Friedrichs *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2017) 2017:7
DOI 10.1186/s13636-017-0103-7

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

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

Klaus Friedrichs[1*], Nadja Bauer[1], Rainer Martin[2] and Claus Weihs[1]

Abstract

The benefit 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 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 impeded by a distorted perception of music signals. Recently, several listening experiments have been conducted to assess the impact of hearing loss on music perception for hearing-impaired listeners (e.g., [1–4]). For many applications like optimization of hearing instruments, it is desirable to measure this impact automatically using 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. Auditory models are computational models which mimic the human auditory process by transforming acoustic signals into neural activity of 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, 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 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 three separate music recognition tasks, i.e., onset detection, pitch estimation, and instrument recognition. Contrary to state-of-the-art methods, here, we extract information from auditory output only. In fact, some recent proposals in the field of speech recognition and music data analysis use auditory models, thus exploiting the superiority of the human auditory system (e.g., [9–11]). 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 [12–15]. In [16, 17], it is shown that

*Correspondence: friedrichs@statistik.tu-dortmund.de
[1] Department of Statistics, TU Dortmund University, 44221 Dortmund, Germany
Full list of author information is available at the end of the article

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 [18] (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 due to the emulation of hearing impairments 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 [19]. In particular, onset detection is still rather error-prone for polyphonic music [20]. 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 optional pre-processing stage, a reduction function (called onset detection function), which is derived at a lower sampling rate, and a peak-picking algorithm [21]. They all can be summarized into one algorithm with several parameters to optimize. In [22], 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 where each channel corresponds to a different frequency band. Here, the additional challenge lies in the combination of different onset predictions of several channels. In [23], 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 [24] where the individual estimation vectors are combined via summing and smoothing. Instead of combining the individual estimations at the end, in [25], 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. Onset detection of the predominant voice 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 improvements are implemented in the popular YIN algorithm [26] which in fact uses the difference function instead of the ACF. A further extension is the pYIN method which is introduced in [27]. It is a two-stage method which takes past estimations into account. First, for every frame, several fundamental frequency candidates are predicted, and second, the most probable temporal path is estimated, according to a hidden Markov model. In [28], a maximum-likelihood approach is introduced in the frequency domain. Another alternative is a statistical classification approach which is proposed in [29].

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 [30, 31]. In those studies, an autocorrelation method is proposed where the individual running ACFs of each channel are combined by summation (averaging) across all channels (SACF). The results of that approach are equivalent to human performance 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 [32], 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 detection 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 [33]. For building a classifier, 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. However, features based on an auditory model have also been introduced in [34]. Also, biomimetic spectro-temporal features, requiring a model of higher central auditory stages, have been successfully investigated for solo music recordings in [35]. Predominant instrument recognition can be solved similarly to the monophonic variant, but is much harder due to the additional "noise" from the accompanying instruments [36]. An alternative is starting with sound source separation in order to apply monophonic instrument recognition afterwards [37]. Naturally, this concept can only work if the sources are well separated, a task which itself is still a challenge.

3 Contribution of the paper

As there exist only very few approaches for music recognition tasks using a comprehensive auditory model, in this study, new methods are proposed. For onset detection, we adapt the ideas of [23, 24] to develop a method for combining onset estimations of different channels. The main drawback of the approach in [23] is that the selection procedure of onset candidates is based on a loudness estimation and a global threshold which makes it unsuitable for music with high dynamics. Instead, in [24] and also in our approach, relative thresholds are applied. However, the proposal in [24] can only combine synchronous onset estimations, i.e., the same sampling rate has to be used for the onset detection functions of all basic estimators. Our new approach can handle asynchronous estimations which enables the use of different hop sizes. 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 [29], we propose a classification approach for peak-picking, where each channel nominates one candidate.

In [29], potential pitch periods derived from the original signal are used as features, whereas in our approach, features need to be derived using the auditory model. Our approach is applicable to temporal autocorrelations as well as to frequency domain approaches. 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 from the auditory output. This is contrary to [34], where the features are defined across all channels. The channel-wise approach preserves more information, can be more easily adapted to the hearing-impaired variants, and enables assessments of the contribution of specific channels to the recognition rates.

