SFB
823

A computational study of auditory models in music recognition tasks for normal-hearing and hearing-impaired listeners

Klaus Friedrichs, Nadja Bauer,
Rainer Martin, Claus Weihs

Nr. 65/2016

Discussion Paper

A Computational Study of Auditory Models in Music Recognition Tasks for Normal-Hearing and Hearing-Impaired Listeners

Klaus Friedrichs[*], Nadja Bauer[*], Rainer Martin[**], and Claus Weihs[*]

[*]Department of Statistics, TU Dortmund University
[**]Institute of Communication Acoustics, Ruhr-Universität Bochum

November 9, 2016

Abstract

The utility of auditory models for solving three music recognition tasks – onset detection, pitch estimation and instrument recognition – is analyzed. Appropriate features are introduced which enable the use of supervised classification. The auditory model-based approaches are tested in a comprehensive study and compared to state-of-the-art methods, which usually do not employ an auditory model. For this study, music data is selected according to an experimental design, which enables statements about performance differences with respect to specific music characteristics. The results confirm that the performance of music classification using the auditory model is at least comparable to the traditional methods. Furthermore, the auditory model is modified to exemplify the decrease of recognition rates in the presence of hearing deficits. The resulting system is a basis for estimating the intelligibility of music which in the future might be used for the automatic assessment of hearing instruments.

Keywords: music recognition, classification, onset detection, pitch estimation, instrument recognition, auditory model, music intelligibility, hearing

impairment

1 Introduction

Hearing-impaired listeners like to enjoy music as well as normal-hearing listeners although this is aggrieved by a distorted frequency resolution. Recently, several listening experiments have been conducted to assess the impact of hearing loss on music perception for hearing-impaired listeners (e.g., [1, 2, 3, 4]). For many applications like hearing instrument optimization it is desirable to measure this impact automatically by the use of a simulation model. Therefore, we investigate the potential of emulating certain normal-hearing and hearing-impaired listeners by automatically assessing their ability to discriminate music attributes via an auditory model in this study. auditory models are computational models which mimic the perception of the human auditory process by transforming acoustic signals into neural activity of several simulated auditory nerve fibers (channels). Since these models do not explain the whole listening comprehension of higher central auditory stages, a back end is needed relying on the output of the auditory periphery. Similar ideas have already been proposed for measuring speech intelligibility in [5] and [6] where this back end is an automatic speech recognition system, resulting in the word-recognition rate as a natural metric. However, no such straightforward method exists to measure the corresponding "music intelligibility" in general. Unlike speech, music spectra are highly variable and it peaks tend to be sharper. Additionally, typical musical inputs have a much greater dynamic range [7]. For estimating "music intelligibility" its constituent elements (pitch, harmony, rhythm and timbre) have to be assessed in an independent manner [8]. Therefore, we focus on the three separate music recognition tasks onset detection, pitch estimation and instrument recognition. Contrary to state-of-the-art methods, here we extract information from the auditory output only. In fact, some recent proposals in the field of speech recognition and music data analysis use auditory models, thus capitalizing on the superiority of human perception (e.g., [9, 10, 11]). However, in most of these proposals, the applied auditory model is not sufficiently detailed to provide adequate options for implementing realistic hearing deficits. In the last decades auditory models have been developed which are more sophisticated and meanwhile can simulate hearing deficits [12, 13, 14]. In [15] and

[16], it is shown that simple parameter modifications in the auditory model are sufficient to realistically emulate auditory profiles of hearing-impaired listeners.

In this study, we restrict our investigation on chamber music which includes a predominant melody instrument and one or more accompanying instruments. For further simplification, we are only interested in the melody track which means that all accompanying instruments are regarded as interferences. This actually means that the three recognition tasks are described more precisely as predominant onset detection, predominant pitch estimation and predominant instrument recognition.

The article is organized as follows: In Section 2 related work is discussed. The contribution of this paper is summarized in Section 3. In Section 4, the applied auditory model of Meddis (Section 4.1) and our proposals for the three investigated music recognition tasks are described (Sections 4.2 - 4.4). At the end of that section, the applied classification methods – Random Forest (RF) and linear SVM – are briefly explained (Section 4.5). Section 5 provides details about the experimental design. Plackett-Burman (PB) Designs are specified for selecting the data set, which enable assessments about performance differences w. r. t. the type of music. In Section 6, we present the experimental results. First, the proposed approaches are compared to state-of-the-art methods, and second, performance losses of the hearing-impaired emulators are investigated. Finally, Section 7 summarizes and concludes the paper, and gives some suggestions for future research.

2 Related Work

Combining predominant onset detection and predominant pitch estimation results in a task which is better known as melody detection. However, the performance of approaches in that research field are rather poor to date compared to human perception [17]. In particular, onset detection is still rather error-prone for polyphonic music [18]. Hence, in this study all three musical attributes of interest are estimated separately, which means the true onsets (and offsets) are assumed to be known for pitch estimation and instrument recognition, excluding error propagation from onset detection.

2.1 Onset Detection

The majority of onset detection algorithms consists of an optional pre-processing, a reduction function (called onset detection function), which is derived at a lower sampling rate, and a peak-picking algorithm [19]. They all can be summarized into one algorithm with several parameters to optimize. In [20], we systematically solve this by using sequential model-based optimization. The onset detection algorithm can also be applied channel-wise to the output of the auditory model. Here, the additional challenge lies in the combination of different onset predictions of several channels. In [21], a filter bank is used for pre-processing, and for each band, onsets are estimated which together build a set of onset candidates. Afterwards, a loudness value is assigned to each candidate and a global threshold and a minimum distance between two consecutive onsets are used to sort out candidates. A similar approach, but this time for combining the estimates of different onset detection functions, is proposed in [22] where the individual estimation vectors are combined via summing and smoothing. Instead of combining the individual estimations at the end, in [23] we propose a quantile-based aggregation before peak-picking. However, the drawback of this approach is that the latency of the detection process varies for the different channels, which is difficult to compensate before peak-picking. The predominant variant of onset detection is a task which to our best knowledge has not been investigated, yet.

2.2 Pitch Estimation

Most pitch estimation algorithms are either based on the autocorrelation function (ACF) or they work in the frequency domain by applying a spectral analysis of potential fundamental frequencies and their corresponding partials. For both approaches, one big challenge is to pick the correct peak which is particularly difficult for polyphonic music where the detection is disturbed by overlapping partials. In order to solve that issue, several extensions to the autocorrelation approach are implemented in the popular YIN algorithm [24] which in fact uses the difference function instead of the ACF. A further extension is the pYIN method which is introduced in [25]. It is a two-stage method which takes past and future estimations into account. First, for every frame several fundamental frequency candidates are predicted, and second, the most convenient temporal path is estimated, according to a hidden Markov model. In [26], a maximum-likelihood approach is introduced in the frequency domain. Another alternative is a statistical classification approach which is proposed in [27].

For pitch estimation, also a few approaches using an auditory model – or at least some of its components – have been introduced. In [11], an outer/middle ear filter is proposed for pre-processing which reduces the number of octave errors. A complete auditory model is applied in [28] and [29]. In those studies, an autocorrelation method is proposed where the individual running ACF's of each channel are combined by summation (averaging) across all channels (SACF). The results of that approach are equivalent to human perception for some specific sounds. However, the approach is not tested for complex music signals, yet. Also here, the challenge of picking the correct peak remains. All previously discussed approaches are originally designed for monophonic pitch detection. However, pitch estimation can be extended to its predominant variant by identifying the most dominant pitch, which many peak-picking methods implicitly calculate.

Also for polyphonic pitch estimation approaches exist. One approach is proposed in [10]. Instead of just picking the maximum peak of the SACF, the strength of each candidate (peak) is calculated as a weighted sum of the amplitudes of its harmonic partials. Another approach is introduced in [30], where the EM-algorithm is used to estimate the relative dominance of every possible harmonic structure.

2.3 Instrument Recognition

The goal of instrument recognition is the automatic distinction of music instruments playing in a given music piece. Different music instruments have different compositions of partial tones, e.g., in the sound of a clarinet mostly odd partials occur. This composition of partials is, however, also dependent on other factors like the pitch, the room acoustic and the performer [31]. For building a classifier the meaningful information of each observation has to be extracted, which is achieved by appropriate features. Timbral features based on the one-dimensional acoustic waveform are the most common features for instrument recognition, yet. However, features based on an auditory model have already been introduced in [32]. Also, biomimetic spectro-temporal features, requiring a model of higher central auditory stages, have been successfully investigated for solo music recordings in [33]. Predominant instrument recognition can be solved similarly to the monophonic variant, but is much harder due to the additional "noise" from the accompanying instruments [34]. An alternative is starting with sound source separation in order to apply monophonic instrument recognition afterwards [35]. Naturally this concept fails if the sources are not separated well, a task which itself is still a challenge.

3 Contribution of the Paper

In this study, we use the comprehensive and well established auditory model of Meddis [36], and its hearing-impaired variants [16]. For onset detection, we adapt the ideas of [21] and [22] to develop a method for combining onset estimations of different channels which can handle asynchronous estimations and which is also suitable for music with dynamics. Furthermore, we propose parameter optimization to adapt the method to predominant onset detection. Sequential model-based optimization (MBO) is applied to find optimal parameter settings for three considered variants of onset detection: (1) monophonic, (2) polyphonic and (3) predominant onset detection. For pitch estimation, inspired by [27], we propose a classification approach for peak-picking, where each channel nominates one candidate. Our approach is applicable to temporal autocorrelations as well as in the frequency domain. Additionally, we test the SACF-method, where we investigate two variants for peak-picking. For instrument recognition, we adapt common timbral features for instrument recognition by extracting them channel-wise – contrary to [32], where the features are defined across all channels – from the auditory output. This channel-wise approach preserves more information, can be more easily adapted to the hearing-impaired variants and enables assessments about the impact of specific channels to the recognition rates.

All approaches are extensively investigated using a comprehensive experimental design. The capability of auditory models to discriminate the three considered music attributes is shown via the normal-hearing-auditory model which is compared to the state-of-the-art methods. In our experiments, the approaches using the auditory model-output for pitch estimation and instrument recognition even perform distinctly better than the common approaches. As a prospect of future research, performance losses based on hearing deficits are exemplified using three so-called hearing-dummies introduced in [16].

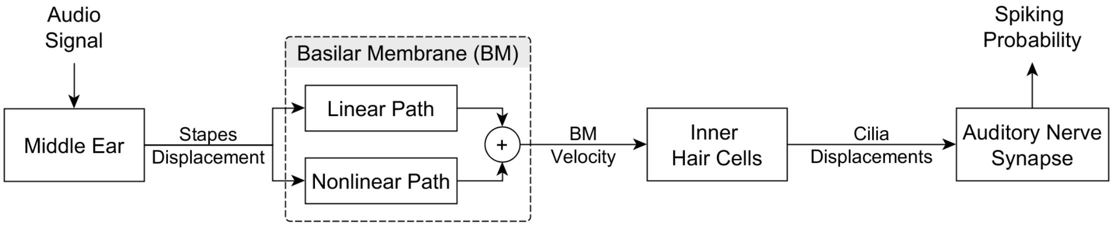

Figure 1: Block diagram of Meddis' model of the auditory periphery.

4 Music Classification using Auditory Models

4.1 Auditory Models

The auditory system of humans and other mammals consists of several stages located in the ear and the brain. While the higher stages located in the brain are difficult to model, the auditory periphery is much better investigated. This stage models the transformation from acoustical pressure waves in the air to release events to the auditory nerve fibers. Out of the several models simulating the auditory periphery, we apply the popular and widely analyzed model of Meddis [36].

The auditory periphery consists of the outer ear, the middle ear and the inner ear. The main task of the outer ear is collecting sound waves and directing them further into the ear. At the back end of the outer ear the ear-drum vibrates. This vibration is transmitted to the stapes in the middle ear and then directed further to the cochlea in the inner ear. Inside the cochlea, the basilar membrane vibrates at specific locations dependent on the stimulating frequencies. On the basilar membrane inner hair cells are located which are activated by the velocity of the membrane and evoke spike emissions (neuronal activity) of the auditory nerve fibers.

