SFB
823

Comparison of parameter optimization techniques for a music tone onset detection algorithm

Nadja Bauer, Julia Schiffner, Claus Weihs

Discussion Paper

Nr. 42/2012

Comparison of parameter optimization techniques for a music tone onset detection algorithm

Nadja Bauer*, Julia Schiffner, Claus Weihs

Chair of Computational Statistics, Department of Statistics, TU Dortmund,
44221 Dortmund, Germany

Abstract

Design of experiments is an established approach to parameter optimization for industrial processes. In many computer applications, however, it is usual to optimize the parameters via genetic algorithms or, recently, via sequential parameter optimization techniques. The main idea of this work is to analyse and compare parameter optimization approaches which are usually applied in industry with those applied for computer optimization tasks using the example of a tone onset detection algorithm. The optimal algorithm parameter setting is sought in order to get the best onset detection accuracy.

We vary in our work essential options of the parameter optimization strategies like size and constitution of the initial designs in order to assess their influence on the evaluation results. Furthermore we test how the instrumentation and the tempo of music pieces affect the optimal parameter setting of the onset detection algorithm.

Keywords: Sequential parameter optimization, Design of experiments,
Tone onset detection

1. Introduction

Parameter optimization is an important issue in almost every industrial process or computer application. It is remarkable that the parameter op-

*Corresponding author

Email addresses: bauer@statistik.tu-dortmund.de (Nadja Bauer),
schiffner@statistik.tu-dortmund.de (Julia Schiffner),
weihs@statistik.tu-dortmund.de (Claus Weihs)

timization strategies which are applied in industry and in computer applications differ significantly. In industry often strong assumptions regarding the relationship between the target variable and the influential parameters are made and then such experimental designs are used, which fulfill special criteria (like A- or D-optimality). Many computer optimization approaches, in contrast, aim to cover the parameter space uniformly by heuristically generated designs (like Latin Hypercube Sampling designs). Furthermore, when planning trial series in industry many aspects are considered (like improving internal and external validity, identifying and controling disturbing factors or modeling interactions between the influential factors) which are often neglected when planning computer optimizations. The number of trials (or function evaluations) is also very different: While in industry often maximally 100 trials are allowed, the number of function evaluations in computer optimization frequently exceeds ten thousands.

The main idea of this work is to combine and compare industrial and computer simulation based parameter optimization techniques for the optimization of a music signal analysis algorithm. The tone onset detection algorithm, which we aim to optimize here, is presented in Section 2. Two important factors that can influence the optimization results are the optimization strategy and the music data set under consideration. The optimization strategy determines how trial points, where the function is evaluated, are selected. In Sections 3 and 4 we present a parameter optimization approach and define characteristics of industrial and computer based parameter optimization which we will compare systematically. Also, we systematically vary characteristics of the music data in order to assess their influence on the evaluation results. Section 5 gives the procedure of the music data set generation. Section 6 presents the simulation results. Finally Section 7 summarizes our work and provides points for future research.

2. Onset detection algorithm

A tone onset is the time point of the beginning of a musical note or other sound. Onset detection is an important step for music transcription and other applications like timbre or meter analysis (see 5, for a tutorial on onset detection).

The algorithm we will use here is based on two approaches proposed in (2): In the first approach the amplitude slope and in the second approach the change of the spectral structure of an audio signal are considered as indicators for tone onsets. The ongoing audio signal is split up into windows of length $\boldsymbol{L}$ samples with an overlap of $\boldsymbol{O}$ per cent. In each window (starting with the second) two features are evaluated: The difference between amplitude maxima ($\boldsymbol{F1}$) and the correlation coefficient between the spectra ($\boldsymbol{F2}$) of the current and the previous window, respectively. Each of the vectors $\boldsymbol{F1}$ and $\boldsymbol{F2}$ is then rescaled into the interval [0,1].

For each window a combined feature $\boldsymbol{CombF}$ is calculated as $\boldsymbol{CombF} = \boldsymbol{W} \cdot \boldsymbol{F1} + (1 - \boldsymbol{W}) \cdot \boldsymbol{F2}$, where the weight $\boldsymbol{W} \in [0, 1]$ is a further parameter, which specifies the influence of each feature on the sum. In (4) we investigated further feature combination approaches, where this approach provided the best results. In order to assess, based on $\boldsymbol{CombF}$, if a window contains a tone onset a threshold is required. We will use here a $\boldsymbol{Q}$%-quantile of the $\boldsymbol{CombF}$-vector as such threshold, where $\boldsymbol{Q}$ is the fourth algorithm parameter. If the $\boldsymbol{CombF}$-value for the current window, but neither for the preceding nor for the succeeding window, exceeds the threshold, an onset is detected in this window. If the threshold is exceeded in multiple, consecutive windows, we assume that there is only one onset, located in that window with the maximal $\boldsymbol{CombF}$-value in this sequence.

