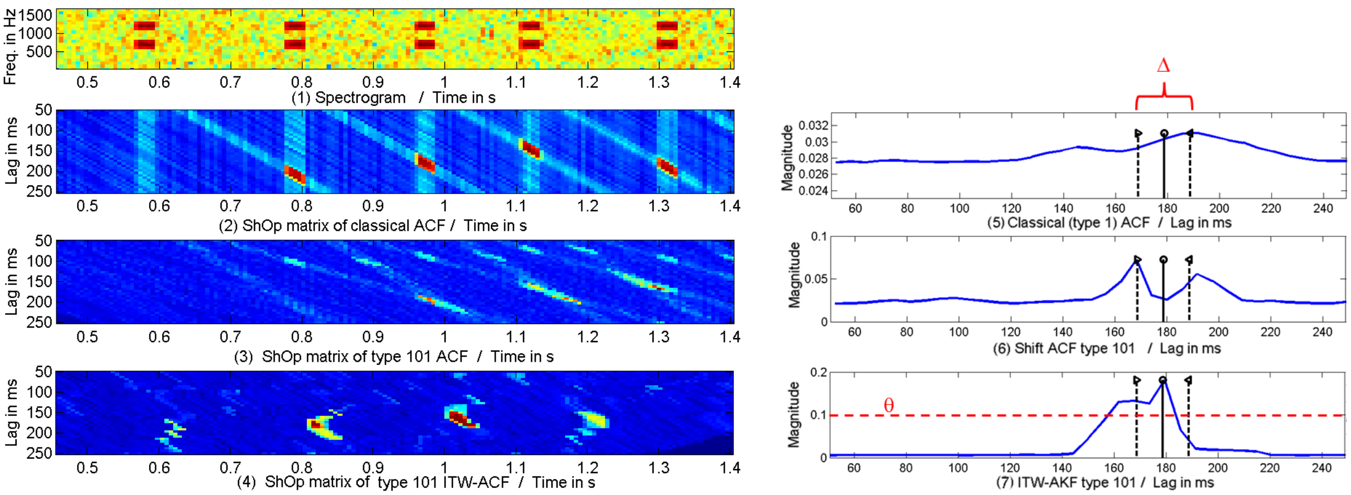

Fig. 1. Left: (1) Spectrogram of 5 subsequent DTMF-tones. (2)-(4) ShOp-matrices for different ACF types (2),(3) and time-warped ACF (4). Right: Different ACFs corresponding to (2)-(4): (5) standard (type 1) ACF, (6) type 101 shift ACF, (7) warped ACF of type 101. Ground truth IOI (circle) and 20 ms tolerance region (dashed lines) are indicated in black. Parameters θ and Δ are explained in Sect.4

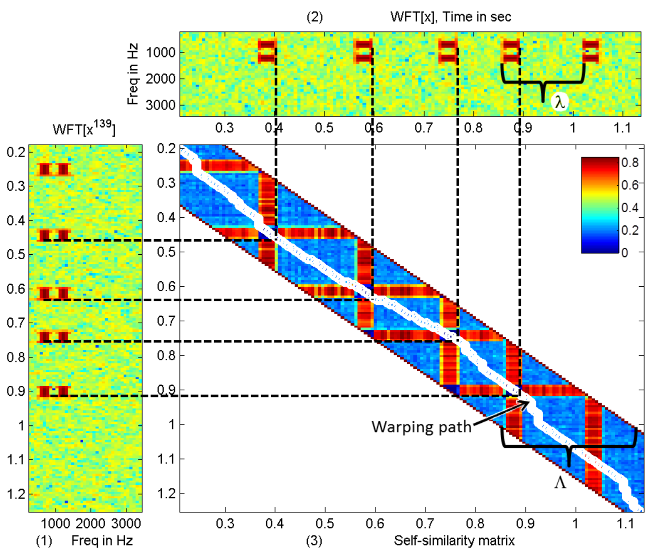

Fig. 2. Illustration of restricted DTW: (1)+(2) Spectrograms of DTMF-signal and delayed version. (3) Restricted self-similarity matrix and warping path (white circles).

of the s-th side diagonal of S_x. This side-diagonal consists of all non-zeros values of the ShOp $\mathcal{O}_s^1[x] = x(k) \cdot \overline{x^s(k)}$. Note that all introduced ACF-concepts carry over to vector-valued signals x where product- and minimum-operations are performed component-wise. Particularly, x will be a time-series of spectrogram columns for the rest of this paper.