All approaches are extensively investigated using a comprehensive experimental design. The experimental setup is visualized in Fig. 1. 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. For instrument recognition, the approach using the auditory model output even performs distinctly better than the approach using standard features. As a prospect of future research, performance losses based on hearing deficits are exemplified using three so-called hearing dummies as introduced in [17].

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 brainstem and cortex are difficult to model, the auditory periphery is much better investigated. This stage models the transformation from acoustical pressure waves to release events of the auditory nerve fibers. Out of the several models simulating the auditory periphery, we apply the popular and widely analyzed model of Meddis [18], for which simulated hearing profiles of real hearing impaired listeners exist [17].

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 eardrum

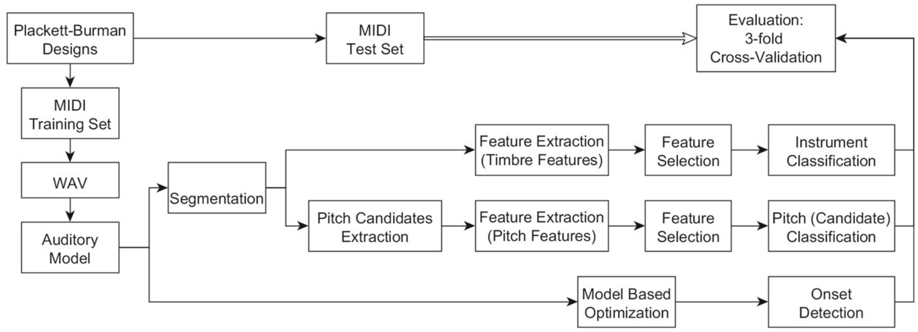
Plackett-Burman Designs
MIDI Test Set
Evaluation: 3-fold Cross-Validation
MIDI Training Set
WAV
Auditory Model
Segmentation
Feature Extraction (Timbre Features)
Feature Selection
Instrument Classification
Pitch Candidates Extraction
Feature Extraction (Pitch Features)
Feature Selection
Pitch (Candidate)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

Fig. 1 Structure of the experiments for the music recognition tasks

transmits vibrations to the stapes in the middle ear and then further to the cochlea in the inner ear. Inside the cochlea, a traveling wave deflects the basilar membrane at specific locations dependent on the stimulating frequencies. On the basilar membrane, inner hair cells are activated by the velocity of the membrane and evoke spike emissions (neuronal activity) to the auditory nerve fibers.

The auditory model of Meddis [18] 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 Fig. 2. 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 41st channel.

In the last plot of Fig. 3, an exemplary output of the model can be seen. The 41 channels are located on the vertical axis according to their best frequencies. The grayscale 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 the Meddis model corresponds to 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 [38]. 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 cell 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 details about the model equations, the reader is refered to the appendix in [18].

For the auditory model with hearing loss, we consider three examples, called "hearing-dummies," which are described in [16, 17]. These are modified versions of the Meddis auditory model. The goal of the hearing-dummies is to mimic the effect of real hearing impairments [39]. In the original proposal [17], channels with best frequencies between 250 Hz and 8 kHz are considered, whereas in the normal-hearing model described above, channel frequencies between 100 Hz and 6 kHz are used. Note that