For each window with an onset detected its beginning and ending time points are calculated and the onset time is then estimated by the centre of this time interval. In this work we assume a tone onset to be correctly detected, if the absolute difference between the true and the estimated onset time is less than 50 ms (see 7).

As quality criterion for the goodness of the onset detection the so called F-value is used here: $F = \frac{2c}{2c+f^{+}+f^{-}}$, where c is the number of correctly detected onsets, f^{+} is the number of false detections and f^{-} denotes the number of undetected onsets (7). Note that the F-value lies always between 0 and 1. The optimal F-value is 1.

The studied ranges of possible settings for the onset detection algorithm parameters are: $\boldsymbol{L}$ (window length in samples): 512, 1024 and 2048, $\boldsymbol{O}$ (overlap in per cent): $0 - 50$ with step size 5, $\boldsymbol{W}$ (weight of the features): $0 - 1$ with step size 0.05, $\boldsymbol{Q}$ (%-quantile): $1 - 30$ with step size 1.

3. Sequential parameter optimization

An experimental design is a scheme that prescribes in which order which trial points are evaluated. One of our aims here is to compare the classical parameter optimization, where all trial points are fixed in advance, with the sequential parameter optimization, where a relatively small initial design is given and the next trial points are chosen according to the results of previous experiments.

We consider a non-linear, multimodal black-box function $f : \mathbb{R}^k \to \mathbb{R}$, $\boldsymbol{x} \mapsto f(\boldsymbol{x})$ of k parameters. We aim to minimize f with respect to $\boldsymbol{x}$. Let $\boldsymbol{V} \subset \mathbb{R}^k$ denote the feasible parameter space. The following procedure of (sequential) parameter optimization is used.

1. Let $D \subseteq \boldsymbol{V}$ denote the initial experimental design with $N_{initial}$ trial points and let $Y = f(D)$ be the set of function values of points in D.
2. Repeat the following sequential step until the termination criterion is fulfilled:
   - 2.1 Generate a random number s from the distribution: $P(s = 0) = p_0$, $P(s = 1) = 1 - p_0$, $0 \leq p_0 \leq 1$.
     - 2.1a If $s = 0$, fit a model M which models the relationship between D and the response $Y = f(D)$. Find the next trial point $d_{next} \subset \boldsymbol{V}$, which minimizes the model prediction.
     - 2.1b If $s = 1$, let $D_{sample} \subseteq \boldsymbol{V}$ denote a design with N_{sample} trial points from the parameter space $\boldsymbol{V}$. For each point in D_{sample} calculate the Euclidean distance to all points in D and sum them up. The next trial point d_{next} is that point in design D_{sample}, which has a maximal sum of Euclidean distances.
   - 2.2 Evaluate $y_{next} = f(d_{next})$ and update $D \longleftarrow D \cup d_{next}$, $Y \longleftarrow Y \cup y_{next}$.
3. Return the optimal value y_{best} of the target variable Y and the associated parameter setting d_{best}.

The challenge in sequential design of experiments is to find the appropriate next trial point to evaluate. The major differences between the existing algorithms for sequential parameter optimization lie in step 2.1. A popular approach here is to just use step 2.1a: Fitting a user-chosen model M and calculating its prediction for a sequential design $D_{step} \subseteq \boldsymbol{V}$ of size $N_{step} \gg N_{initial}$. The next trial point then is the point in D_{step} with the best predicted value (1). However, choosing the next trial points in this way may lead to convergence to a local optimum of f. A suitable approach here might be to take into account not only the model prediction for each point in D_{step} but also the distances of these points to already evaluated trial points. Such a methodology is already used in the case of Kriging models (expected improvement criterion, 10), which is unfortunately suitable exclusively for these models. Nevertheless, in order to consider the above mentioned distances of new points to already evaluated points we implement here a simple approach for the exploration of the parameter space: In step 2.1 the next trial point is chosen according to the model prediction (step 2.1a) with a user-defined probability p_0 and according to the distance to already evaluated trial points (step 2.1b) with probability $1 - p_0$.

Our settings for the parameter optimization approach presented above are: D_{step} is a Latin Hypercube Sampling (LHS) design (a design which covers the parameter space uniformly[1], 15) with $N_{step} = 20.000$ points and D_{sample} is an LHS design with $N_{sample} = 500$. Note that, as described in step 2.1b, for each point in D_{sample} we calculate distances to the points in D, therefore N_{sample} should not be chosen too large. The termination criterion in step 2 is defined by the total number of evaluations (N_{total}) of the function f. The probability p_0 is set to 0.9. Details regarding the model M are discussed in Section 4.