Fig. 1 on the left shows (1) the spectrogram of a sequence of 5 identical DTMF-tones (Dual-tone multi-frequency signaling) with some added Gaussian background noise. The basic IOI is about 180 ms but the single tones are temporally jittered. The rows of the *ShOp-matrix* (2) show the ShOp magnitudes (vectors are replaced by their 1-norms) of the classical $\mathrm{ACF}^1[x]$. More explicitly, the s-th row of (2) contains the sequence $(\|\mathcal{O}_s^1[x](k)\|_1)_k$. The different vertical (lag-) positions show the different IOIs of the DTMF-tones. On the right, Fig. 1 shows ACFs corresponding to (2)-(4): (5) shows $\mathrm{ACF}^1[x]$ which is obtained simply from the row-sums of the ShOp-matrix. As a result of the jittered IOIs, no clear peak at the basic IOI (indicated by a circle) nor in a 10 ms neighborhood (indicated by dashed lines) can be observed. Fig. 1 (6) shows that the type 101 shift-ACF, with corresponding ShOps shown in (3) also fails to properly represent the basic IOI around 180 ms.

2.2. Dynamic Time Warping

To compensate for jitter, we propose to compute ShOps of sequences that are temporally aligned using DTW. For two vector sequences $x = (x_1, \dots, x_N)$ and $x' = (x'_1, \dots, x'_M)$, the task of DTW can be summarized as follows. First, a similarity matrix $S_{x,x'} := (d(x_k, x'_\ell))_{1\le k\le N, 1\le \ell\le M}$ is computed, where d is a suitable distance measure. In our case we will use the cosine measure $d(\xi,\zeta) = \langle \xi,\zeta\rangle / (\|\xi\|_2 \cdot \|\zeta\|_2)$. An alignment then amounts to finding an optimal *warping path* $\mathrm{DTW}(x,x') := w := ((a_1,b_1), \dots (a_P,b_P))$ through $S_{x,x'}$, such that the *path similarity* $\delta(w) := \sum_{i=1}^{P} d(x_{a_i}, x'_{b_i})$ of w is maximized. Additionally, warping paths are restricted to start in the upper left corner, $(a_1,b_1) = (1,1)$, end in the lower right, $(a_P,b_P) = (N,M)$, and obey the step condition, $(a_{i+1},b_{i+1}) = (a_i,b_i) + \sigma$. where $\sigma \in \{(0,1),(1,0),(1,1)\}$. The warping path can be found efficiently using dynamic programming. We refer to [4] for details.

In this paper, we will use DTW-based alignment of equal-length sequences, i.e., $N = M$, with the global constraint that the warping path is restricted to a band-region of size Λ around the main diagonal, the Sakoe-Chiba band [3].

Fig. 2 illustrates an alignment of a DTMF-signal with spectrogram shown in (1) and a shifted version of itself (2). The signal consists of 5 identical tones at a basic IOI of $\lambda = 139$ ms with random jitter of $|\delta_i| < 20$ ms. In (3) the similarity matrix is shown. Regions outside the Sakoe-Chiba band of width Λ are left blank. The optimal warping path inside the band is shown by white circles. The dashed lines indicate how DTW compensates for jitter and thus properly aligns the events of the DTMF-signal and its shifted version.

3. ITERATED TIME-WARPED ACF

DTW is now used to align a signal x and its shifted version x^s prior to the shift operation. For practical implementation, x and x^s will be replaced by finite, equal length sequences x^H and x^T in the subsequently described algorithm. The general idea is that DTW compensates the misalignment of repeating components due to jitter. By using a restricted version of DTW with a Sakoe-Chiba band of size $\Lambda \leq \lambda$, an alignment is only possible between uniquely determined pairs of components in x and x^s, if the above maximum individual event deviation of $\max_i |\delta_i| < \lambda/2$ holds. More precisely, Λ-restricted DTW can compensate temporal jitter with maximum individual event deviation $< \Lambda/2$, but not more.