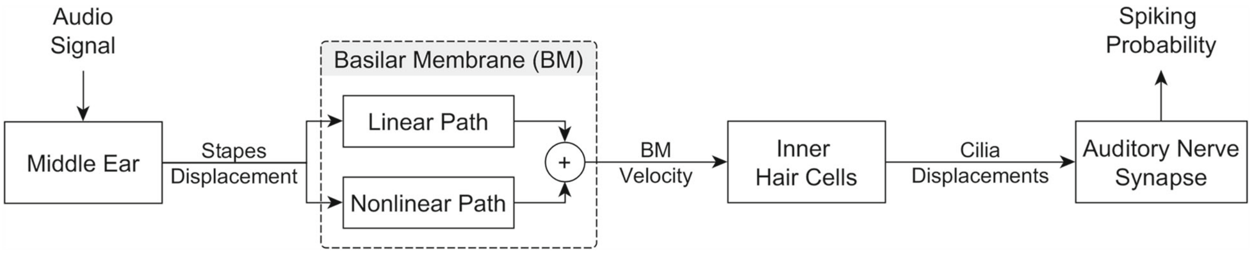

Fig. 2 Block diagram of Meddis' model of the auditory periphery

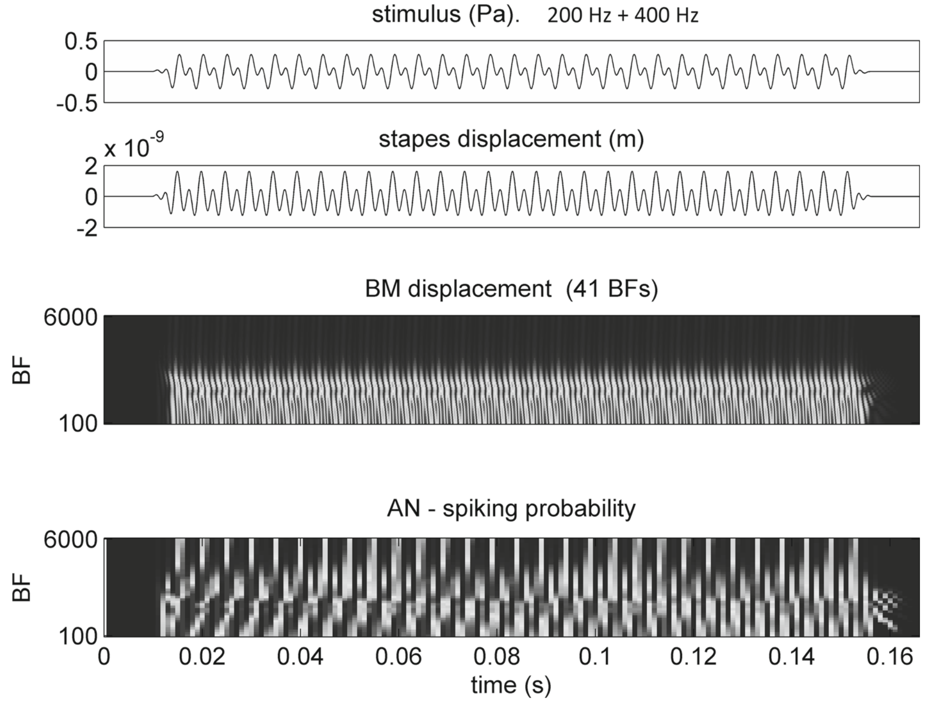

Fig. 3 Exemplary output of Meddis' model of the auditory periphery: (*1*) original signal (200 Hz + 400 Hz), (*2*) middle ear output (stapes displacement), (*3*) basilar membrane (BM) output with respect to the channels' best frequencies (BF), (*4*) auditory nerve (AN) output with respect to the BFs

this difference is just a matter of the user's interesting frequency range and not influenced by any hearing damage. For a better comparison, the same best frequencies will 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 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 $x[t]$. Second, we adapt this algorithm to the auditory model output in a channel-wise manner. Third, 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 Fig. 4. The corresponding parameters, used for the optimization, are shown in parentheses.

In the first step, the incoming signal is split into small frames with a frame size of M samples and a hop size h which is the distance in samples between the starting

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

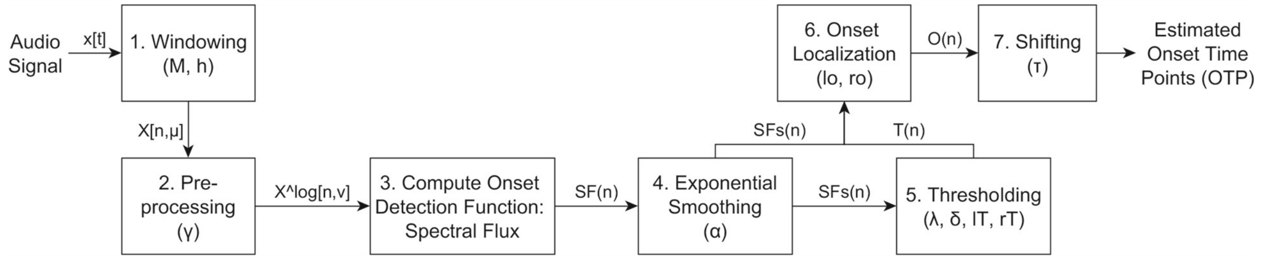
Audio Signal
x[t]
1. Windowing (M, h)
X[n,μ]
2. Pre-processing (γ)
X^log[n,v]
3. Compute Onset Detection Function: Spectral Flux
SF(n)
4. Exponential Smoothing (α)
SFs(n)
5. Thresholding (λ, δ, lT, rT)
SFs(n)
T(n)
6. Onset Localization (lo, ro)
O(n)
7. Shifting (τ)
Estimated Onset Time Points (OTP)