The auditory model of Meddis [36] is a cascade of several consecutive modules, which emulate the spike firing process of multiple auditory nerve fibers. A block diagram of this model can be seen in Figure 1. Since auditory models use filter banks, the simulated nerve fibers are also called channels within the simulation. Each channel corresponds to a specific point on the basilar membrane. In the standard setting of the Meddis model, 41 channels are examined. As in the human auditory system, each channel has an individual best frequency (center frequency) which defines the frequency that evokes maximum excitation. The best frequencies are equally spaced on a log scale with 100 Hz for the first and 6000 Hz for the 41th channel.

In the last plot of Figure 2, an exemplary output of the model can be seen. The 41 channels are located on the vertical axis according to their best frequencies, and the gray-scale indicates the probability of spike emissions (white means high probability). The acoustic stimulus of this example is a harmonic tone which is shown in the first plot of the figure. The first module of Meddis' model is the middle ear where sound waves are converted into stapes displacement. The resulting output of the sound example is shown in the second plot. The second module emulates the basilar membrane where stapes displacement is transformed into the velocity of the basilar membrane at different locations, implemented by a dual-resonance-non-linear (DRNL) filter bank, a bank of overlapping filters [37]. The DRNL filter bank consists of two asymmetric bandpass filters which are processed in parallel: one linear path and one nonlinear path. The output of the basilar membrane for our sound example can be seen in the third plot of the figure. Next, time dependent basilar membrane velocities are transformed into time dependent Inner Hair Cells cilia displacements. Afterwards these displacements are transformed by a calcium-controlled transmitter release function into spike probabilities $p(t, k)$, the final output of the considered model, where t is the time, and k is the channel number.

For the auditory model with hearing loss we consider three examples, called hearing dummies, which are described in [15] and [16]. These are modified versions of the Meddis auditory model. The goal of hearing dummies is to mimic the perception of real hearing impairments. In future, they might be used to evaluate psychological inspired hearing-aids [38]. In the original proposal, channels with best frequencies between 250 Hz and 8 kHz are examined, whereas in the normal-hearing model described above this range is between 100 Hz and 6 kHz. Note that this difference is not influenced by any hearing damage, it is just a matter of design perspective. For a better comparison, the same best frequencies have to be taken into account for all models. Since the range between 100 Hz and 6 kHz seems to be more suitable to music, we adjust the three hearing dummies accordingly.

The first hearing dummy simulates a strong mid- and high-frequency hearing loss. In the original model, this is implemented by retaining the channel with the best frequency of 250 Hz only and by disabling the nonlinear

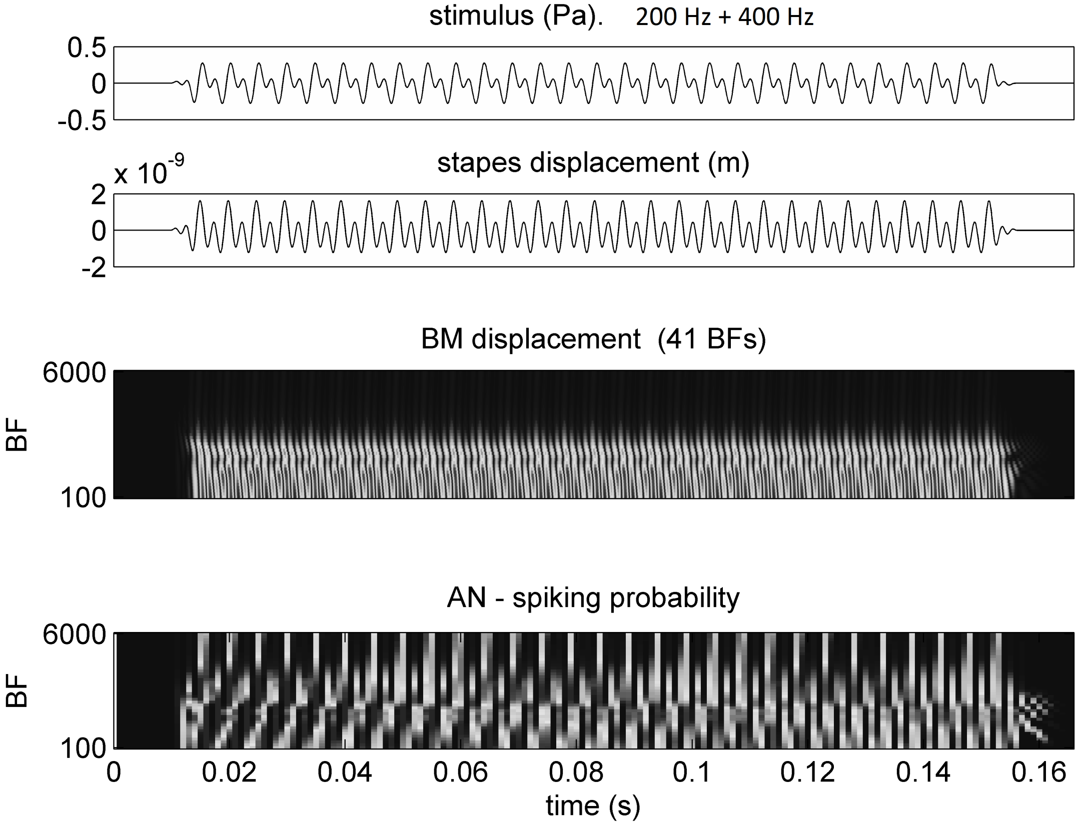

Figure 2: Exemplary output of Meddis' model of the auditory periphery: (1) original signal (200 Hz + 400 Hz), (2) middle ear output (stapes displacement), (3) basilar membrane (BM) output with respect to the channels' best frequencies (BF), (4) auditory nerve (AN) output with respect to the BFs.

Table 1: Parameterization of the three considered hearing dummies and the normal-hearing model.

	Remaining Channels	Nonlinear Path
Normal Hearing	1 - 41	yes
Hearing Dummy 1	1 - 10	no
Hearing Dummy 2	1 - 16 and 33 - 41	yes
Hearing Dummy 3	1 - 29	yes

path. In our modified version of that dummy, the first ten channels are retained – all of them having best frequencies lower than or equal to 250 Hz – and the nonlinear path is disabled for all of them. The second hearing dummy simulates a mid-frequency hearing loss indicating a clear dysfunction in a frequency region between 1 and 2 kHz. Therefore, we disable 16 channels (channels 17 to 32) for the modified version of the hearing dummy. The third hearing dummy is a steep high-frequency loss, which is implemented by disabling all channels with best frequencies above 1750 Hz corresponding to the last 12 channels in the model. The parameterization of the three hearing dummies is summarized in Table 1.

4.2 Onset Detection

The task of onset detection is to identify all time points where a new tone begins. For predominant onset detection, just the onsets of the melody track are of interest. First, we define the baseline algorithm which operates on the acoustic waveform x and which we use for comparison reasons. However, this algorithm can also be adapted to the auditory model output in a channel-wise manner. Second, we describe the performed parameter tuning which we apply to optimize onset detection. Last, we introduce our approaches using the auditory model by aggregating the channel-wise estimations.

4.2.1 Baseline Onset Detection Approach

The baseline onset detection approach we use in our study consists of seven steps illustrated in Figure 3. The corresponding parameters, we want to optimize, are shown in parentheses.

In the first step, the ongoing signal is split into small windows with a window size of M samples and a hop size h which is the distance in samples

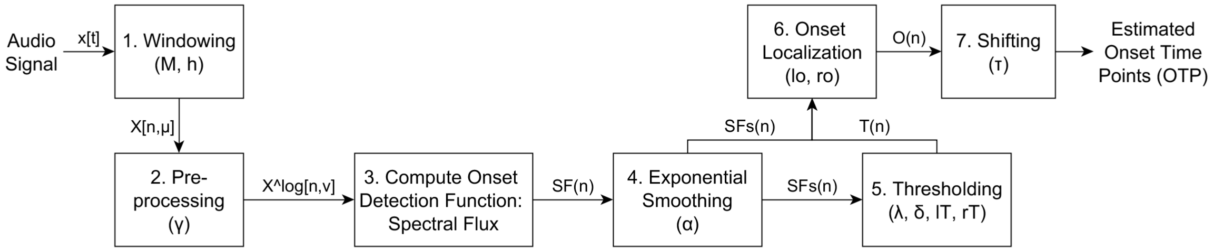
Audio Signal
x[t]
1. Windowing (M, h)
X[n,μ]
2. Pre-processing (γ)
X^log[n,ν]
3. Compute Onset Detection Function: Spectral Flux
SF(n)
4. Exponential Smoothing (α)
SFs(n)
5. Thresholding (λ, δ, lT, rT)
SFs(n)
T(n)
6. Onset Localization (lo, ro)
O(n)
7. Shifting (τ)
Estimated Onset Time Points (OTP)

Figure 3: Block diagram for classical onset detection (without auditory model).

between the starting points of subsequent windows. For each window the magnitude spectrum of the discrete Fourier transform (DFT) $|X[n, \mu]|$ is computed where n denotes the window index and μ the frequency bin index. Afterwards, two preprocessing steps are applied (step 2). First, a filter-bank $F[\mu, \nu]$ filters the magnitude spectrum according to the note scale of western music [39]. The filtered spectrum is given by

$$X_{filt}[n, \nu] = \sum_{\mu=1}^{M} |X[n, \mu]| \cdot F[\mu, \nu], \tag{1}$$

where ν is the bin index of this scale which consists of $B = 82$ frequency bins (12 per octave), spaced in semitones for the frequency range from 27.5 Hz to 16 kHz. Second, the logarithmic magnitude of the spectrum is computed:

$$X^{log}[n, \nu] = \log(\gamma \cdot X_{filt}[n, \nu] + 1), \tag{2}$$

where $\gamma \in]0, 20]$ is a compression parameter to optimize.

Afterwards, a feature is computed in each window (step 3). Here we use the Spectral Flux ($SF(n)$) feature, which is the best feature for onset detection w. r. t. the F-Measure according to recent studies. In [40], this is shown on a music data set with 1,065 onsets covering a variety of musical styles and instrumentations, and in [39], this is verified on an even larger data set with 25,966 onsets. Spectral Flux describes the degree of positive spectral changes between consecutive windows and is defined as:

$$\begin{aligned} SF(n) = \sum_{\nu=1}^{B} H(X^{log}[n, \nu] - X^{log}[n-1, \nu]) \\ \text{with } H(x) = (x + |x|)/2, \end{aligned} \tag{3}$$

Joining the feature values over all windows consecutively yields the SF vector.

Next, exponential smoothing (step 4) is applied, defined by

$$\begin{aligned} &SF_{\mathrm{s}}(1) = SF(1) \text{ and} \\ &\qquad SF_{\mathrm{s}}(n) = \alpha \cdot SF(n) + (1-\alpha) \cdot SF_{\mathrm{s}}(n-1) \\ &\qquad\qquad \text{for } n = 2, \ldots, L, \end{aligned} \tag{4}$$

where L is the number of windows and $\alpha \in [0, 1]$.

A threshold function (step 5) distinguishes between relevant and non-relevant maxima. To enable reactions to dynamic changes in the signal, a moving threshold is applied, which consists of a constant part δ and a local part weighted by λ [40]. The threshold function is defined as

$$\begin{aligned} T(n) = \delta + \lambda \cdot mean(SF_{\mathrm{s}}(n - l_T), \ldots, SF_{\mathrm{s}}(n + r_T)), \\ \text{for } n = 1, \ldots, L, \end{aligned} \tag{5}$$

where l_T and r_T are the numbers of windows to the left and to the right, respectively, defining the subset of considered windows.