Further important issues here are the initial design and the number of sequential steps. We propose different settings for the initial designs using an experimental scheme, in which the size of the sequential design and its type are considered as control variables. For construction of initial designs information about the experimental parameters is required. According to Section 2 there are four parameters to be optimized: $\boldsymbol{L}$, $\boldsymbol{O}$, $\boldsymbol{W}$ and $\boldsymbol{Q}$.

As classical parameter optimization strategy we use here a full factorial design with 3 levels for each parameter (81 trial points). After evaluating f in these 81 points a so called verification step is conducted: We identify the next trial point (in this case just with step 2.1a) and evaluate f at this point. The total number of evaluations, $N_{total} = 82$, should not be exceeded by all further parameter optimization strategies in order to facilitate comparability. Other settings for the size of the initial designs are approx. one half and approx. one third of the evaluation budget (N_{total}). Here we aim to investigate, whether and (if so) how the size of initial designs influences the optimization results.

We consider two different types of initial designs: "textbook" designs, which fulfill special criteria, and LHS designs. LHS initial designs are commonly used in (sequential) parameter optimization of computer applications,

[1] We use here `randomLHS` command from *R*-package `lhs` (6).

Table 1: Strategies of (sequential) parameter optimization where N_{seq_step} is the number of sequential steps

strategy	initial design	$N_{initial}$	N_{seq_step}	N_{total}
Classic	3^4 full factorial design	81	1	82
Orth	orthogonal design with 3 (for $\boldsymbol{L}$) or 4 (for $\boldsymbol{O}$, $\boldsymbol{A}$ and $\boldsymbol{Q}$) factor levels	48	34	82
Centr	central composite design with inner star	25	57	82
$\mathbf{LHS_{81}}$	LHS design	81	1	82
$\mathbf{LHS_{48}}$	LHS design	48	34	82
$\mathbf{LHS_{25}}$	LHS design	25	57	82

while the “textbook” designs are often applied to optimization of industrial processes. We employ both in order to assess which leads to better results.

Table 1 presents our parameter optimization strategies. We decided to implement two widely used “textbook”-designs (in addition to the full factorial design mentioned above): A central composite design with inner star (17) with 25 trial points and an orthogonal design with 48 trial points[2]. The size of the central composite design ($k^2 + 1 + 2 \cdot k$) depends on the number of parameters k, which is here 4. The size of the orthogonal design depends on the number of parameters and the number of their levels. For the generation of the orthogonal design we use the *R*-package `DoE.base` (9) where for our number of parameters and number of levels (see Table 1) only a design with 48 trial points was possible. The disadvantage of most “textbook”-designs in comparison with LHS-designs is their inflexibility regarding the design size.

4. Model combination

In step 2.1a in the sequential parameter optimization procedure in Section 3 both a single model and a combined model can be used. In this work we will use four model combination strategies that were introduced and investigated in (3). In the following we will briefly review the main ideas.

Let us assume that m models $M_1, M_2, \ldots, M_m$ are given with response Y and design D which includes the settings of the influential parameters. For each model we first compute a model prediction accuracy criterion (10-fold

[2] We do not present the trial schemes for the initial designs.

cross-validated mean squared error) and then calculate model predictions for each point d_j, $j = 1, \ldots, N_{step}$, of the sequential design D_{step}.

As first model combination method we will use the weighted average approach (**WeightAver**): For each point d_j the weighted sum of the m model predictions is calculated, where the model weights are defined by the associated values of the prediction accuracy criterion. In each sequential step the next evaluation is done at that point d_j which has the best weighted sum of predictions.

In the second combination approach (**BestModel**) we will just choose the best model according to the model prediction accuracy criterion. Then the function f is evaluated at that point d_j which has the best model prediction value.

The third combination method (**Best2Models**) is similar to the second method but in each step we evaluate two points according to the predictions of the two best models. We take care that we do not carry out more function evaluations than allowed (see the termination criterion in Section 3).

In the last model combination approach we determine for each model the ten trial points with the best predicted values (**Best10**). Here for each point d_j we do not only consider the model predictions with associated accuracy criteria but also the number of models for which this point has one of the ten best predicted values. The core idea is to prefer points which belong to the best predictions of many models at the same time. For more details see (3).