Fig. 4 Block diagram for classical onset detection (without auditory model)

points of subsequent frames. For each frame, the magnitude spectrum of the discrete Fourier transform (DFT) $|X[n,\mu]|$ is computed where n denotes the frame 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 [40]. 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:

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

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

Afterwards, a feature is computed in each frame (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 [41], this is shown on a music data set with 1065 onsets covering a variety of musical styles and instrumentations, and in [40], this is verified on an even larger data set with 25,966 onsets. Spectral flux describes the degree of positive spectral changes between consecutive frames and is defined as:

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

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

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

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

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

A threshold function (step 5) distinguishes between relevant and nonrelevant 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 [41]. The threshold function is defined as

$$T(n) = \delta + \lambda \cdot \mathrm{mean}(\mathrm{SF_s}(n - l_T), \ldots, \mathrm{SF_s}(n + r_T)), \quad \textit{for } n = 1, \ldots, L, \tag{5}$$

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

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

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

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

Frames 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 [42]. In the same manner, these rise times also affect the maximum value of spectral flux and other features.

Then, $\mathbf{OTP} = (\mathrm{OTP}_1, \ldots, \mathrm{OTP}_{C_{\mathrm{est}}})$ 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 which is also used in other studies [40].

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

$$F = \frac{2 \cdot m_{T^+}}{2 \cdot m_{T^+} + m_{F^+} + m_{F^-}}, \quad F \in [0, 1], \tag{7}$$

where m_{T^+} is the number of correctly detected onsets, m_{F^+} is the number of false alarms, and m_{F^-} 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 number of true onsets $C_{\text{true}} = m_{T^+} + m_{F^-}$ and the F value for the scenario where no false alarm is produced:

$$m_{F^+} = 0 \quad \Longrightarrow \quad m_{F^-} = \left(1 - \frac{F}{2-F}\right) \times C_{\text{true}}. \quad (8)$$

In a listening experiment, we would assume a relatively low number of false alarms. Hence, we will use the scenario $m_{F^+} = 0$ for a comparison to human perception.

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 (5 to 15 min on the used Linux-HPC cluster system [43]), 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 [44, 45].

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

Parameter name	Minimum value	Maximum value
Frame 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

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 $\mathbf{OTP}_k$. Now, the additional challenge arises how to combine different onset predictions of several channels. We compare two approaches. First, as a simple variant, we just consider the channel which achieves the best F-value on the training data. Second, we introduce a variant which combines the final results of all channels. This approach is illustrated in Fig. 5. Again, the 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 that fewer channels lead to a reduction of computation time, we allow the omission of the lowest and the highest channels by defining the minimum $k_{\min}$ and the maximum channel $k_{\max}$. All estimated onset time points of the remaining channels are pooled into one set of onset candidates:

$$\mathbf{OTP}_{\text{cand}} = \bigcup_{k=k_{\min}}^{k_{\max}} \mathbf{OTP}_k. \quad (9)$$

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 above a global threshold. The threshold is defined by

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

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 $[\,n - t_{\text{loc}}, \ldots, n + t_{\text{loc}}]$, 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