The localized tone onsets are selected by two conditions (step 6):

$$O(n) = \begin{cases} 1, & \text{if } SF_{\mathrm{s}}(n) > T(n) \text{ and } SF_{\mathrm{s}}(n) = \\ & \max(SF_{\mathrm{s}}(n - l_O), \ldots, SF_{\mathrm{s}}(n + r_O)) \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

$\boldsymbol{O} = (O(1), \ldots, O(L))^T$ is the tone onset vector and l_O and r_O are additional parameters, namely the number of windows to the left and right of the actual window.

Windows with $O(n) = 1$ are converted into time points by identifying their beginnings (in seconds). Finally, all estimated onset time points are shifted by a small time constant τ (step 7) to account for the latency of the detection process. Compared to the physical onset, which is the target in our experiments, the perceptual onset is delayed, affected by the rise times of instrument sounds [41]. In the same manner, these rise times also affect the maximum value of spectral flux and other features. $\boldsymbol{OTP} = (OTP_1, \ldots, OTP_{C_{est}})$ denotes the resulting vector of these final estimates, where C_{est} is the number of estimated onsets. A found tone onset is correctly identified if it is inside a tolerance interval around the true onset. We use ±25 ms as the tolerance.

The performance of tone onset detection is measured by the F-measure taking into account the tolerance regions:

$$F = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}, \quad F \in [0, 1], \tag{7}$$

where TP is the number of correctly detected onsets, FP is the number of false alarms, and FN is the number of missed onsets. $F = 1$ represents an optimal detection, whereas $F = 0$ means that no onset is detected correctly. Apart from these extremes, the F-measure is difficult to interpret. Therefore, we exemplify the dependency of the number of missed onsets on the F-value and the number of true onsets $C_{true} = TP + FN$ for the scenario where no false alarm is produced:

$$FP = 0 \quad \Longrightarrow \quad FN = (1 - \frac{F}{2 - F}) \cdot C_{true}. \tag{8}$$

4.2.2 Parameter Optimization

The baseline onset detection algorithm contains the 11 parameters summarized in Table 2. Parameter optimization is needed to find the best parameter setting w. r. t. a training data set and to adapt the algorithm to predominant onset detection and to the auditory model output. Since evaluation of one parameter setting – also called point in the following – is time consuming (several minutes on the used Linux-HPC cluster system [42]), we apply sequential model-based optimization (MBO). After an initial phase, i.e., an evaluation of some randomly chosen starting points, new points are proposed and evaluated iteratively w. r. t. a surrogate model fitted to all previous evaluations, and an appropriate infill criterion decides which point is the most promising. The most prominent infill criterion is expected improvement (EI) which looks for a compromise of surrogate model uncertainty in one point and its expected function value. For a more detailed description of MBO see [43] and [44].

4.2.3 Onset Detection using an auditory model

The baseline onset detection algorithm can also be performed on the output of each channel of the auditory model ($p(t, k)$). Again, we use MBO to optimize the algorithm on the data, this time individually for each channel k, getting the estimation vector $\boldsymbol{OTP}_k$. Now, the additional challenge arises how to

Table 2: Parameters and their ranges of interest for the classical onset detection approach.

Parameter Name	Minimum Value	Maximum Value
window size M	2^{10}	2^{12}
hop size h	400	1600
γ	0.01	20
α	0	1
λ	0	1
δ	0	10
l_T	0 s	0.5 s
r_T	0 s	0.5 s
l_O	0 s	0.25 s
r_O	0 s	0.25 s
τ	-0.025 s	0.025 s

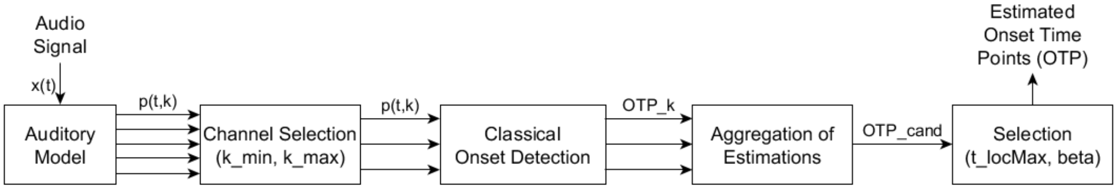
Audio Signal
x(t)
Auditory Model
p(t,k)
Channel Selection (k_min, k_max)
p(t,k)
Classical Onset Detection
OTP_k
Aggregation of Estimations
OTP_cand
Selection (t_locMax, beta)
Estimated Onset Time Points (OTP)

Figure 4: Block diagram for the proposed approach for onset detection using an auditory model.

combine different onset predictions of several channels. We compare two approaches. First, as a simple variant, we just consider the channel which performs best during the training phase. Second, we introduce a variant which combines the final results of all channels. This approach is illustrated in Figure 4. Again, the corresponding parameters, we want to optimize, are shown in parentheses.

Since particularly the performance of the highest channels are rather poor as we will see in Section 6, and furthermore, considering fewer channels leads to a reduction of computation time, we allow to omit the lowest and the highest channels by defining the minimum k_{min} and the maximum channel k_{max} to consider. All estimated onset time points of the remaining channels

Table 3: Parameters and their ranges of interest for the aggregation approach (onset detection with auditory model).

Parameter Name	Minimum Value	Maximum Value
t_{lM}	0 s	0.125 s
β	0	1
k_{min}	1	20
k_{max}	21	41

are pooled into one set of onset candidates:

$$\boldsymbol{OTP}_{cand} = \bigcup_{k=k_{min}}^{k_{max}} \boldsymbol{OTP}_k. \tag{9}$$

Obviously, in this set many estimated onsets occur several times, probably with small displacements, which have to be combined to a single estimation. Additionally, estimations which just occur in few channels might be wrong and should be deleted. Hence, we develop the following method to sort out candidates. For each estimation we count the number of estimations in their temporal neighborhood, defined by an interval of ±25 ms (corresponding to the tolerance of the F-measure). In a next step only estimations remain where this count is a local maximum and above a global threshold. The threshold is defined by

$$\beta \cdot (k_{max} - k_{min} + 1), \tag{10}$$

where β is a parameter to optimize. For each candidate time point n, the interval within which it must fulfill the maximum condition is set to $[n - t_{loc}, \ldots, n + t_{loc}]$, where t_{loc} is another parameter to optimize.

This results in four free parameters which we optimize in a second MBO run. The ranges of interest for these parameters are listed in Table 3. Since optimizing just four parameters is much faster than optimizing the eleven parameters of the conventional method, the overhead of computation time can be ignored.

The adaption to predominant onset detection using the auditory model output is again just performed by searching the best parameter setting with respect to the reduced target time points (not including the onset time points of the accompaniment).

4.3 Predominant Pitch Estimation

Here, we understand pitch estimation as a synonym for fundamental frequency (F_0) estimation, where we allow a tolerance of $1/2$ semitone. This is equivalent to a relative error of approximately 3% in the frequency scale (Hz). In the predominant variant, we are just interested in the pitch of the melody instrument. As already mentioned above, we assume to know the onsets and offsets of each melody tone. This information is used to separate the auditory output of each song temporally into individual melody tones (including the accompaniment at this time).

Our tested approaches using the auditory model can be divided into two groups – autocorrelation approach and spectral approach – which are described in the following. Additionally, we use the YIN algorithm [24], which works without an auditory model, for comparison reasons in our experiments.

4.3.1 Autocorrelation Approach

One challenge of autocorrelation analysis of the auditory output is again the combination of several channels. In [28] and [29], this is achieved by first computing the individual running autocorrelation function (ACF) of each channel and combining them by summation (averaging) across all channels (SACF). The SACF is defined by

$$s(t, l) = \frac{1}{K} \sum_{k=1}^{K} h(t, l, k), \tag{11}$$

where K is the number of considered channels and $h(t, l, k)$ is the running ACF of each channel k at time t and lag l. The peaks of the SACF are indicators for the pitch where the maximum peak is a promising indicator for the fundamental frequency. The model is successfully tested for several psychophysical phenomena like pitch detection with missing fundamental frequency [28, 29]. However, for complex musical tones, often the maximum peak of the SACF is not located at the fundamental frequency, but instead at one of its multiples. Hence, we propose an improved peak picking version which takes the first peak of the SACF which is above an optimized threshold:

$$\min[t \in t_{\text{lM}} \colon SACF(t) > \lambda \cdot \max(SACF(t))], \tag{12}$$

where t_{lM} is the set of all local maxima of the SACF and $\lambda \in [0, 1]$ has to be optimized on a training set.

4.3.2 Spectral Approach

We propose a classification method partly based on features which we introduced in [45] and [46] for detecting the frequencies of all partials. Here, the feature set is adapted for pitch estimation and some additional features are added. At first, the DFT magnitude spectrum $|P[\mu, k]|$ of each channel k is computed where each maximum peak within an interval around the channel's best frequency – limited by the best frequencies of the two neighboring channels – is considered as the channel's pitch candidate:

$$\mu^*[k] = \underset{\mu \in \{BF[k-1],\ldots,BF[k+1]\}}{\arg\max} |P[\mu, k]|, \quad k = 1, \ldots, K, \tag{13}$$

where $BF[k]$ is the frequency bin which comprises the best frequency of channel k (for $k = 1, \ldots, K$), which is between 100 Hz for the first and 6 kHz for the last channel. For the limits of the first and the last channel, we additionally define $BF[0]$ as the frequency bin which comprises 50 Hz and $BF[K+1]$ as the frequency bin which comprises 10 kHz. The center frequency $CF(\mu)$ of the frequency bin $\mu^*[k]$ is the candidate $c[k] = CF(\mu^*[k])$.

The classification target is to identify the channel with minimal distance between its best frequency and the fundamental frequency. The frequency candidate of this channel is returned as the estimated pitch. The following features are computed individually for each channel respectively each candidate:

- The frequency of the candidate $c[k]$,
- The spectral amplitude of the candidate's frequency bin: $a_c[k] = |P[\mu^*[k], k]|$,
- The bandwidth $b[k]$ of the candidate, defined by the distance between the two closest frequency bins to the left and right of the candidate, where the spectral amplitude is below 10% of the candidate's amplitude (see also Figure 5):

$$b[k] = CF(\mu^*_{right}[k]) - CF(\mu^*_{left}[k]), \tag{14}$$

where the band edges are defined by

$$\mu^*_{right}[k] \;=\; \min(\mu \;\in\; \{\mu^*[k], ..., {}^{M}\!/_{2}\} \colon {}^{a_c[k]}\!/_{10} \;>\; |P[\mu, k]|), \tag{15}$$

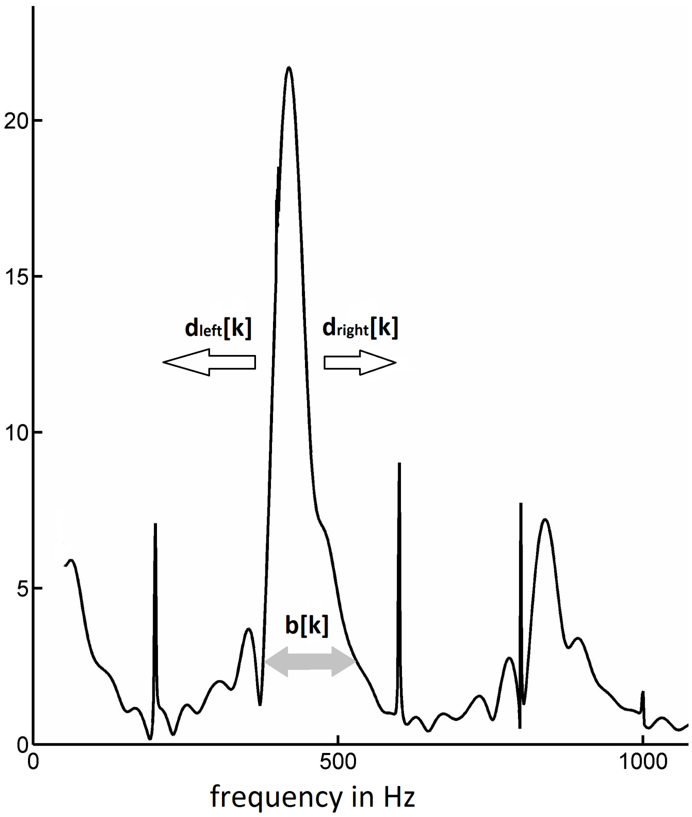

Figure 5: Features for pitch estimation: a) bandwidth $b[k]$ of the candidate peak, b) distance to maximum left $d_{left}[k]$ and c) distance to maximum right $d_{right}[k]$.