The time-warped ShOp for x at lag s and type $t' \in \{0,1\}$ is computed as follows:

1. *Obtain subsequences to be DTW-aligned* as $x^T := x_{s+1:N}$ ("tail" sequence of x) and $x^H := x_{1:N-s}$ ("head" sequence of x^s).
2. *Use Λ-band-restricted DTW* to align x^H and x^T resulting in a warping path $w = \mathrm{DTW}(x^H, x^T)$, where $w := ((a_1,b_1),\dots(a_P,b_P))$ and $a_i, b_j \in [1:N-s]$, refering to the common length $N-s$ of x^H and x^T.
3. *Compute the w-warped t'-operation* by defining the sequence $y(k) := O_0^{t'}(x^H(a_k), x^T(b_k))$, for $1 \leq k \leq P$.
4. *Unwarp y* with respect to x^H, by using the path projection $p := p^1 := (a_1,\dots,a_P)$ on the first component of w. To this end, an unwarping operation u_p is defined for $p:[1:P] \to [1:N-s]$ and applied to the sequence y. Let $p^{-1}(k) := \{i \,|\, p(i) = k\}$ be the index set of all positions contributing to $k \in [1:N-s]$. Because of the start- und end-conditions that hold for the warping path, i.e., $(a_1,b_1) = (1,1)$ and $(a_P,b_P) = (N,N)$, as well as the admissible step sizes $\sigma \in \{(0,1),(1,0),(1,1)\}$, it follows that $|p^{-1}(k)| \geq 1$ for all k. Hence we can define

$$u_p[y](k) := \frac{1}{|p^{-1}(k)|} \sum_{\ell \in p^{-1}(k)} y(\ell). \tag{3}$$

5. The *DTW-ShOp* is then defined by unwarping y w.r.t. the first component p^1, this is, $\tilde{O}_s^{t'}[x] := u_{p^1}[y]$. Note that alternatively, unwarping w.r.t. the second component $p^2 := (b_1,\dots,b_P)$ can be used. Our experiments show that this leads to essentially the same results.

Then, the iterated time-warped ACF (ITW-ACF) of type $t = (t_1,\dots,t_n)$ is defined as

$$\text{ITW-ACF}^{\,t}[x](s) := \sum_{k \in \mathbb{Z}} \tilde{O}_s^{\,t}[x](k), \tag{4}$$

where $\tilde{O}_s^{\,t}[x] := \tilde{O}_s^{\,t_1}[\tilde{O}_s^{(t_2,\dots,t_n)}[x]]$.

Fig. 1 (7) shows ITW-ACF$^{101}[x]$ for the jittered DTMF signal, clearly exhibiting a peak close to the basic IOI. The rows of Fig. 1 (4) show the type 101 DTW-ShOps. As compared to the corresponding shift-ACF ShOps (3), the energy is clearly concentrated at regions corresponding to aligned events. As compared to classical ACF (5), those regions share a common lag range, leading to the clear peak in the ITW-ACF of Fig. 1 (7).

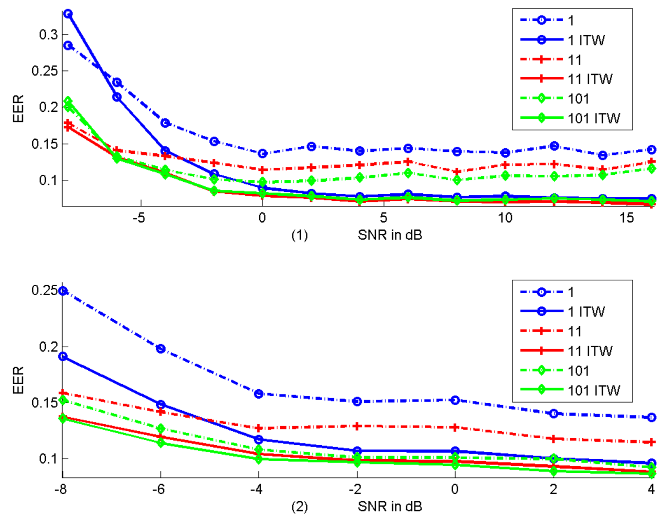

Fig. 3. Detection-EERs for (1) DTMF-signals and (2) bioacoustic events depending on SNR-level of added Gaussian noise.