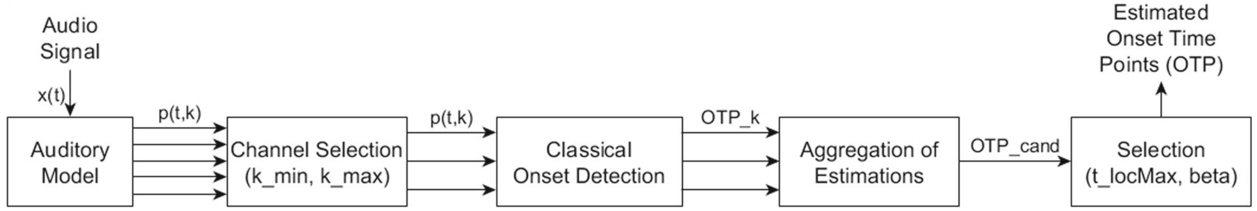
Audio Signal
x(t)
Auditory Model
p(t,k)
Channel Selection (k_min, k_max)
p(t,k)
Classical Onset Detection
OTP_k
Aggregation of Estimations
OTP_cand
Selection (t_locMax, beta)
Estimated Onset Time Points (OTP)

Fig. 5 Block diagram for the proposed approach for onset detection using an auditory model

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 a half semitone (50 cents). This is equivalent to a relative error of approximately 3% on 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 that the onsets and offsets of each melody tone are known. 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 [26] and its extension pYIN [27], which do not employ an auditory model, for comparison reasons in our experiments. The mean error rate over all tones is applied to measure the performances of the approaches, i.e., it is assumed that all tones are equally important, regardless of their length.

4.3.1 Autocorrelation approach

One challenge of autocorrelation analysis of the auditory output is again the combination of several channels. In [30, 31], 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 auditory 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 [30, 31]. 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:

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

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.

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

Parameter name	Minimum value	Maximum value
t_{locMax}	0 s	0.125 s
β	0	1
k_{min}	1	20
k_{max}	21	41

4.3.2 Spectral approach

We propose a classification method partly based on features which we introduced in [46, 47] 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 auditory 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:

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

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

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 and for 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 Fig. 6):

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

The band edges are defined by

$$\mu^*_{\mathrm{right}}[k] = \min\left(\mu \in \left\{\mu^*[k], \ldots, \frac{M}{2}\right\} : \frac{a_c[k]}{10} > |P[\mu, k]|\right), \tag{15}$$

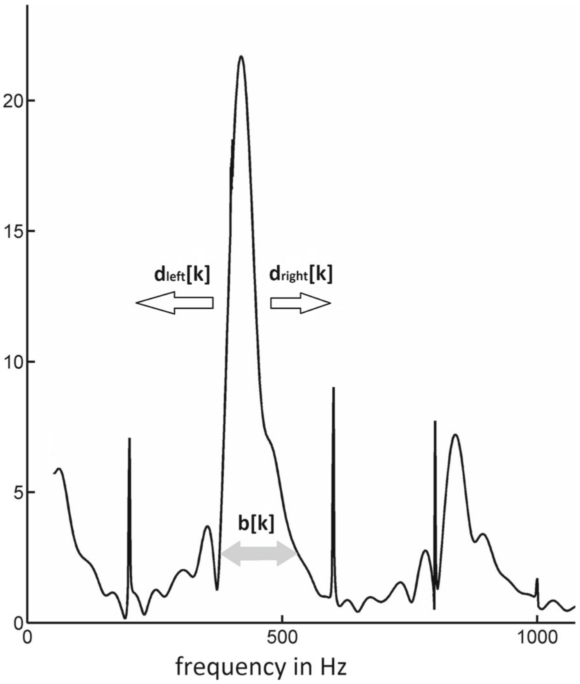

Fig. 6 Features for pitch estimation. **a** Bandwidth $b[k]$ of the candidate peak, **b** distance to maximum left $d_{\mathrm{left}}[k]$, and **c** distance to maximum right $d_{\mathrm{right}}[k]$the candidate) across

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

$$\mu^*_{left}[k] = \max\left(\mu \in \{1, \ldots, \mu^*[k]\} : \frac{a_c[k]}{10} > |P[\mu, k]|\right), \tag{16}$$

where $\mu^*_{\mathrm{left}}[k]$ 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 candidate's band edges (two features: $d_{\mathrm{left}}[k]$ and $d_{\mathrm{right}}[k]$, see also Fig. 6):

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

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

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

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

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

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

Studies have shown that humans can resolve the first 7 to 11 partials [48, 49]. Hence, the first nine partials might be beneficial for pitch estimation.