where $\mu^*_{right}[k]$ is set to $M/2$, if no such μ exists, and

$$\mu^*_{left}[k] \quad = \quad \max(\mu \quad \in \quad \{1, ..., \mu^*[k]\} : \, a_c[k]/10 \quad > \quad |P[\mu, k]|), \quad (16)$$

where $\mu^*_{left}[k]$ is set to 0, if no such μ exists,

- The distances of the candidate's frequency to the maxima to the left and right, respectively, restricted by the candidates band edges (two features: $d_{left}[k]$ and $d_{right}[k]$, see also Figure 5):

$$d_{left}[k] = c[k] - CF(max_{left}[k]), \text{ where} \quad max_{left}[k] = \underset{\mu \in \{1 \ldots \mu^*_{left}[k]\}}{\arg\max} (P[\mu, k]) \text{ and} \tag{17}$$

$$d_{right}[k] = CF(max_{right}[k]) - c[k], \text{ where} \quad max_{right}[k] = \underset{\mu \in \{\mu^*_{right}[k] \ldots M/2\}}{\arg\max} (P[\mu, k]), \tag{18}$$

- The spectral amplitude of these two maxima (2 features): $|P[max_{left}[k]]|$ and $|P[max_{right}[k]]|$.
- Average and maximum spike probabilities of the channel: $p_{mean}[k]$ and $p_{max}[k]$,
- Average and maximum spectral magnitude of the first nine partials ($pl = 1, \ldots, 9$) across all channels:

$$P^{mean}_{pl}[k] = \frac{1}{K} \sum_{n=1}^{K} P[a(pl \cdot c[k]), n], \tag{19}$$

where $a(i)$ is the frequency bin which comprises frequency i, and

$$P^{max}_{pl}[k]) = \max_{n \in \{1, \ldots, K\}} (P[a(pl \cdot c[k]), n],), \tag{20}$$

- In the same manner, average and maximum spectral magnitude of the first undertone (half frequency of the candidate) across all channels: $P^{mean}_{1/2}[k]$ and $P^{max}_{1/2}[k]$.

Altogether this results in 29 features for each channel, i. e. $29 \cdot 41 = 1189$ features for the auditory model.

As a third method for pitch estimation, this classification approach is also applied in the same way to the ACF. Here, the same 29 features are extracted, but this time based on the ACF instead of the DFT.

Table 4: Features for instrument recognition

feature no.	feature name
1	Root-Mean-Square Energy
2	lowenergy
3	mean spectral flux (see equation 3)
4	standard deviation of spectral flux
5	spectral rolloff
6	spectral brightness
7	irregularity
8 - 20	Mel-Frequency Cepstral Coefficients (mfcc): first 13 coefficients
21	entropy

4.4 Predominant Instrument Recognition

Although one could assume the same predominant instrument during one song we do not use the information about previous tones, since we want to use instrument recognition as an indicator for correctly perceived timbre. We think this is best characterized by tone-wise classification without using additional knowledge. Hence, also here the auditory output of each song is separated into temporal segments defined by the individual tones of the predominant instrument, and for each segment – corresponding to one melody tone – features are extracted separately.

We use 21 features, listed in Table 4, which we already considered in previous studies [47] and which are common for instrument recognition based directly on the time domain waveform. For our approach using an auditory model, they are computed on each of the 41 channels, thus getting $41 \cdot 21 = 861$ features for each tone. The first 20 features are computed by means of the *MIRtoolbox* [48]. The last feature is the Shannon-Entropy:

$$H(X) = -\sum_{\mu=1}^{M} pr(|X[\mu]|) \log_2 pr(|X[\mu]|), \qquad (21)$$

where $X[\mu]$ is the DFT of the time signal (respectively the DFT of a channel output in the auditory model variant) and $pr(|X[\mu]|) = {}^{|X[\mu]|}/_{\sum_{\nu=1}^{M}|X[\nu]|}$ is the share of the μth frequency bin with respect to the cumulated spectral magnitudes of all bins. $H(X)$ measures the degree of spectral dispersion of an acoustic signal and is taken as a measure for tone complexity.

4.5 Classification Methods

Supervised classification is required for our approaches in pitch estimation and instrument recognition. Formally, a classifier is a map $f : \Phi \rightarrow \Psi$, where Φ is the input space containing characteristics of the entities to classify, and Ψ is the set of categories or classes. Here, Φ is a (reduced) set of features and Ψ is a set of labels of musical instruments or channels (pitch candidates).

In our experiment we apply two important classes of methods, namely linear large margin methods (represented by the linear Support Vector Machine, SVM) and ensembles of decision trees (Random Forests, RF).

4.5.1 Decision Trees and Random Forests

Decision trees are one of the most intuitive models used in classification. The model is represented as a set of hierarchical "decision rules", organized usually in a binary tree structure. When a new observation needs to be classified, it is propagated down the tree taking either the left or right branch in each decision node of the tree, depending on the decision rule of the current node and the corresponding feature value. Once a terminal node has been reached, a class label is assigned. For a more detailed description of decision trees see [49].

Sometimes, a single classification rule is not powerful enough to sufficiently predict classes of new data. Then, one idea is to combine several rules to improve prediction. This leads to so-called ensemble methods. One example is Random Forests (RF), a combination of many decision trees (see, e.g., [50]). The construction of the different classification trees has random components - i.e., for each tree only a random subset of observations and for each decision node only a random subset of features is considered -, leading to the term Random Forests.

4.5.2 Support Vector Machines

Support Vector Machines (SVMs) [51] are among the state-of-the-art machine learning methods for linear and non-linear classification. They are often among the strongest available predictors, and they come with extensive theoretical guarantees. To simplify our experimental design, we consider only linear SVMs.

The linear SVM separates two classes indicated by labels $\psi \in \{-1, +1\}$ with an affine function $f(\vec{\phi}) = \vec{w}^T \vec{\phi} + b$, given by a weight vector $\vec{w} \in \mathbb{R}^p$ and a bias or offset term $b \in \mathbb{R}$. An input $\vec{\phi}$ is classified according to $\text{sign}(f(\vec{\phi}))$. The SVM classifier is defined as the (affine) linear function f that maximizes a safety margin between the classes. As we cannot exclude the existence of outliers, slack variables ξ_i are applied, one per training point, measuring the amount of constraint (or margin) violation:

$$\min_{\vec{w},b} \quad \frac{1}{2}\|\vec{w}\|^2 + C \cdot \sum_{i=1}^{n} \xi_i \qquad \text{s.t. } \psi_i \cdot (\vec{w}^T \vec{\phi}_i + b) \geq 1 - \xi_i \text{ and } \xi_i \geq 0. \quad (22)$$

The solution $(\vec{w}^*, b^*)$ of this problem is defined as the (standard) linear SVM. It has a single parameter, $C > 0$, trading maximization of the margin against minimization of margin violations.

Many practical problems – like our music recognition tasks – involve three or more classes $(G > 2)$. Therefore the large margin principle has been extended to multiple classes. We apply the one-versus-one approach, where the G-class problem is converted into ${}^{G(G-1)}/_{2}$ binary problems. For each pair of classes, a SVM decision function is trained for separating the two specific classes. The prediction rule then picks the class which is voted the most.

4.5.3 Feature Selection

Feature selection filters the important features in order to reduce computation time for feature extraction as well as for the classification process itself. Another advantage of feature selection is a better interpretability of a classification model based on lesser features. Knowing which features are important might also help to design improved feature sets. Lastly, feature selection can even improve classification results since classifiers have problems with meaningless or redundant features.

Two basic approaches exist for feature selection: forward selection and backward selection [52]. Forward selection is a greedy search approach which starts with an empty set of features. In each iteration the feature which yields the most improvement w. r. t. the error rate is added to the set until no feature yields an improvement higher than a specified threshold. Backward selection works the other way round. It starts with all features and in each iteration the feature is removed which yields the least improvement. Here, the stopping threshold is usually a small negative value allowing also small decreases of the error rate in order to simplify the model.

Both approaches have a complexity of $O(n^2)$ which results in too much computation time when dealing with $n \approx 1000$ features as we consider for pitch estimation and instrument recognition. Hence, we propose to group the features into feature groups and to handle each group as one single feature for forward and backward selection, respectively. There are two natural grouping mechanisms since the features can be categorized by two dimensions: the channel index and the feature name. The first approach is to combine the related features across all channels into one group and the second approach is to combine all features generated in the same channel into one group. The first approach results in 29 feature groups for pitch estimation and 21 groups for instrument recognition. For both tasks, the second approach results in K feature groups. An additional benefit of channel-based grouping is the potential of sorting out entire channels which also reduces computation time for the simulated auditory process. In our experiments, we set the minimum improvement for forwards selection to 0.01 and for backward selection to -0.001.

5 Design of Experiments

5.1 Data

Our data base consists of 100 chamber music pieces recorded in MIDI which include a specific melody instrument and one or more accompanying instruments, either piano or strings. The ISP toolbox in Matlab with the "Fluid (R3) General MIDI SoundFont“ is applied for synthesizing MIDI files in a sample based way [53]. For simplification reasons, only standard playing styles are considered, e.g., bowed for cello. Naturally, real music recordings would be preferable, but the chosen concept provides a labeled data base – including onset times, pitches and the playing music instruments – which is sufficiently large to apply our experimental design.

In most studies of music data, experiments are performed on a rather arbitrary data base of music samples where it is difficult to determine how well it represents the whole entity of music. Instead, we construct a more structured data base using an experimental design based on 8 musical factors which might have an influence on music intelligibility respectively the classification tasks. This enables identification of music which is the most problematic w. r. t. classification performance. We apply Plackett-Burman (PB) designs which are experimental designs requiring just two levels for each factor [54]. After all experiments (music samples) are evaluated, a linear regression model is fitted for predicting the error rates w. r. t. the factor levels. The goal is to identify the factors where the target variable, e.g., the error rate of pitch detection, has a significantly different expectation w. r. t. the chosen level of that factor. The values of these factors are crucial for the value of the target variable with a high probability. If no factor has a significant influence on the target variable, we can assume that the approach works equally well for all kinds of considered music. The goodness of fit of the regression model is measured by the so-called R-squared ($R^2 \in [0, 1]$) which indicates the proportion of variance that is explained by the factors. $R^2 = 1$ means that the results are completely explained by the considered factors, whereas $R^2 = 0$ means that the factor values do not influence the results, i. e. the results are independent of the type of music. Since the R-squared also depends on the number of factors, the adjusted R-squared is an attempt to compensate this effect [55]. It is defined as

$$R_a^2 = 1 - \frac{n_{\text{exp}} - 1}{n_{\text{exp}} - p_{\text{fac}} - 1}(1 - R^2), \tag{23}$$

where n_{exp} is the number of experiments and p_{fac} is the number of factors [56].

In the context of music, influence factors can be separated into two groups: factors where changes produce unnatural new tone sequences and factors where changes mostly preserve a given composition. Obviously, the problematic group is the first one since we are not interested to analyze music which sounds unnatural, and hence, we keep these factors constant. Instead, we identify original music extracts for each possible combination of these factor levels. Only the factors of the second group are changed in the MIDI annotation to get every desired combination of factor levels. We define four factors which belong to the first group and four factors which belong to the second group. The factor levels are determined by identifying typical values, considering our data base of 100 chamber music pieces. They are chosen such that the numbers of song extracts which belong to the two levels are rather equal, and in addition, a clear gap between the two levels is ensured. The factors of the first group are:

- *Mean interval size*: This is the mean interval step between two consecutive tones of the melody, measured in semitones. We define two factor levels: < 2.5 and > 3.5.