5. Data base

Since one of the aims of this work is to determine the influence of the music signal characteristics on the optimal parameter settings of the onset detection algorithm, we designed a special music data set. There are many characteristics which describe a music signal like tempo, genre, instrumentation or sound volume. We consider only the instrumentation and the tempo as control variables when designing the data set. The special characteristic of this data set is that the same tone sequences are recorded by different music instruments with different tempo settings, so that we can explicitly measure the influence of these two control variables on the optimal parameter settings of the onset detection algorithm.

As we need the information about the true onset times and in order to vary the tempo and instrumentation of tone sequences we will work with

Table 2: Used music instruments and their pitch ranges in English tone notation and MIDI-coding

instrument	pitch range	MIDI-coding
guitar	E2-E5	40-76
piano	A0-C8	21-108
flute	C4-C7	60-96
clarinet	D3-F6	50-89
trumpet	E3-D6	52-86
violin	G3-E5	55-76

MIDI-files[3]. However, the MIDI-files are not converted to WAVE-files[4] using synthetic tones (which is the case for most free and commercial converter programs), but using a specially developed program, which employs recordings of real tones for the WAVE-file generation[5]. The challenge here is finding such music pieces, which can be played by all music instruments under consideration. Table 2 presents the music instruments we use in our work as well as their pitch ranges in English tone notation and in MIDI-coding. According to this table the common pitch range includes 17 tones, from 60 to 76 (in MIDI-coding). We found two German folk songs, which fulfill the tone range condition: $S1$[6] with 122 tone onsets from the tone interval [60, 76] and $S2$[7] with 138 tone onsets from the tone interval [65, 74].

The tempo of a music piece can be measured by Beats Per Minute (BPM). We will set the tempo for each piece to 90 BPM (classical tempo marking: andante) and 200 BPM (classical tempo marking: presto). The sampling rate of the recordings is set to 44100 Hz. The names of the music signals follow the pattern: $S1_{\text{tempo}}$_instrument (for example: $S1_{90}$_piano). The total number of music pieces in the data set is 24 (2 tempi, 2 music pieces and 6 instruments).

[3] http://www.midiworld.com/basics/, date 01.06.2012.

[4] http://www.sonicspot.com/guide/wavefiles.html, date 01.06.2012.

[5] This program is introduced in detail by (4).

[6] http://www.ingeb.org/Lieder/haidschi.mid, date 01.06.2012.

[7] http://www.ingeb.org/Lieder/esgetein.mid, date 01.06.2012.

6. Results

In order to compare different (sequential) parameter optimization approaches (see Section 3) we generate an experimental scheme with the following three meta-parameters: *model type*, *model combination type* and *initial design*. We use the programming language *R* (version 2.15.0, 14) for calculation.

The meta-parameter *model type* determines the model which describes the relationship between the onset detection algorithm parameters ($\boldsymbol{L}$, $\boldsymbol{O}$, $\boldsymbol{W}$ and $\boldsymbol{Q}$) and the target variable (F-value, see Section 2). We employ six model types: A full second order model (**FSOM**, *R*-package `rsm`, 12), Kriging (**KM**, *R*-package `DiceKriging`, 8), random forests (**RF**, *R*-package `randomForest`, 13), support vector machines (**SVM**, *R*-package `kernlab`, 11), neural networks (**NN**, *R*-package `nnet`, 16) and the combination of these five models (**COMB**).

The second meta-parameter – *model combination type* – is just meaningful for the sixth model type and has four options (see Section 4): weighted average (**WeightAver**), best model (**BestModel**), best two models (**Best2Models**) and best ten points (**Best10**).

The last meta-parameter – *initial design* – is related to the initial designs of the optimization strategies and has six levels: Three "textbook"-designs with different numbers of trial points and three associated Latin Hypercube Sampling designs (see Table 1). Note that for the initial designs **Classic** and $\mathbf{LHS_{81}}$ the model combination approach **Best2Models** is not possible, because in these cases the number of function evaluations (83) would exceed the experimental budged (82 evaluations).

For each optimization strategy and for each music piece the evaluation is carried out ten times. This is done in order to average out the influence of chance on the outcome. We actually have a maximization problem here, the sign of F will be reversed hence to get a minimization problem (see Section 3).

As we aim to know for each music piece and for each optimization strategy, how close the estimated optima and the true optima are to each other, we find the true optima using a time-consuming grid search. The full factorial design *Grid* consists – according to the ranges of the possible parameter settings (see the last paragraph of Section 2) – of 20790 trial points. For each of 24 music pieces a vector of function values for the *Grid*-design is computed.