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

Altogether, this results in 29 features for each channel, i.e., $29 \times 41 = 1189$ 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.

4.4 Predominant instrument recognition

Instrument recognition is a typical supervised classification task. First, meaningful features need to be extracted, and second, a classification model is learned which maps the feature space to the instrument categories. 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 [50]. These features are common for instrument recognition based directly on the time domain waveform and are linked to timbre in the literature [51]. For our approach using an auditory model, they are computed on each of the 41 channels, thus, we obtain $41 \times 21 = 861$ features for each tone. The first 20 features are computed by means of the *MIRtoolbox* [52]. The last feature is the Shannon-Entropy:

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

where $X[\mu]$ is the DFT of a signal frame (respectively the DFT of a channel output in the auditory model variant) and $\mathrm{pr}(|X[\mu]|) = \frac{|X[\mu]|}{\sum_{\nu=1}^{M}|X[\nu]|}$ is the share of the μth frequency bin with respect to the cumulated spectral magnitudes of all bins. $H(X[\mu])$ 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).

Table 4 Features for instrument recognition [50]

Feature no.	Feature name
1	Root-mean-square energy
2	Low energy
3	Mean spectral flux (see Eq. 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

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 [53].

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., [54]). 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) [55] 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 by a hyperplane maximizing a so-called safety margin between the classes. As we cannot exclude the existence of outliers, so-called slack variables are applied, one per training point, measuring the amount of margin violation. Overall, maximization of the margin is traded 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 $\frac{G(G-1)}{2}$ 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 [56]. 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 increases 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 [57]. 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 with onset times, pitches, and the instruments being played 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 a whole entity of music. Instead, we construct a more structured data base using an experimental design based on eight musical factors which might have an influence on music intelligibility. This enables identification of the most problematic music w.r.t. classification performance. We apply Plackett-Burman (PB) designs which require just two levels for each factor [58]. After all experiments (music samples) are evaluated, a linear regression model is fitted to predict the target variables w.r.t. the factor levels. For onset detection, the target variable is the F measure; for instrument recognition, it is the classification error rate; and for pitch estimation, it is the mean error rate using a tolerance of a half semitone. If no factor has a significant influence on the target variable, we can assume that the approach works equally well for all considered music pieces. 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 R^2 also depends on the number of factors, adjusted R^2 are used to compensate this effect [59]:

$$R_a^2 = 1 - \frac{n_{\text{exp}} - 1}{n_{\text{exp}} - p_{\text{fac}} - 1}(1 - R^2), \tag{22}$$

where n_{exp} is the number of experiments and p_{fac} is the number of factors [60].

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 each of 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 as follows:

- *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 common 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 absolute 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 factor tone duration by the duration of the song extracts in order to maintain 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 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.
- *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 semitones). The two considered levels are defined by the intervals $[-6,6]$ and $[-24, -12]$ measured in semitones. 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 three times, changing the melody instrument each time. Overall, this results in $3 \times 12 \times 30 = 1080$ 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 $n_{\text{exp}} = 3 \times 12 = 36$ experiments altogether. To ensure that the accompaniment is not louder than the melody, we use a melody to accompaniment ratio of 5 dB.

5.2 Structure of the comparison experiments

At first, the approaches described in the previous section are compared using the original auditory model without a simulated hearing loss. The structure of the whole process is illustrated in Fig. 1.

For all experiments, threefold 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

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:	$[-6, 6]$	$[12, 24]$
melody − accompaniment	Half tones	Half tones

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, features are extracted from the auditory model or the original signal which results in four considered variants altogether. For pitch estimation, eleven approaches are compared: four classification approaches with auditory features—RF or SVM combined with DFT or ACF features—(Section 4.3.1), two peak-picking variants for the SACF approach (Section 4.3.2) and five variants of the YIN (respectively pYIN) algorithm as the state-of-the-art approaches without an auditory model.