- *Onsets in accompaniment*: This factor defines the share of individual onsets produced by the accompanying instrument(s) which do not occur in the track of the melody instrument w. r. t. to all onsets. We apply two factor levels: < 0.4 and > 0.6.

- *Dynamics*: We define the dynamics of a song by the mean loudness difference of consecutive melody tones, measured in MIDI velocity numbers. We consider two factor levels: < 0.5 and > 1.0.

- *Accompanying instrument*: We consider two instruments as factor levels: piano and strings.

The four factors of the second group can take values which are, within limits, freely adjustable:

- *Melody instrument*: We consider three instruments of different instrument groups as factor levels: cello, trumpet and clarinet. Here, no natural aggregation into two factor levels exist. Hence, it is not considered within the PB designs, and instead the designs are repeated three times, one repetition for each instrument.

- *Mean pitch of the melody*: We restrict the minimum and maximum allowed pitches for the melody to the pitch range of the three considered instruments which is from E3 (165 Hz) to A6 (1047 Hz). For the experimental design we define two levels. The first level transposes the song extract (including the accompaniment) such that the average pitch of the melody is D4 (294 Hz) and the second level transposes the song extract such that the average pitch of the melody is D5 (587 Hz). Afterwards, we apply the following mechanism to prevent unnatural pitches w. r. t. the instruments. If the pitch of one tone violates the allowed pitch range the pitch of all tones within the considered song extract is shifted until all pitches are valid.

- *Tone duration*: We define the tone duration by the duration of the song extracts in order to remain the rhythmic structure. If this factor is modified, all tone lengths of the song extract are adjusted in the same way. We consider two factor levels: 12 s and 25 s which, for our data, results in tone lengths between 0.1 and 0.5 s for the first level and between 0.2 and 1.0 s for the second level.

Table 5: Plackett-Burman designs: factor levels.

Factors	1st level	2nd level
mean interval	<2.5	>3.5
onsets accompaniment	<0.4	>0.6
dynamic	<0.5	>1.0
accompaniment	piano	strings
mean pitch	D4	D5
song (tone) duration	12 s	25 s
pitch difference: melody - accompaniment	$[-6, 6]$ half tones	$[12, 24]$ half tones

- *Mean pitch of accompaniment*: This factor is the difference of the average pitch of the accompaniment compared to the average pitch of the melody. For changing this factor, we only permit transpositions of the accompaniment tracks by full octaves (12 half-tones). The two considered levels are defined by the intervals [-6,6] and [-24,-12]. If the pitches of melody and accompaniment are similar we expect higher error rates for the considered classification tasks. The case where the accompaniment is significantly higher than the melody is neglected since this is rather unusual at least for western music.

The factors and their specified levels are summarized in Table 5. We apply PB designs with 12 experiments and $p_{\text{fac}} = 7$ factors (as noted above the melody instrument is not considered within the PB design) to generate appropriate song extracts. Each experiment defines one specific combination of factor levels. First, for each experiment all possible song extracts with a length of 30 melody tones are identified from our data base of 100 MIDI songs w. r. t. the specification of the first factor group. Second, for each experiment one of these song extracts is chosen and the factors of the second group are adjusted as defined by the design. Finally, each song extract is replicated 3 times, changing the melody instrument each time. Overall, this results in $3 \cdot 12 \cdot 30 = 1080$ melody tones for each PB design. We apply three independent designs and choose different song excerpts in order to enable cross-validation. Hence, we get $n_{\text{exp}} = 3 \cdot 12 = 36$ experiments altogether. To ensure that the accompaniment is not louder than the melody, we use a melody to accompaniment ratio of 5 dB.

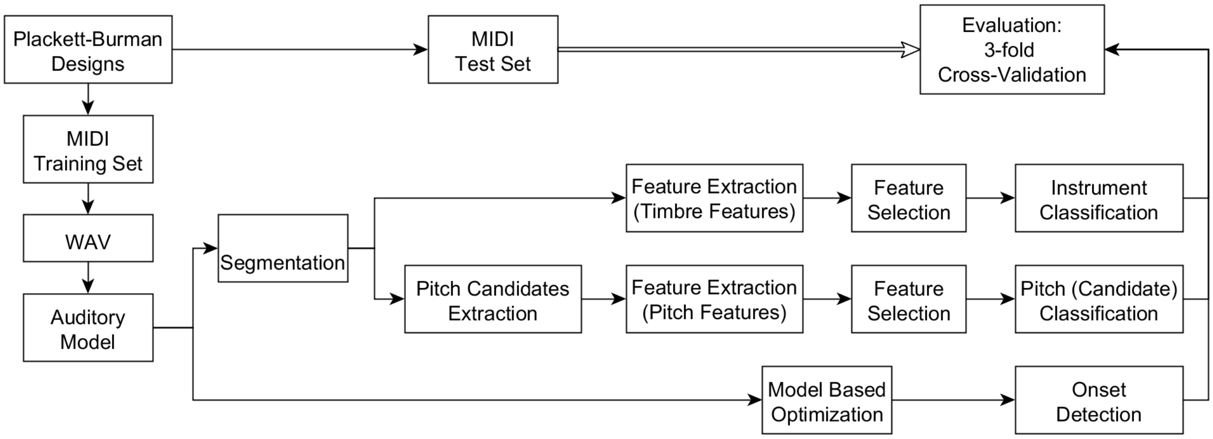
Plackett-Burman Designs
MIDI Test Set
Evaluation: 3-fold Cross-Validation
MIDI Training Set
WAV
Auditory Model
Segmentation
Feature Extraction (Timbre Features)
Feature Selection
Instrument Classification
Pitch Candidates Extraction
Feature Extraction (Pitch Features)
Feature Selection
Pitch (Candidate) Classification
Model Based Optimization
Onset Detection

Figure 6: Structure of the experiments for the music recognition tasks.

5.2 Structure of the Comparison Experiments

At first, the approaches described in the previous section are tested and compared using the original auditory model without a simulated hearing loss. The structure of the whole process is illustrated in Figure 6.

For all experiments 3-fold cross-validation is applied which means the excerpts of two designs are used for training of the classification models – or in the optimization stage in case of onset detection – and the remaining excerpts of the third design are used for testing. Additionally, the approaches are also compared on monophonic data using the same excerpts but without any accompanying instruments. Without any distortion by the accompaniment, misclassification rates should be marginal.

Since the predominant variant of onset detection is a novel issue, a comparison to existing approaches is difficult. Searching for all onsets, as well as the monophonic case are the standard problems of onset detection. Hence, apart from the monophonic and the predominant variant, we also investigate the approaches w. r. t. usual polyphonic onset detection (all onsets). All nine cases – three approaches (two with and one without an auditory model) combined with the three variants – are individually optimized using MBO with 200 iterations, which means 200 different parameter settings are tested on the training data.

For pitch estimation and instrument recognition all classification approaches are tested in two variants: RF and linear SVM (Section 4.5). For instrument recognition, this results in four considered variants altogether – features extracted from the auditory model output versus features extracted directly on the original signal – (Section 4.4). For pitch estimation, seven approaches are compared: four classification approaches with auditory features – RF or SVM and DFT or ACF features – (Section 4.3.1), two peak-picking variants for the SACF approach (Section 4.3.2), and the YIN algorithm as the state-of-the-art approach without an auditory model. However, note that we do not optimize all parameters of the YIN algorithm on our specific music data so that its outcome might be somewhat suboptimal. We use the standard settings except for the lower and the upper limits of the search range which we set to 155 Hz and 1109 Hz, respectively. These values corresponds to the pitch range of the melody in rhe considered song extracts.

For pitch and instrument recognition the feature selection approaches, described in Section 4.5.3, are used to investigate the importance of channels (best frequencies) and features. Finally, all experiments conducted for the auditory model without hearing loss are repeated for the three hearing dummies described in Section 4.1.

5.3 Software

For classification the R package *mlr* [57] is applied using the package *randomForest* [58] for RFs and the package *kernlab* [59] for SVMs. MBO is performed by using the R package *mlrMBO* [60]. Finally, the huge number of experiments performed is managed by the R packages *BatchJobs* and *BatchExperiments* [61].

6 Results

First, we present the main results regarding the auditory model for the normal-hearing person in comparison to the reference approaches (Section 6.1). Second, we consider the performance loss of models with hearing deficits exemplified by the three hearing-dummies (Section 6.2).

6.1 Comparison of Proposed Approaches

We will look at the results of onset detection, pitch estimation and instrument recognition, consecutively.

Table 6: Results (mean F-Measure) for Onset Detection with and without an Auditory Model (AM).

Design	all	melody	monoph.
w/o AM: cello	0.65	0.57	0.80
: clarinet	0.79	0.72	0.80
: trumpet	0.87	0.84	0.97
: **mean**	**0.77**	**0.71**	**0.86**
AM, best ch.: cello	0.44	0.37	0.68
: clarinet	0.65	0.61	0.80
: trumpet	0.70	0.79	0.99
: **mean**	**0.60**	**0.59**	**0.82**
AM, aggr.: cello	0.53	0.46	0.79
: clarinet	0.71	0.72	0.76
: trumpet	0.85	0.87	0.98
: **mean**	**0.69**	**0.68**	**0.84**

6.1.1 Onset Detection

Table 6 shows the results of onset detection for the three considered approaches: (1) common onset detection on the original signal (without any auditory model), (2) onset detection using the auditory model output by choosing the output of the best single channel, and (3) onset detection where the estimated onset time points of several channels are combined. For all approaches the relevant parameters are separately optimized for three tasks: monophonic onset detection (songs without accompaniment), predominant onset detection where we are just interested in the melody onsets, and onset detection where we are interested in all onsets.

All approaches perform worse than expected, even the reference approach without the auditory model, which is the state-of-the-art method for monophonic data. Solving onset detection by using only one of the auditory channels performs very differently from channel to channel as can be seen in Figure 7. For the predominant task, channels with a medium best frequency are better than low and high channels. The best performance is achieved by using the output of channel 23 resulting in an average F-value of 0.59. However, the approach which aggregates the final estimations of all channels improves this result. Interestingly, in the optimum all channels are considered, also the highest ones which individually perform very poorly as we

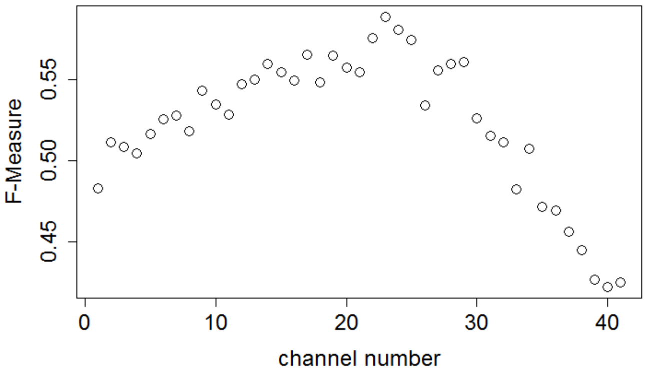

Figure 7: Results (mean F-Measure) for predominant onset detection using the output of just 1 channel.

have seen above. The average F-value of 0.68 in the predominant variant is still slightly worse than the common onset detection approach based on the original signal. However, the aggregation is based on a relatively simple classification approach, which uses just the number of estimations in the neighborhood as a single feature.

In all variants the performance for trumpet – which has a clear attack – is by far the best, whereas in most variants the performance for cello is the worst. In the predominant variant the detection of cello tones is even more difficult if it is distorted by string accompaniment. Note that a comparison of different approaches for a specific instrument should be done with care, since only the overall performance is optimized. This means, e.g., a small loss of performance for trumpet might be beneficial if this leads to a bigger gain for cello or clarinet. As expected, the results for the polyphonic variants are distinctly worse than for the monophonic variant. Furthermore, finding all onsets seems to be simpler than finding just the melody onsets, at least for the considered melody to accompaniment ratio of 5 dB.

In Table 7 the evaluation of the experimental design for the channel-aggregating method, averaged over the 3 instruments, can be seen. In the monophonic variant the adjusted R-squared (R_a^2) is negative, which indicates that the performance is independent to the type of music. This is also supported by the p-values, since neither of them shows a significant impact.