In order to assess the goodness of the optimization strategies the following procedure is conducted:

- Let i denote the index of a music song, $i = 1, ..., 24$:
    - let q_i be the 99%-quantile of the vector Y^i_{Grid}, where Y^i_{Grid} is the vector of function values for the *Grid*-design for the i-th song,
    - determine the number of replications (nr_i) of the current optimization strategy, in which an F-value that exceeds the q_i-value was found,
    - compute the relative frequency of these "successful" replications by $freq_i = nr_i/10$,
- compute the goodness of the optimization strategy by $\frac{1}{24} \cdot \sum_{i=1}^{24} freq_i$.

Table 3 shows the goodness values for the parameter optimization strategies. The strategies whose goodness-values exceed 0.95 are marked with an asterisk. One of the most important findings when considering Table 3 is that the strategies with LHS-initial designs (in almost all cases) achieve better goodness-measures than the associated strategies with "textbook"-initial designs. The best result is achieved by the strategy with ID 38, a single Kriging model with initial design $\mathbf{LHS_{48}}$, and the second best result is given by the strategy with ID 52, a combined model (**Best10**) with initial design $\mathbf{LHS_{48}}$.

When considering only the "textbook"-initial designs we observe in contrast to LHS-initial designs that firstly a model combination approach (**Best10**, ID 22) is better than the best single model (Kriging, ID 7), and secondly that initial designs of size 25 seem to be better than designs of size 48. For "textbook" as well LHS-initial designs it is obvious that the classical optimization strategies (where 81 of 82 design points are fixed in advance) are considerably worse than the sequential optimization strategies.

Further we look at the number of function evaluations which are required by each optimization strategy for finding an F-value close to the true optimum. For this purpose we conduct the following procedure for all optimization strategies with initial designs of size 25 and 48:

- Let i denote the index of a music song, $i = 1, \ldots, 24$:
    - let q_i be the 99%-quantile of the vector Y^i_{Grid},

Table 3: Frequency of finding an F-value close to the optimum by the parameter optimization strategies under consideration (the strategies whose goodness-values exceed 0.95 are marked with an asterisk)

model type	model combination type	initial design	[ID]	goodness	initial design	[ID]	goodness
FSOM	-	**Classic**	[1]	0.3875	**LHS$_{81}$**	[27]	0.6292
KM	-	**Classic**	[2]	0.3958	**LHS$_{81}$**	[28]	0.7042
RF	-	**Classic**	[3]	0.3958	**LHS$_{81}$**	[29]	0.6542
SVM	-	**Classic**	[4]	0.3917	**LHS$_{81}$**	[30]	0.6375
NN	-	**Classic**	[5]	0.3833	**LHS$_{81}$**	[31]	0.6583
FSOM	-	**Centr**	[6]	0.2875	**LHS$_{25}$**	[32]	0.6292
KM	-	**Centr**	[7]	0.9125	**LHS$_{25}$**	[33]	0.9625*
RF	-	**Centr**	[8]	0.7167	**LHS$_{25}$**	[34]	0.7125
SVM	-	**Centr**	[9]	0.7458	**LHS$_{25}$**	[35]	0.7958
NN	-	**Centr**	[10]	0.6542	**LHS$_{25}$**	[36]	0.8375
FSOM	-	**Orth**	[11]	0.4667	**LHS$_{48}$**	[37]	0.6500
KM	-	**Orth**	[12]	0.8250	**LHS$_{48}$**	[38]	0.9875*
RF	-	**Orth**	[13]	0.6333	**LHS$_{48}$**	[39]	0.7708
SVM	-	**Orth**	[14]	0.7000	**LHS$_{48}$**	[40]	0.8125
NN	-	**Orth**	[15]	0.7208	**LHS$_{48}$**	[41]	0.8083
COMB	**WeightAver**	**Classic**	[16]	0.3958	**LHS$_{81}$**	[42]	0.6417
COMB	**BestModel**	**Classic**	[17]	0.3917	**LHS$_{81}$**	[43]	0.6708
COMB	**Best10**	**Classic**	[18]	0.3833	**LHS$_{81}$**	[44]	0.6792
COMB	**WeightAver**	**Centr**	[19]	0.6750	**LHS$_{25}$**	[45]	0.8458
COMB	**BestModel**	**Centr**	[20]	0.9375	**LHS$_{25}$**	[46]	0.9375
COMB	**Best2Models**	**Centr**	[21]	0.9167	**LHS$_{25}$**	[47]	0.9501*
COMB	**Best10**	**Centr**	[22]	0.9625*	**LHS$_{25}$**	[48]	0.9585*
COMB	**WeightAver**	**Orth**	[23]	0.7083	**LHS$_{48}$**	[49]	0.8625
COMB	**BestModel**	**Orth**	[24]	0.8500	**LHS$_{48}$**	[50]	0.9458
COMB	**Best2Models**	**Orth**	[25]	0.8625	**LHS$_{48}$**	[51]	0.9625*
COMB	**Best10**	**Orth**	[26]	0.8792	**LHS$_{48}$**	[52]	0.9667*