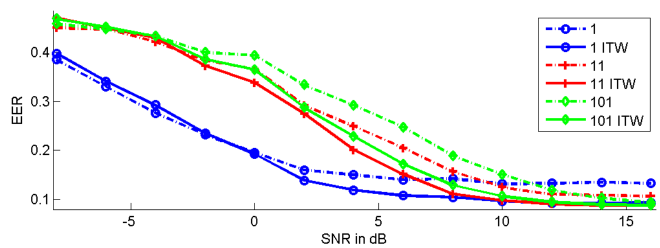

Fig. 4. Detection-EERs for bioacoustic events depending on SNR-level of added realistic bioacoustic background noise.

4. EVALUATION

As ITW-ACF can be used as a basic building block for various kinds of detection algorithms, we perform an evaluation by comparing ITW-ACF to classical ACF(as a baseline) and different types of shift-ACF. We evaluate detection performance of the different methods based on ACFs as depicted in Fig. 1 (5)-(7) computed for various synthesized event sequences with known basic IOI. First, all ACFs a are normalized to $a/\|a\|_1$. Then, all lags inside a fixed tolerance region of width Δ around the known basic IOI (circles) with ACF-value above a fixed threshold $0 \leq \theta \leq 1$ will be considered as true positives (TP). Outside this tolerance region, an ACF greater than θ is considered as a false positive (FP). Parameters $\theta = 0.1$ and $\Delta = 20$ ms are indicated in red in Fig. 1. By varying θ, a ROC curve showing TP-rate vs FP-rate is computed for each ACF. Each experiment is repeated, interpolated ROC curves are averaged and an equal error rate (EER) is computed as single quality measure. EER represents the point on the interpolated ROC curve where the FP-rate equals 1-TP-rate (missed detection rate).

As a first experiment, we use a completely synthetic setting of 5 identical DTMF-tones repeated at an initial (random) IOI between 80 and 200 ms, where each event is offset by a maximum individual deviation of ± 20 ms, i.e., $\lambda/2 = 20$. After shifting the events according to the random jitter, the ground truth IOI is adjusted by choosing a best fitting IOI in a least squares sense.

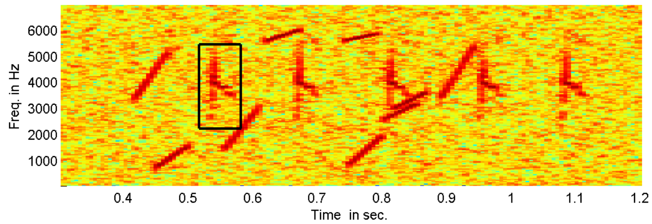

Fig. 5. 5-fold repeated bioacoustic event (framed) with 9 added chirps at 0 dB Gaussian background noise.

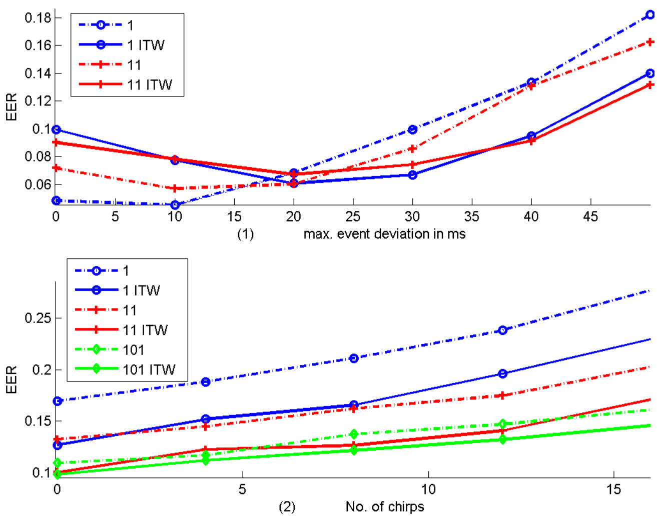

Fig. 6. Detection-EERs for bioacoustic events with added (1) realistic bioacoustic background noise at 10 dB, depending on maximum event deviation in ms, and (2) Gaussian noise at -5 dB for different numbers of added chirp signals.

For the first experiment, Gaussian background noise is added at different SNRs. Here, the SNR is computed w.r.t. the event duration, so at 0 dB the noise during an event has the same energy as the event itself. Fig. 3 (1) shows detection-EERs based on different types of shift-ACF and corresponding ITW-ACFs, depending on the SNR-level of the added Gaussian noise averaged over 750 trials. The DTW band-restriction – valid for the first 3 experiments – is set to $\Lambda := \lambda$, and the evaluation tolerance is $\Delta := 20$ ms. Except for very strong noise, ITW-ACF (bold lines - in the legend indicated by ACF-type and *'ITW'*) outperforms shift-ACF, where type 1 ACF corresponds to classical ACF. Furthermore, higher length ACFs outperform the classical length 1 ACF.