Table 7: Evaluation over all instruments and all Plackett-Burman designs for the proposed aggregation approach. The average F-value is the target variable – **a**: monophonic onset detection, **b**: predominant onset detection, and **c**: polyphonic onset detection (**bold** = significant at 10%-level).

	a		**b**		**c**	
Fit	$R^2 = 0.13, R_a^2 = -0.09$		$R^2 = 0.65, R_a^2 = 0.56$		$R^2 = 0.61, R_a^2 = 0.51$	
Factors	Estimates	p-value	Estimates	p-value	Estimates	p-value
(Intercept)	0.8448	**<2e-16**	0.6815	**<2e-16**	0.6945	**<2e-16**
mean interval	-0.0015	0.90	-0.0041	0.76	0.0308	0.17
onsets accompaniment	-0.0021	0.87	-0.0636	**4e-05**	-0.0448	**0.05**
dynamic	-0.0146	0.25	-0.0186	0.17	-0.0019	0.93
accompaniment	0.0177	0.16	-0.0109	0.41	-0.1313	**2e-06**
mean pitch	0.0029	0.81	0.0510	**6e-04**	0.0198	0.37
tone duration	-0.0087	0.49	-0.0348	**0.01**	-0.0026	0.91
pitch: mel. - acc.	0.0051	0.68	-0.0224	**0.10**	-0.0213	0.34

Obviously, this was expected for some factors which correspond to the accompaniment so that they should only have an impact in the polyphonic case. However, before the experiments, we expected that greater values of the *mean interval* should simplify onset detection.

For the other two variants of onset detection, the goodness of fit is relatively high ($R_a^2 > 0.5$) – note that we describe music pieces by just 8 dimensions which explains a relatively high amount of noise in all evaluation models of the experimental design. Thus, we can identify some important influence factors w. r. t. the performance of the proposed algorithm. In the predominant variant, the performance is better if the number of onsets produced by the accompaniment is low, which obviously was expected. However, a higher mean pitch and shorter tones also seem to be beneficial. In the polyphonic variant piano accompaniment is better than string accompaniment. This effect is explained by the bad performance of onset detection for string instruments in general as we have already seen for cello. Furthermore, also in this scenario, a smaller number of individual onsets produced by the accompaniment is beneficial, probably because mutual onsets of melody and accompaniment are easier to identify.

Comparison to Human Perception Although there is a wide range of publications dealing with the human perception of rhythm (see [62] for an overview), none of them analyzes the human ability to recognize onsets in musical pieces. Reason for this might be the fact that onset detection is a rather trivial task for normal-hearing listeners at least for chamber music. This is the case particular for monophonic music where only the detection of very short tones and the separation of two identical consecutive tones of bowed instruments seem to be challenging. According to Krumhansl, the critical duration between two tones for event separation is 100 ms [62], a threshold which is exceeded for all pairs of tones in this study.

An informal listening test with our monophonic music data indicates that even all onsets of identical consecutive tones can be identified by a trained normal-hearing listener. However, to study a worst case scenario, let us assume (s)he does not recognize these onsets in case of the cello. That means, 94 out of the 3240 onsets are missed which corresponds to a misclassification rate of 2.9% and an F-value of 0.99. Contrary, even the state-of-the-art method without the auditory model, achieves a mean F-value of only 0.86 which, according to (8), means that 24.6% of all onsets are missed, if we assume that the algorithm does not produce any false alarm. In conclusion, in the field of automatic onset detection big improvements are necessary to simulate human perception.

6.1.2 Pitch Estimation

Table 8 lists the average error rates of pitch detection using the methods described in Section 4.3 for the three instruments. Additionally, also the results for the monophonic data are listed. Our approach using spectral features of the auditory output and a linear SVM for classification performs best and even clearly outperforms the YIN algorithm. In all cases, the error rates for clarinet are clearly the lowest, whereas cello tones seem to be the most difficult ones. The pitch of clarinet tones is easier to estimate because these tones have a relatively low intensity of the even partials which might prevent several octave errors. For trumpet and cello tones, often the frequency of the second partial is wrongly estimated as the fundamental one. Again, pitches of cello tones which are accompanied by string instruments are especially difficult to estimate. As expected, in the monophonic variant all approaches perform clearly better than in the polyphonic one. Here, again the spectral approach performs best. However, in this case RF and linear SVM perform

Table 8: Mean error rates of pitch detection methods.

	polyphonic				mono.
Method	cello	clar.	trump.	**mean**	mean
SACF max.	0.55	0.52	0.54	0.54	0.20
SACF thresh.	0.24	0.12	0.17	0.18	0.05
DFT + RF	0.14	0.02	0.08	0.08	0.02
DFT + SVM	0.11	0.01	0.08	**0.07**	0.02
ACF + RF	0.24	0.08	0.30	0.20	0.05
ACF + SVM	0.21	0.05	0.24	0.17	0.04
YIN	0.36	0.15	0.32	0.28	0.05

Table 9: Feature selection for pitch classification with auditory model and DFT: number of selected features and error rates.

Method	no selection	channel groups		feature groups	
		forward	backward	forward	backward
RF: number of features	$41 \cdot 29 = 1189$	$4 \cdot 29 = 116$	$35 \cdot 29 = 1015$	$41 \cdot 2 = 82$	$41 \cdot 28 = 1148$
RF: error rate	0.08	0.10	0.07	0.09	0.08
SVM: number of features	$41 \cdot 29 = 1189$	$5 \cdot 29 = 145$	$23 \cdot 29 = 667$	$41 \cdot 2 = 82$	$41 \cdot 9 = 369$
SVM: error rate	0.07	0.10	0.07	0.09	0.07

equally well.

For the best method – the classification approach using spectral features and either linear SVM or RF – group-based feature selection, as introduced in Section 4.5.3, is performed. The corresponding results are listed in Table 9. Especially, feature-based grouping shows good results. For both classification methods, the forward variant finishes with just 2 feature groups – instead of 29 without feature selection – where the performance reduction is only small. Interestingly, the two classifiers choose different features. For RF, $c[k]$ and $d_{right}[k]$ are picked, whereas for SVM, $p_{mean}[k]$ and $P_1^{mean}[k]$ are chosen. In the backward variant, the SVM just needs the following 9 feature groups to achieve the same error rate as with all features: $c[k]$, $p_{mean}[k]$, $p_{max}[k]$, $b[k]$, $d_{left}[k]$, $d_{right}[k]$, $P_4^{mean}[k]$, $P_8^{mean}[k]$ and $P_9^{mean}[k]$. All other features might be meaningless or redundant.

Also some channels can be omitted: For classification with SVM, 23 channels instead of all 41 are sufficient to get the best error rate of 0.07. The ignored channels are located in all regions, which means no priority to lower

Table 10: Evaluation over all instruments and all Plackett-Burman Designs. The error rate is the target variable – **a**: Pitch Estimation and SVM (auditory model + DFT), – **b**: Instrument Recognition and SVM (auditory model features) – (**bold** = significant at 10%-level).

	a		**b**	
Fit	$R^2 = 0.30, R_a^2 = 0.12$		$R^2 = 0.37, R_a^2 = 0.21$	
Coefficients	Estim.	p-value	Estim.	p-value
(Intercept)	0.0660	**2e-08**	0.0111	**9e-04**
interval	-0.0074	0.40	0.0049	0.11
onsets acc.	0.0056	0.53	0.0037	0.23
dynamic	-0.0142	0.11	-0.0019	0.54
acc.	0.0148	**0.10**	0.0056	**0.07**
mean pitch	-0.0068	0.44	0.0062	**0.05**
tone dur.	-0.0025	0.78	0.0025	0.42
mel. - acc.	-0.0185	**0.04**	-0.0056	**0.07**

or higher channels can be observed, and the crucial information is redundant in neighboring (overlapping) channels.

Table 10a shows the evaluation of the experimental design. The goodness of fit ($R_a^2 = 0.12$) is rather low but some weakly significant influence factors can be identified. For example, a bigger distance between the average pitch of melody and accompaniment seems to be advantageous. This was expected, since a bigger distance leads to a lesser number of overlapping partials. Additionally, there is a small significant influence regarding the kind of accompaniment: piano accompaniment seems to be beneficial. Again this sounds logical as it is difficult to distinguish cello tones from tones of other string instruments.

Comparison to Human Perception There exist several studies which investigate the ability of human pitch perception (see [62] and [63] for an overview). In most of these studies the ability to recognize relative changes of consecutive tones is quantified. Frequency differences of about 0.5% can be recognized by a normal-hearing listener [64]. However, quantifying these differences, is a much harder challenge. Discriminating thresholds for this task are in the magnitude of a semitone for listeners without musical training which corresponds to a frequency difference of approximately 6%[65]. The ability to recognize such relative changes is called relative pitch which is the natural way most people perceive pitches. However, relative pitch remains poorly understood, and the standard view of the auditory system corresponds to absolute pitch since common pitch models make absolute, rather than relative, features of a sound's spectrum explicit [63]. In fact, also some humans can perceive absolute pitch which is the ability to label pitches without a reference point. It is assumed that this requires acquisition early in life. Also absolute pitch possessors make errors - most times octave and semitone errors - whose rate varies strongly between individuals [66].

In conclusion, comparing the results of our study to human data is a big challenge. Nevertheless, considering the ability of a normal-hearing listener for relative pitch, we can assume, that (s)he might be able to perceive the pitches almost perfectly w. r. t. the tolerance of 1/2 semitone at least in the monophonic case. This estimation approximately corresponds to the result of the classification method with DFT-features which yields an error rate of 2% in our study. The human ability for the perception of polyphonic music has not yet been adequately researched to make any estimations. Hence, in future studies appropriate listening tests are necessary.

6.1.3 Instrument Classification

The error rates for instrument classification are listed in Table 11. Here, the auditory model-based features perform distinctly better than the standard features. In both cases, the linear SVM performs slightly better than the RF. Distinguishing trumpet from the rest seems to be slightly more difficult than identifying cello or clarinet. In the monophonic variant, the results are nearly perfect for all variants. Since the auditory model based features are only beneficial in the polyphonic case, we conclude that these features enhance the ability to separate individual voices or instruments.

Table 12 shows the result of feature selection for instrument recognition. Here, both backward variants even slightly improve the no-selection result for RF. Using only the features of 12 channels leads to the best result which is equally well as the SVM with all features. The selected channels are 8, 12, 19, 21, 22, 24, 26, 27, 28, 32, 33 and 41. Comparing the best frequencies of these channels and the pitch range of the melody explains why the low channels are unimportant. The fundamental frequency of the considered melody tones is between 165 Hz and 1047 Hz, corresponding to the channels

Table 11: Mean error rates of instrument recognition methods.

	polyphonic				monophonic
Method	Cello vs. all	Clarinet vs. all	Trumpet vs. all	**Overall**	Overall
AM features, RF	0.012	0.017	0.029	0.019	0.002
AM features, SVM	0.007	0.007	0.014	**0.011**	0.001
Standard features, RF	0.044	0.034	0.052	0.063	0.000
Standard features, SVM	0.025	0.019	0.054	0.035	0.002

Table 12: Feature Selection for instrument recognition with auditory model features: number of selected features and error rates.

Method	no selection	channel groups		feature groups	
		forward	backward	forward	backward
RF number of features	$41 \cdot 21 = 861$	$2 \cdot 21 = 42$	$12 \cdot 21 = 420$	$41 \cdot 3 = 123$	$41 \cdot 17 = 697$
RF error rate	0.019	0.034	0.011	0.058	0.016
SVM number of features	$41 \cdot 21 = 861$	$2 \cdot 21 = 42$	$12 \cdot 21 = 420$	$41 \cdot 3 = 123$	$41 \cdot 8 = 328$
SVM error rate	0.011	0.030	0.017	0.045	0.015

6 to 24 which have best frequencies between 167 Hz and 1053 Hz. Also some of the higher channels are important which supply information about overtones and possibly the fine structure. However, the deselection of several channels also illustrates the redundancy of neighboring channels.