  - determine for each replication $j, j = 1, \dots, 10$, of the current optimization strategy the number of function evaluations which were sufficient to find an F-value that exceeds the q_i-value. Collect these numbers into the vector $NR_i = (nr_1, nr_2, \dots, nr_{10})'$,
  - compute the mean of vector NR_i: $\overline{NR}_i = \frac{1}{10} \cdot \sum_{j=1}^{10} nr_j$,

- calculate the mean and the standard deviation of the vector $(\overline{NR}_1, \overline{NR}_2, \dots, \overline{NR}_{24})'$

The results are shown in Table 4. According to Table 4 all LHS-initial design strategies (with exception of the strategy with model type **FSOM** and initial design **Orth**) are faster than the associated "textbook"-design strategies (by 5.8 evaluations on average). Frequently the mean number of function evaluations in Table 4 does not exceed the size of the associated initial design. This is caused by the fact that in many cases a sufficiently large F-value has been already achieved in the initial design. The standard deviations (of the number of steps) of the LHS-initial designs is 1.6 to 3.7 times smaller than the standard deviations of the associated "textbook"-designs. This seems to be a further advantage of LHS-initial designs.

Our further analysis of the simulation results refers to the optimal parameter settings of the tone onset detection algorithm. As we mentioned above, we calculated the best parameter setting for each music piece using the full factorial design *Grid*. These optimal settings and the corresponding F-values are presented in Table 5. There is no identifiable system regarding the best settings for parameters ***L*** and ***O*** with respect to instrumentation and tempo. On the one hand this can be caused by the fact that the function, which we aim to optimize, has many local optima and if we consider e.g. ten best trial points for one music piece, we can note a certain inhomogeneity within these points. On the other hand, we generated music pieces using a MIDI to WAVE converter and such music pieces can be only seen as more or less acceptable alternatives to true recordings. At the moment it is for example impossible to model legato (playing musical notes smoothly and connected). This disadvantage is significant especially for wind instruments.

However we can see a system for the optimal settings for parameter ***W***: For the instruments whose tone onsets are characterised by an amplitude increase (piano and guitar) ***W*** has a value greater than 0.5. This means that the amplitude based feature ***F1***, see Section 2) has the major weight in the combined feature (***CombF***). For the wind instruments (flute, clarinet

Table 4: The mean (and the standard deviation) of the number of function evaluations required to find an F-value close to the true optimum

model type	comb. type	initial design	mean (sd)	initial design	mean (sd)
FSOM	-	**Centr**	43.65 (28.62)	$\mathbf{LHS_{25}}$	30.78 (8.22)
KM	-	**Centr**	37.34 (14.55)	$\mathbf{LHS_{25}}$	29.71 (5.03)
RF	-	**Centr**	39.61 (15.79)	$\mathbf{LHS_{25}}$	30.19 (7.52)
SVM	-	**Centr**	41.41 (15.33)	$\mathbf{LHS_{25}}$	34.76 (9.05)
NN	-	**Centr**	42.11 (18.40)	$\mathbf{LHS_{25}}$	34.34 (7.57)
FSOM	-	**Orth**	29.07 (19.96)	$\mathbf{LHS_{48}}$	34.85 (8.98)
KM	-	**Orth**	45.12 (22.99)	$\mathbf{LHS_{48}}$	41.16 (6.06)
RF	-	**Orth**	40.49 (22.00)	$\mathbf{LHS_{48}}$	38.31 (7.09)
SVM	-	**Orth**	44.03 (22.39)	$\mathbf{LHS_{48}}$	39.72 (9.12)
NN	-	**Orth**	44.25 (22.66)	$\mathbf{LHS_{48}}$	39.37 (9.38)
COMB	**WeightAver**	**Centr**	45.50 (17.09)	$\mathbf{LHS_{25}}$	31.70 (5.62)
COMB	**BestModel**	**Centr**	37.17 (12.35)	$\mathbf{LHS_{25}}$	30.94 (5.17)
COMB	**Best2Models**	**Centr**	36.25 (11.62)	$\mathbf{LHS_{25}}$	30.58 (4.07)
COMB	**Best10**	**Centr**	37.61 (12.28)	$\mathbf{LHS_{25}}$	30.91 (4.74)
COMB	**WeightAver**	**Orth**	48.93 (25.33)	$\mathbf{LHS_{48}}$	39.14 (7.11)
COMB	**BestModel**	**Orth**	46.17 (22.84)	$\mathbf{LHS_{48}}$	40.52 (6.68)
COMB	**Best2Models**	**Orth**	44.17 (21.68)	$\mathbf{LHS_{48}}$	42.38 (5.94)
COMB	**Best10**	**Orth**	43.77 (21.56)	$\mathbf{LHS_{48}}$	41.23 (6.23)

