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

Frank Kurth and Kevin Wilkinghoff

Fraunhofer FKIE, 53433 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 \( (t_o + l, t_o + 2l, \ldots, t_o + K \lambda) \). In the ACF-based approach the inter-onset-interval (IOI), or lag, \( \lambda \) is then estimated by comparing a signal \( x \) containing the events, with shifted versions of \( x \). In real application scenarios, the IOI \( \lambda \) usually is not perfectly constant but may vary by some jitter \( \delta = (\delta_0, \ldots, \delta_K) \). 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 \( x^\delta \) 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

\[
\text{ACF}[x](n) := \sum_{k \in \mathbb{Z}} x(k) \cdot x(k + n).
\]

The basic principle of the classical ACF is that signal components repeating at a lag of \( n \) samples within an analyzed signal \( x \) are emphasized by a shift-product \( O_s^1[x](k) := x(k) \cdot x^s(k) \), 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 \( |\text{ACF}[x]| \) 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 \( O_s^1 \) iteratively to amplify repeating components. Secondly, \( O_s^1 \) is complemented by a type 0, shift-minimum operator \( O_s^0[x](k) := \min(|x(k)|, |x^s(k)|) \) in order to suppress non-repeating components. This can be generalized by arbitrarily iterating \( n \) operators \( O_s^t := O_s^{t_1} \circ \cdots \circ O_s^{t_n} \) where the type \( t = (t_1, \ldots, t_n) \in \{0, 1\}^n \) specifies which sequence of operators is applied. The shift-ACF of type \( t \) and length \( n \) is then defined by

\[
\text{ACF}^t[x](n) := \sum_{k \in \mathbb{Z}} O_s^1[x](k),
\]

with shift-operations (ShOps) \( O_s^t[x] \) 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 \( S_x := (x(k) \cdot x^s(l))_{0 \leq k, l < N} \). Then, \( \text{ACF}[x](n) \) is the sum
of the \(s\)-th side diagonal of \(S_\Delta\). This side-diagonal consists of all
non-zeros values of the ShOp \(O^s_s[x] = x(k) \cdot x^s(k)\). 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 ACF \(1_x\). More explicitly, the \(s\)-th
row of (2) contains the sequence \(\|O^s_s[x(k)]\|_1\). 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
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 se-
quencies that are temporally aligned using DTW. For two vector
sequences \(x = (x_1, \ldots, x_N)\) and \(x' = (x'_1, \ldots, x'_M)\), the task
of DTW can be summarized as follows. First, a similarity ma-
trix \(S_{x,x'} := (d(x_k, x'_k))_{1 \leq k \leq N, 1 \leq \ell \leq M}\) is computed, where \(d\)
is a suitable distance measure. In our case we will use the co-
sine measure \(d(\xi, \zeta) = (\xi, \zeta)/(||\xi|| \cdot ||\zeta||)\). An alignment then
amounts to finding an optimal warping path \(DTW(x, x') := w :=
((a_1, b_1), \ldots, (a_P, b_P))\) through \(S_{x,x'}\), such that the path simi-
arity \(\delta(w) := \sum_{i=1}^{P} d(x_{a_i}, x'_{b_i})\) of \(w\) is maximized. Addi-
tionally, 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 effi-
ciently 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 \(\Lambda\) around the main diago-
nal, the Sakoe-Chiba band [3].

Fig. 2 illustrates an alignment of a DTMF-signal with spectro-
gram shown in (1) and a shifted version of itself (2). The signal con-
ists of 5 identical tones at a basic IOI of \(\lambda = 139\) ms with random
jitter of \(|\delta| < 20\) ms. In (3) the similarity matrix is shown. Regions
outside the Sakoe-Chiba band of width \(\Lambda\) 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 $\max_i |b_i| < \lambda/2$ holds. More precisely, $\Lambda$-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 $x$).
2. Use $\Lambda$-band-restricted DTW to align $x^H$ and $x^T$ resulting in a warping path $w = \text{DTW}(x^H, x^T)$, where $w := ((a_1, b_1), \ldots (a_P, b_P))$ and $a_i, b_i \in [1 : N - s]$, referring to the common length $N - s$ of $x^H$ and $x^T$.
3. Compute the $w$-warped $t'$-operation by defining the sequence $y(k) := O^w_t(x^H(a_k), x^T(b_k)), \text{ for } 1 \leq k \leq P$.
4. Unwarp $y$ with respect to $x^H$, by using the path projection $p := p^{-1} := (a_1, \ldots, a_P)$ on the first component of $w$. To this end, an unwarping operation $u_p$ is defined for $p : [1 : P] \rightarrow [1 : N - s]$ 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- and 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

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

5. The DTW-ShOp is then defined by unwarping $y$ w.r.t. the first component $p^1$, this is, $\hat{O}_s^w[x] := u_{p^1}[y]$. Note that alternatively, unwarping w.r.t. the second component $p^2 := (b_1, \ldots, b_P)$ can be used. Our experiments show that this leads to essentially the same results.

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

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

where $\hat{O}_s^w[x] := \hat{O}_s^w[O_{t_1}^w(\ldots O_{t_n}^w)[x]]$.

Fig. 1 (7) shows ITW-ACF $101^t[x]$ 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).

### 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 $\theta$ are normalized to $\theta/||a||_1$. Then, all lags inside a fixed tolerance region of width $\Delta$ around the known basic IOI $\theta$ (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 $\theta$ is considered as a false positive (FP). Parameters $\theta = 0.1$ and $\Delta = 20$ ms are indicated in red in Fig. 1. By varying $\theta$, 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 $\pm 20$ 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.
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