In a second experiment we manually annotated a dataset of 61 bioacoustic events, each taken from sequences of repeated event. For this we used audio recordings from the Reference System of Animal Vocalisations of Museum für Naturkunde Berlin[1] and the xeno-canto[2] platform. By randomly selecting single events from this dataset in 750 trials, we repeated the first experiment. The resulting EER-graphs shown in Fig. 3 (2) largely confirm the findings of the first experiment.

For the third experiment we annotated 10 passages from the above recordings containing no dominant animal vocalizations. Those were randomly used as source for added background noice instead of the Gaussian noise resultung in EER-graphs shown in Fig. 4. While again ITW-ACF outperforms classical resp. shift-ACF, longer shift-ACF types are beneficial only for better SNRs. Note that due to the way of background noise normalization we use, absolute SNRs of this and the first two experiments cannot be compared directly.

In the fourth experiment, we evaluate the effect of the amount of jitter by changing the maximum individual event deviation from 0 to 50 ms. The setting is as in experiment 3, i.e., random bioacoustic events and random real bioacoustic background are added at a fixed SNR of 10 dB. The EER-graphs in Fig. 6 (1) show a transition from better performance of classical resp. shift-ACF for low deviations to better performance of ITW-ACF for maximum deviations exceeding 20 ms. While this is perfectly reasonable, a decrease in performance of ITW-ACF towards low deviations is noteable. The main reason for this is that for low deviations, the ITW-ACF by construction results in box-shaped regions of width Λ around the basic IOIs, in our setting leading to worse EERs. For higher deviations, those regions get more peaky as shown in Fig. 1 (5), hence improving EERs. In the fifth experiment, we evaluate detection performance in the presence of secondary events. For this we selected short-time chirp signals as they are similar to a frequent type of bird vocalization. In our evaluation we added 0 - 16 chirps to each event sequence generated from the above bioacoustic dataset. Here, each added chirp has the same energy as the bioacoustic event. Chirp durations were selected to be the same as the respective bioacoustic events, start and end frequency were randomly and independently selected. Gaussian background noise at -5 dB is added, $\lambda/2 = 20$, $\Lambda := \lambda$, and $\Delta := 20$ are used as in experiments 1–3. As an example, Fig. 5 shows a 5-fold repeated bioacoustic event with 9 added chirps at 0 dB Gaussian background noise. The first occurrence of the event is marked by a black frame. Fig. 6 (2) shows resulting EER-graphs where the advantage of ITW-ACF over classical ACFs, particularly in combination with higher order ACF-types is obvious.

5. CONCLUSIONS

In this paper we have proposed to combine shift-ACF with restricted DTW in order to robustly detect jittered multiply repeating audio events. By iterating ACF and DTW we have derived the ITW-ACF algorithm. In our evaluations we compared ITW-ACF with classical ACF and shift-ACF for the application of detecting repeated bird vocalizations. To this end we designed different test scenarios using combinations of both artificial and realistic audio events as well as artificial and realistic background noises. It turns out that the proposed approach significantly improves EERs in many scenarios w.r.t. to previous approaches. For future work it will be hence very promising to identify further scenarios where ITW-ACF can complement or substitute ACF-based approaches. Furthermore, in future work the combination of DTW and ACF proposed in this paper could be improved with respect to its computational complexity by using concepts of subsequence DTW ([4], Chapter 7).

6. REFERENCES

[1] F. Kurth, "Robust Detection of Multiple Bioacoustic Events with Repetitive Structures," in *Proc. Interspeech*, 2016.

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

[1]http://www.animalsoundarchive.org/RefSys/

[2]http://www.xeno-canto.org/

[3] H. Sakoe and S. Chiba, “Dynamic programming algorithm optimization for spoken word recognition,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, Feb 1978.

[4] M. Müller, *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. Springer Publishing Company, Incorporated, 2015.

[5] A. Härmä, M. Karjalainen, L. Savioja, V. Välimäki, U. K. Laine, and J. Huopaniemi, “Frequency-warped signal processing for audio applications,” *J. Audio Eng. Soc*, vol. 48, no. 11, pp. 1011–1031, 2000.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.