and trumpet) and for violin we can note small optimal $\boldsymbol{W}$-values ($\boldsymbol{W} < 0.5$ with the exception of two cases). This means that in these cases the spectral based feature has more influence on the combined feature. This confirms the observations reported in (4). Moreover, it is interesting that for the instruments piano, guitar, clarinet and trumpet very large optimal F-values can be achieved (larger than 0.95), whereas the optimal F-values for flute and violin lie between 0.831 and 0.974.

Furthermore the relationship between the optimal settings for parameters $\boldsymbol{Q}$, $\boldsymbol{L}$ and $\boldsymbol{O}$ and the tempo of music pieces is investigated. The number of windows in one second depends on $\boldsymbol{L}$ and $\boldsymbol{O}$. It is monotonically decreasing in $\boldsymbol{L}$ and increasing in $\boldsymbol{O}$. For a fixed number of onsets in one second (music tempo) the following can be supposed: Since $\boldsymbol{Q}$ is a quantile of the $\boldsymbol{CombF}$-vector (which rules the number of windows, in which an onset can be detected), with increasing number of windows the parameter $\boldsymbol{Q}$ should decrease.

However, for a fixed number of windows in a second we suppose the following: With increasing tempo we expect that there are more windows that contain a tone onset (parameter $\boldsymbol{Q}$ should be set higher). We fit a linear regression with $\boldsymbol{Q}$ as response and $\boldsymbol{L}$, $\boldsymbol{O}$ and tempo of music piece (in BPM) as influencing parameters:

$$\boldsymbol{Q} = \beta_0 + \beta_1 \cdot \boldsymbol{L} + \beta_2 \cdot \boldsymbol{O} + \beta_3 \cdot \mathit{tempo} + \boldsymbol{\varepsilon}.$$

For the model parameter estimation we use the data from Table 5. The model with estimated parameters is

$$\widehat{\boldsymbol{Q}} = -4.263 + 0.008 \cdot \boldsymbol{L} - 0.05 \cdot \boldsymbol{O} + 0.036 \cdot \mathit{tempo}.$$

The value for the adjusted R^2 is 0.926. This indicates a good model fit. The above presented suppositions regarding the direction of the relationship between response and influencing variables can be confirmed by the estimated model. For our further research we will estimate parameter $\boldsymbol{Q}$ through this model in order to reduce the number of parameters to optimize.

Table 5: Best parameter settings according to the evaluation of full-factor design *Grid*

inst.		$S1_{90}$	$S1_{200}$	$S2_{90}$	$S2_{200}$	inst.	$S1_{90}$	$S1_{200}$	$S2_{90}$	$S2_{200}$
piano	$\boldsymbol{L}$	512	2048	512	1024	guitar	1024	2048	512	2048
	$\boldsymbol{O}$	0	30	0	0		0	0	0	5
	$\boldsymbol{W}$	1.00	0.55	0.95	0.85		0.85	0.90	0.60	0.55
	$\boldsymbol{Q}$	3	18	2	10		5	18	2	18
	F-v.	0.996	0.992	0.996	0.996		0.996	0.997	0.996	0.996
clarinet	$\boldsymbol{L}$	512	1024	1024	1024	flute	2048	1024	512	2048
	$\boldsymbol{O}$	5	50	30	50		20	15	30	50
	$\boldsymbol{W}$	0.65	0.35	0.20	0.35		0.30	0.05	0.00	0.00
	$\boldsymbol{Q}$	3	6	6	6		17	12	2	18
	F-v.	1.000	0.996	0.996	0.967		0.911	0.875	0.974	0.934
trumpet	$\boldsymbol{L}$	1024	1024	512	512	violin	1024	1024	512	512
	$\boldsymbol{O}$	0	50	0	20		50	30	50	0
	$\boldsymbol{W}$	0.60	0.15	0.05	0.10		0.20	0.15	0.05	0.10
	$\boldsymbol{Q}$	6	7	4	8		3	8	2	9
	F-v.	1.000	0.992	1.000	1.000		0.843	0.832	0.923	0.902