According to the results of forward selection, two channels are sufficient to get error rates of about 3%. Channels 26 and 41 are chosen for RF and channels 29 and 41 for SVM. The gain of higher channels for instrument recognition is further illustrated in Figure 8. Applying the features of one of the first channels leads to an error rate of almost 40%, whereas the features of the 41st channel generate a model with an error rate below 5%. This is also interesting for our examination of auditory models with hearing loss since usually particularly the higher channels are degraded the most. Also in the backward variant of channel-based grouping, the lowest channels are omitted.

In the feature-based forward variant, the same three feature groups are selected for SVM and RF, respectively: *mean spectral flux*, *root-mean-square energy* and *spectral rolloff*. In the backward variant using the SVM, these three features are also chosen and five additional ones: *irregularity* and the 1st, the 3rd, the 4th, and the 7th MFCC coefficients.

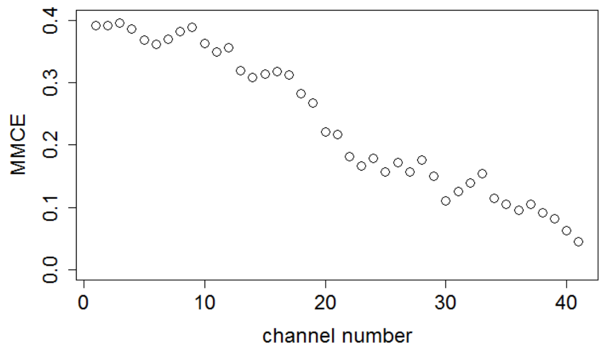

Figure 8: Mean misclassification error (MMCE) for predominant instrument recognition using the features of just 1 channel and the linear SVM.

Table 10b shows the evaluation of the experimental design for predominant instrument recognition. Here, the goodness of fit is moderate ($R_a^2 = 0.21$) and three weakly significant influence factors can be identified. The most significant influence has the mean pitch, i. e. lower tones can be distinguished better. Also string accompaniment affects the error rates more than piano accompaniment. Again, the reason might be the difficulty to distinguish cello from other string instruments. Additionally, a bigger distance between the pitches of melody and accompaniment also seems to be beneficial.

Comparison to Human Perception Most studies about timbre in the field of music psychology try to quantify dissimilar ratings and analyze their correlations to physical features, whereas the common task in the field of music information retrieval is instrument recognition. Although both perceptions are very similar, there exist one important difference which causes diverging results of the two disciplines. Dissimilar ratings are subjective measures which rely on judgements of humans, wheras instrument recognition is a well-defined task [67]. Nevertheless, also some studies have conducted experiments about the human ability to distinguish music instruments (see [68] for a tabular overview). The most comprehensive experiment is reported in [69],

Table 13: Results (mean F-measure) of Onset Detection for hearing-dummies (HD) compared to the normal-hearing (NH) model.

Task	**Monophonic**				**Melody Onsets**				**All Onsets**			
Hearing Impairment	NH	HD1	HD2	HD3	NH	HD1	HD2	HD3	NH	HD1	HD2	HD3
Cello	0.79	0.67	0.74	0.78	0.46	0.37	0.44	0.45	0.53	0.46	0.50	0.53
Clarinet	0.76	0.75	0.77	0.72	0.72	0.58	0.70	0.69	0.71	0.62	0.70	0.69
Trumpet	0.98	0.99	0.98	0.98	0.87	0.70	0.80	0.86	0.85	0.74	0.81	0.83
Mean	0.84	0.80	0.83	0.83	0.68	0.55	0.65	0.67	0.69	0.61	0.67	0.68

a listening experiment with music experts. The subjects had to distinguish isolated notes of 27 instruments. The recognition accuracy was 46% for individual instruments and 92% for instrument families which included the five categories string, brass, double reed, clarinet and flutes. The latter result can be compared to the monophonic variant in this study although the task here is distinctly easier since only three categories have to be distinguished and for each category only one representantive instrument is considered. Some informal listening experiments indicate that a trained normal-hearing listener might distinguish the three instruments as perfectly as the classification approach does. To our best knowledge no experiments exist which study the human ability for instrument classification in a polyphonic scenario. As for pitch estimation, this is a crucial topic for future studies.

6.2 Evaluation of Hearing Dummies

The results of onset detection for the three hearing dummies (HD) described in Section 4.1 are listed in Table 13. For all three considered tasks – monophonic, predominant and usual polyphonic – HD2 and HD3 perform just a little worse than the normal-hearing model. This is an indicator that these moderate hearing losses have no big impact on the recognition rates of tone onsets, although this result should be considered with care due to the overall relative poor results of automatic onset detection. However, HD1 performs distinctly worse, particularly in the case of predominant onset detection.

In Table 14, the error rates of predominant pitch estimation for hearing dummies are listed. For all considered approaches the results are as expected: the greater the hearing deficit is, the greater are the error rates. Even HD3 performs a little worse than the model without hearing loss, although the kind of hearing loss affects only frequencies above the fundamental frequencies of

Table 14: Mean error rates of pitch detection methods for hearing dummies (HD).

Method	NH	HD1	HD2	HD3
SACF max.	0.54	0.67	0.60	0.56
SACF thresh.	0.18	0.44	0.34	0.22
DFT + RF	0.08	0.32	0.29	0.10
DFT + SVM	0.07	0.32	0.24	0.09
ACF + RF	0.20	0.91	0.42	0.21
ACF + SVM	0.17	0.90	0.40	0.18

Table 15: Mean error rates of instrument recognition methods for hearing dummies (HD).

Method	NH	HD1	HD2	HD3
AM and RF	0.02	0.28	0.03	0.05
AM and SVM	0.01	0.26	0.02	0.04

all considered tones. However, this is consistent with results of psychoacoustic experiments which also report an important impact of higher partials (and channels) on pitch estimation [70].

For instrument recognition similar results can be observed as shown in Table 15. However, this time HD2 performs better than HD3, since here, higher channels are the most relevant ones as we have already seen in Figure 8.

7 Summary and Conclusion

Music intelligibility is simplified into three tasks of music classification: onset detection, pitch estimation and instrument recognition. We can conclude that pitch estimation and instrument recognition are solved well by using the output of an auditory model. In our experiments, the performance of the proposed approach is distinctly better than the performances of the reference approaches without an auditory model.

The results for onset detection are disappointing, but this is also true for the reference approach. State-of-the-art in onset detection performs rather poorly especially when dealing with polyphonic music. Especially, the detection of cello onsets is problematic, where the average F-value in the predominant variant is just 0.57. Nevertheless, we think also these results imply information about the level of difficulty for tone recognition, e.g., also for a human listener tone onsets of a trumpet are easier to identify than onsets of a cello. Furthermore, one could also just analyze the results for musical instruments which perform satisfactorily, e.g., for trumpet, the average F-value is 0.84 in the predominant case, and in the monophonic case, an almost perfect value of 0.97 is achieved.

Classical onset detection can be easily adapted to a single channel output of the auditory model. The challenge arises how to combine the estimations of several channels. Our approach which handles each proposed onset time point as a candidate and subsequently classifies whether it is in fact an onset seems to be promising. Although the results for all three considered scenarios are still slightly worse than the results of onset detection without the auditory model, there are many possible resources for improvements since the proposed classification method is as simple as possible by just considering one feature. Therefore, the approach might be extended to combine estimations of additional features apart from spectral flux.

For predominant pitch detection, our introduced approach which applies spectral features and reduces the problem to a classification problem performs clearly better than the autocorrelation method. The linear SVM performs best with the error rate of 7%. The number of features can be drastically reduced without decreasing the prediction rate by applying group-based feature selection. The features of 23 channels (instead of 41) or the reduction to 9 types of features (instead of 29) lead to identical error rates. For future studies it would be interesting to combine the two feature selection strategies which might reduce computation time even more. The features corresponding to the average spectral amplitude over all channels of the partials ($P_{pl}^{mean}[k]$) seem to be more meaningful than the features corresponding to the maximum amplitude ($P_{pl}^{max}[k]$). However, also most of these former features are excluded by feature selection. Nearly all other features described in Section 4.3 seem to be important and are included by feature selection.

For predominant instrument recognition, the three considered instruments can be almost perfectly distinguished with an error rate of 1.1% by using the auditory features and either linear SVM or RF. Particularly important are the features of the higher channels. For the RF, twelve channels are sufficient to achieve the best error rate. Since the common features (without auditory model) are competitive in the monophonic variant, the benefit of auditory model-features seems to be an enhanced ability for separating different instruments in a polyphonic environment.

For all three considered hearing-dummies, the error rates increase for all classification tasks. The degree of impairment seems to be plausible with respect to the specific hearing deficits. In future studies, these results should be compared to and verified by listening tests, which were beyond the scope of this study.

Applying an experimental design for selecting the examined song excerpts offers the interesting possibility to identify the type of music for which specific tasks are significantly harder or easier to solve than on average. We got some unexpected results, e.g., higher pitches and shorter tones are beneficial for predominant onset detection, whereas lower pitches improve the results of predominant instrument recognition. In future studies, the experimental design could be enhanced by further factors, e.g., varying the melody to accompaniment ratio might be interesting.

In a next step, we want to combine the proposed approaches to estimate an overall measure for music intelligibility, which could be applied for assessing and optimizing hearing instruments for music with several parameters to adjust. Such optimization shall result in several promising parameter setting candidates, which, finally, can be verified and ranked by a listening test.

Acknowledgements

The paper is based on investigations of the projects B3 and C2 of the Collaborative Research Center SFB 823, which is kindly supported by Deutsche Forschungsgemeinschaft (DFG).

References

[1] McDermott, H.J.: Music perception with cochlear implants: a review. Trends in amplification **8**(2), 49–82 (2004). doi:10.1177/108471380400800203

[2] Gfeller, K.E., Olszewski, C., Turner, C., Gantz, B., Oleson, J.: Music perception with cochlear implants and residual hearing. Audiology and Neurotology **11**(Suppl. 1), 12–15 (2006). doi:10.1159/000095608

[3] Emiroglu, S., Kollmeier, B.: Timbre discrimination in normal-hearing and hearing-impaired listeners under different noise conditions. Brain research **1220**, 199–207 (2008). doi:10.1016/j.brainres.2007.08.067

[4] Fitz, K., Burk, M., McKinney, M.: Multidimensional perceptual scaling of musical timbre by hearing-impaired listeners. In: Proceedings of Meetings on Acoustics, vol. 6 (2009). doi:10.1121/1.3186749. Acoustical Society of America

[5] Jürgens, T., Ewert, S.D., Kollmeier, B., Brand, T.: Prediction of consonant recognition in quiet for listeners with normal and impaired hearing using an auditory modela. The Journal of the Acoustical Society of America **135**(3), 1506–1517 (2014). doi:10.1121/1.4864293

[6] Karbasi, M., Kolossa, D.: A microscopic approach to speech intelligibility prediction using auditory models. In: Proc. Annual Meeting of the German Acoustical Society (DAGA) (2015)

[7] Chasin, M., Russo, F.A.: Hearing aids and music. Trends in Amplification **8**(2), 35–47 (2004)

[8] Fitz, K., McKinney, M.: Music through hearing aids: perception and modeling. In: Proceedings of Meetings on Acoustics, vol. 9 (2015). doi:10.1121/1.3436580. Acoustical Society of America

[9] Maganti, H.K., Matassoni, M.: Auditory processing-based features for improving speech recognition in adverse acoustic conditions. EURASIP Journal on Audio, Speech, and Music Processing **2014**(1), 1–9 (2014). doi:10.1186/1687-4722-2014-21

[10] Klapuri, A.: Multipitch analysis of polyphonic music and speech signals using an auditory model. Audio, Speech, and Language Processing, IEEE Transactions on **16**(2), 255–266 (2008). doi:10.1109/TASL.2007.908129

[11] McLeod, P.: Fast, accurate pitch detection tools for music analysis. PhD Thesis, University of Otago. Department of Computer Science (2009)