7. Conclusion

In the following we will summarize our work. Different strategies for sequential parameter optimization were compared on the basis of an algorithm for tone onset detection. We systematically tested the influence of initial design characteristics and model types on different goodness-measures of the optimization strategies. The LHS-initial designs yield the better results both by achieving a sufficiently good value of the target variable and by their "speed" in comparison with the "textbook"-designs. Furthermore we noticed for the LHS-initial design strategies that by using initial designs with 48 trial points (approx. one half of the evaluation budget) slightly better goodness-values (according to Table 3) could be achieved in comparison with the initial designs of smaller size (25 trial points, approx. one third of the evaluation budget) but that the number of necessary function evaluations for finding a sufficiently large F-value rises by about 5 evaluations. Regarding the meta-parameter *model* we can see that the best model is Kriging (**KN**), but that the model combination approaches **Best2Models** and **Best10** also perform well. Nevertheless, these model combination approaches are more time-consuming than Kriging since the model prediction accuracies have to be calculated in each sequential step.

Please note that it might not be unproblematic to generalize the above results. This is because we used a very specific data base, which in fact increases the internal validity of our study (regarding the analysis of best parameter settings) but reduces the external validity.

For our further research it is important, on the one hand, to apply the different parameter optimization strategies defined here on other real or artificial optimization problems (by using e.g. Black-Box Optimization Benchmarking (BBOB) templates which are available on the COCO (COmparing Continuous Optimisers) platform[8]). On the other hand, we have to investigate the properties of the proposed onset detection algorithm by applying it to a wider range of data sets. For this reason it is important firstly to define interesting music characteristics (e.g. monophony/polyphony, instrumentation, classic/modern, slow/fast) and then find appropriate music pieces for each combination of these characteristics.

Furthermore, a parameter optimization of a more complex music signal analysis algorithm like an algorithm for music transcription is planned. In

[8] http://coco.gforge.inria.fr/doku.php, date 01.06.2012.

this case, however, a multi-objective parameter optimization will be required.

Acknowledgements

This work has been supported by the Collaborative Research Centre "Statistical Modelling of Nonlinear Dynamic Processes" (SFB 823) of the German Research Foundation (DFG), within the framework of Project C2, "Experimental Designs for Dynamic Processes".

References

[1] Bartz-Beielstein, T., Lasarczyk, C., Preuß, M., 2005. Sequential parameter optimization, in: McKay, B. (Ed.), Proceedings 2005 Congress on Evolutionary Computation (CEC'05), Piscataway NJ: IEEE Press, Edinburgh. pp. 773–780.

[2] Bauer, N., Schiffner, J., Weihs, C., 2010. Einsatzzeiterkennung bei polyphonen Musikzeitreihen. Discussion Paper 22/2010. SFB 823, TU Dortmund.

[3] Bauer, N., Schiffner, J., Weihs, C., 2012a. Comparison of classical and sequential design of experiments in note onset detection, in: Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin Heidelberg. Accepted.

[4] Bauer, N., Schiffner, J., Weihs, C., 2012b. Einfluss der Musikinstrumente auf die Güte der Einsatzzeiterkennung. Discussion Paper 10/2012. SFB 823, TU Dortmund.

[5] Bello, J.P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., Sandler, M.B., 2005. A tutorial on onset detection in music signals. IEEE Transactions on speech and audio processing 13, 1035–1047.

[6] Carnell, R., 2009. lhs: Latin Hypercube Samples. R package version 0.5.

[7] Dixon, S., 2006. Onset detection revisited, in: Proc. of the 9th Int. Conference on Digital Audio Effects (DAFx-06), pp. 133–137.

[8] Ginsbourger, D., Roustant, O., 2010. DiceOptim: Kriging-based optimization for computer experiments. R package version 1.0.

[9] Groemping, U., 2011. Relative projection frequency tables for orthogonal arrays. Technical Report. Reports in Mathematics, Physics and Chemistry 1/2011.

[10] Jones, D., Schonlau, M., Welch, W., 1998. Efficient global optimization of expensive black-box functions. J. Global Optimization 13, 455–492.

[11] Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A., 2004. kernlab – an S4 package for kernel methods in R. Journal of Statistical Software 11, 1–20.

[12] Lenth, R.V., 2009. Response-surface methods in R, using rsm. Journal of Statistical Software 32, 1–17.

[13] Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2, 18–22.

[14] R Core Team, 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0.

[15] Stein, M., 1987. Large sample properties of simulations using latin hypercube sampling. Technometrics 29, 143–151.

[16] Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S. Springer, New York. fourth edition. ISBN 0-387-95457-0.

[17] Weihs, C., Jessenberger, J., 1999. Statistische Methoden zur Qualitätssicherung und -optimierung in der Industrie. Wiley-VCH, Weinheim.