[12] Heinz, M.G., Zhang, X., Bruce, I.C., Carney, L.H.: Auditory nerve model for predicting performance limits of normal and impaired listeners. Acoustics Research Letters Online **2**(3), 91–96 (2001). doi:10.1121/1.1387155

[13] Zilany, M.S., Bruce, I.C.: Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery. The Journal of the Acoustical Society of America **120**(3), 1446–1466 (2006). doi:10.1121/1.2225512

[14] Jepsen, M.L., Dau, T.: Characterizing auditory processing and perception in individual listeners with sensorineural hearing loss. The Journal of the Acoustical Society of America **129**(1), 262–281 (2011). doi:10.1121/1.3518768

[15] Meddis, R., Lecluyse, W., Tan, C.M., Panda, M.R., Ferry, R.: Beyond the audiogram: Identifying and modeling patterns of hearing deficits, 631–640 (2010). doi:10.1007/978-1-4419-5686-6_57

[16] Panda, M.R., Lecluyse, W., Tan, C.M., Jürgens, T., Meddis, R.: Hearing dummies: Individualized computer models of hearing impairment. International journal of audiology **53**(10), 699–709 (2014). doi:10.3109/14992027.2014.917206

[17] Salamon, J., Gómez, E.: Melody extraction from polyphonic music signals using pitch contour characteristics. Audio, Speech, and Language Processing, IEEE Transactions on **20**(6), 1759–1770 (2012). doi:10.1109/TASL.2012.2188515

[18] Schluter, J., Bock, S.: Improved musical onset detection with convolutional neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6979–6983 (2014). doi:10.1109/ICASSP.2014.6854953. IEEE

[19] Bello, J.P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., Sandler, M.B.: A tutorial on onset detection in music signals. IEEE Transactions on Speech and Audio Processing **13**(5), 1035–1047 (2005). doi:10.1109/TSA.2005.851998

[20] Bauer, N., Friedrichs, K., Bischl, B., Weihs, C.: Fast model based optimization of tone onset detection by instance sampling. In: Adalbert F.X. Wilhelm, H.A.K. (ed.) Analysis of Large and Complex Data. Springer, Bremen, Germany (2016)

[21] Klapuri, A.: Sound onset detection by applying psychoacoustic knowledge. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 6, pp. 3089–3092 (1999). doi:10.1109/ICASSP.1999.757494. IEEE

[22] Holzapfel, A., Stylianou, Y., Gedik, A.C., Bozkurt, B.: Three dimensions of pitched instrument onset detection. IEEE Transactions on Audio, Speech, and Language Processing **18**(6), 1517–1527 (2010). doi:10.1109/TASL.2009.2036298

[23] Bauer, N., Friedrichs, K., Kirchhoff, D., Schiffner, J., Weihs, C.: Tone onset detection using an auditory model. In: Spiliopoulou, M., Schmidt-Thieme, L., Janning, R. (eds.) Data Analysis, Machine Learning and Knowledge Discovery vol. Part VI, pp. 315–324. Springer, Hildesheim, Germany (2014). doi:10.1007/978-3-319-01595-8_34

[24] De Cheveigné, A., Kawahara, H.: Yin, a fundamental frequency estimator for speech and music. The Journal of the Acoustical Society of America **111**(4), 1917–1930 (2002). doi:10.1121/1.1458024

[25] Mauch, M., Dixon, S.: pyin: A fundamental frequency estimator using probabilistic threshold distributions. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 659–663 (2014). IEEE

[26] Duan, Z., Pardo, B., Zhang, C.: Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions. IEEE Transactions on Audio, Speech, and Language Processing **18**(8), 2121–2133 (2010). doi:10.1109/TASL.2010.2042119

[27] Klapuri, A.: A classification approach to multipitch analysis. In: 6th Sound and Music Computing Conference, Porto, Portugal (2009)

[28] Meddis, R., Hewitt, M.J.: Virtual pitch and phase sensitivity of a computer model of the auditory periphery. i: Pitch identification. The Journal of the Acoustical Society of America **89**(6), 2866–2882 (1991). doi:10.1121/1.400725

[29] Meddis, R., O'Mard, L.: A unitary model of pitch perception. The Journal of the Acoustical Society of America **102**(3), 1811–1820 (1997). doi:10.1121/1.420088

[30] Goto, M.: A real-time music-scene-description system: Predominant-f0 estimation for detecting melody and bass lines in real-world audio signals. Speech Communication **43**(4), 311–329 (2004). doi:10.1016/j.specom.2004.07.001

[31] Sandrock, T.: Multi-label feature selection with application to musical instrument recognition. PhD thesis, Stellenbosch: Stellenbosch University (2013)

[32] Martin, K.D., Kim, Y.E.: Musical instrument identification: A pattern-recognition approach. The Journal of the Acoustical Society of America **104**(3), 1768–1768 (1998). doi:10.1121/1.424083

[33] Patil, K., Elhilali, M.: Biomimetic spectro-temporal features for music instrument recognition in isolated notes and solo phrases. EURASIP Journal on Audio, Speech, and Music Processing **2015**(1), 1–13 (2015). doi:10.1186/s13636-015-0070-9

[34] Wieczorkowska, A., Kubera, E., Kubik-Komar, A.: Analysis of recognition of a musical instrument in sound mixes using support vector machines. Fundamenta Informaticae **107**(1), 85–104 (2011)

[35] Bosch, J.J., Janer, J., Fuhrmann, F., Herrera, P.: A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals. In: ISMIR, pp. 559–564 (2012)

[36] Meddis, R.: Auditory-nerve first-spike latency and auditory absolute threshold: a computer model. The Journal of the Acoustical Society of America **119**(1), 406–417 (2006). doi:10.1121/1.2139628

[37] Lopez-Poveda, E.A., Meddis, R.: A human nonlinear cochlear filterbank. The Journal of the Acoustical Society of America **110**(6), 3107–3118 (2001). doi:10.1121/1.1416197

[38] Jürgens, T., N., C., W., L., R., M.: Exploration of a physiologically-inspired hearing-aid algorithm using a computer model mimicking impaired hearing. International journal of audiology **55**, 346–357 (2016). doi:10.3109/14992027.2015.1135352

[39] Böck, S., Krebs, F., Schedl, M.: Evaluating the online capabilities of onset detection methods. In: ISMIR, pp. 49–54 (2012)

[40] Rosao, C., Ribeiro, R., De Matos, D.M.: Influence of peak selection methods on onset detection. In: ISMIR, pp. 517–522 (2012)

[41] Vos, J., Rasch, R.: The perceptual onset of musical tones. Perception & psychophysics **29**(4), 323–335 (1981)

[42] High Performance Computer-Cluster LiDOng. `http://lidong.itmc.tu-dortmund.de/ldw/index.php?title=System_overview&oldid=322`

[43] Jones, D.R., Schonlau, M., Welch, W.J.: Efficient global optimization of expensive black-box functions. Journal of Global optimization **13**(4), 455–492 (1998). doi:10.1023/A:1008306431147

[44] Bischl, B., Wessing, S., Bauer, N., Friedrichs, K., Weihs, C.: Moi-mbo: multiobjective infill for parallel model-based optimization. In: Learning and Intelligent Optimization, pp. 173–186. Springer, Gainesville, FL, USA (2014). doi:10.1007/978-3-319-09584-4_17

[45] Friedrichs, K., Weihs, C.: Auralization of auditory models. In: Classification and Data Mining, pp. 225–232. Springer, Florence, Italy (2013). doi:10.1007/978-3-642-28894-4_27

[46] Weihs, C., Friedrichs, K., Bischl, B.: Statistics for hearing aids: Auralization. In: Second Bilateral German-Polish Symposium on Data Analysis and Its Applications (GPSDAA), pp. 183–196 (2012)

[47] Friedrichs, K., Weihs, C.: Comparing timbre estimation using auditory models with and without hearing loss. Technical Report 51/2012, Department of Statistics, TU Dortmund University (2012). doi:10.17877/DE290R-10355

[48] Lartillot, O., Toiviainen, P.: A matlab toolbox for musical feature extraction from audio. In: Proc. of the 10th Int. Conference on Digital Audio Effects (DAFx-07), pp. 1–8 (2007)

[49] Breiman, L.: Bagging predictors. Machine learning **24**(2), 123–140 (1996). doi:10.1007/BF00058655

[50] Breiman, L.: Random forests. Machine Learning Journal **45**(1), 5–32 (2001). doi:10.1023/A:1010933404324

[51] Vapnik, V.: Statistical Learning Theory. John Wiley and Sons, USA (1998)

[52] Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial intelligence **97**(1), 273–324 (1997). doi:10.1016/S0004-3702(97)00043-X

[53] Jensen, J.H., Christensen, M.G., Jensen, S.H.: A framework for analysis of music similarity measures. In: Proc. European Signal Processing Conf, pp. 926–930 (2007)

[54] Plackett, R.L., Burman, J.P.: The design of optimum multifactorial experiments. Biometrika, 305–325 (1946). doi:10.2307/2332195

[55] Fahrmeir, L., Kneib, T., Lang, S.: Regression: Modelle, Methoden und Anwendungen. Springer, Berlin Heidelberg New York (2007)

[56] Yin, P., Fan, X.: Estimating r 2 shrinkage in multiple regression: A comparison of different analytical methods. The Journal of Experimental Education **69**(2), 203–224 (2001)

[57] Bischl, B., Lang, M., Richter, J., Bossek, J., Judt, L., Kuehn, T., Studerus, E., Kotthoff, L.: Mlr: Machine Learning in R. R package version 2.5. `https://github.com/mlr-org/mlr`

[58] Liaw, A., Wiener, M.: Classification and regression by randomforest. R News **2**(3), 18–22 (2002)

[59] Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A.: kernlab – an S4 package for kernel methods in R. Journal of Statistical Software **11**(9), 1–20 (2004). doi:10.18637/jss.v011.i09

[60] Bischl, B., Bossek, J., Horn, D., Lang, M.: mlrMBO: Model-Based Optimization for Mlr. R package version 1.0. `https://github.com/berndbischl/mlrMBO`

[61] Bischl, B., Lang, M., Mersmann, O., Rahnenführer, J., Weihs, C.: BatchJobs and BatchExperiments: Abstraction mechanisms for using R in batch environments. Journal of Statistical Software **64**(11), 1–25 (2015). doi:10.18637/jss.v064.i11

[62] Krumhansl, C.L.: Rhythm and pitch in music cognition. Psychological bulletin **126**(1), 159 (2000)

[63] McDermott, J.H., Oxenham, A.J.: Music perception, pitch, and the auditory system. Current opinion in neurobiology **18**(4), 452–463 (2008)

[64] Wier, C., Jesteadt, W., Green, D.: Frequency discrimination as a function of frequency and sensation level. Journal of the Acoustical Society of America **61**(1), 178–184 (1977). doi:10.1121/1.381251

[65] Burns, E.M., Ward, W.D.: Categorical perception–phenomenon or epiphenomenon: evidence from experiments in the perception of melodic musical intervals. The Journal of the Acoustical Society of America **63**(2), 456–68 (1978)

[66] Levitin, D.J., Rogers, S.E.: Absolute pitch: perception, coding, and controversies. Trends in cognitive sciences **9**(1), 26–33 (2005)

[67] Siedenburg, K., Fujinaga, I., McAdams, S.: A comparison of approaches to timbre descriptors in music information retrieval and music psychology. Journal of New Music Research **45**(1), 27–41 (2016)

[68] Brown, J.C., Houix, O., McAdams, S.: Feature dependence in the automatic identification of musical woodwind instruments. The Journal of the Acoustical Society of America **109**(3), 1064–1072 (2001)

[69] Martin, K.D.: Sound-source recognition: A theory and computational model. PhD thesis, Massachusetts Institute of Technology (1999)

[70] Oxenham, A.J.: Pitch perception and auditory stream segregation: implications for hearing loss and cochlear implants. Trends in amplification **12**(4), 316–331 (2008)