For the YIN algorithm, standard settings are used, except for the lower and the upper limits of the search range which are set to 155 and 1109 Hz, respectively. These values corresponds to the pitch range of the melody in the considered song extracts. In contrast to the other tested approaches, the YIN and the pYIN algorithms estimate the fundamental frequency for short frames and not for complete tones. Hence, an aggregation mechanism is needed to aggregate the fundamental frequency estimations of several frames into one estimation for the complete tone. For the YIN algorithm, three aggregation approaches are tested. The first method selects the estimation of the frame which has the smallest “aperiodic power” component which might be an indicator for the estimation uncertainty [26]. The second method returns the median estimation. This method is also tested for the pYIN algorithms which, however, often wrongly estimates a higher partial. This effect might be rectified by lower quantiles, and hence, we also test the 10% quantile for YIN and pYIN.

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. This means also the optimization stage of onset detection and the training stages of pitch and instrument recognition are conducted separately for each hearing model.

5.3 Software

For classification, the R package *mlr* [61] is applied using the package *randomForest* [62] for RFs and the package *kernlab* [63] for SVMs. MBO is performed by using the R package *mlrMBO* [64]. Finally, the huge number of experiments performed is managed by the R packages *BatchJobs* and *BatchExperiments* [65].

6 Results

First, we present the main results regarding the normal hearing auditory model 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.

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),

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

(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 Fig. 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 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 when it is disturbed 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 three instruments, can be seen. In the monophonic variant, the adjusted R^2 (R_a^2) is negative, which indicates that the performance is independent of the type of music. This is also supported by the p values, since none of them shows a significant impact. 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 conducting 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 $\left(R_a^2 > 0.5\right)$—note that we describe music pieces by just eight dimensions which explains the relatively high amount of noise in all evaluation models of the experimental design. Nevertheless, 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 solely produced by the accompaniment is low. Obviously, this was expected since false alarms caused by tones of the accompaniment are decreased in that case. 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

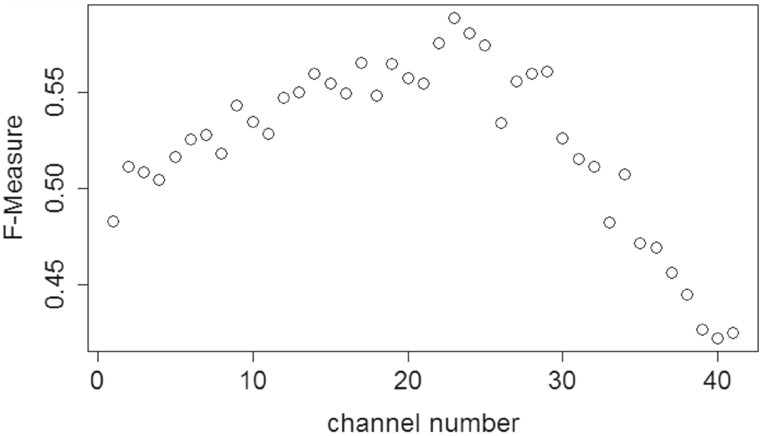

Fig. 7 Results (mean F measure) for predominant onset detection using the output of just one channel

Table 7 Evaluation over all instruments and all Plackett-Burman designs for the proposed aggregation approach (factors as in Table 5)

	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

The average *F* value is the target variable—a: monophonic onset detection, b: predominant onset detection, and c: polyphonic onset detection (*italics* = significant at 10% level)

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 [66] 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 particularly the case 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 [66], 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 in order to emulate 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, the results for the monophonic data are listed. Out of the approaches with auditory model, our approach using spectral features of the auditory output and a linear SVM for classification performs best with a mean error rate of 7% in the polyphonic and 2% in the monophonic case. The YIN algorithm with median aggregation performs even better with a mean error rate of only 3% for polyphonic music, whereas the "aperiodic power" variant performs extremely poor. The pYIN algorithm performs worse than the YIN algorithm. Reason for this is that it more often confuses the frequency of a higher partial with the fundamental frequency. Paradoxically, this effect occurs even more often in the monophonic variant. Contrary, all other approaches perform as expected clearly better in the monophonic variant than in the polyphonic one. Applying the 10% quantile instead of the median decreases the confusion of the higher partials and improves the performance which, nevertheless,

Table 8 Mean error rates of pitch detection methods (italics indicates the best results for the mean performance)

	Polyphonic/predominant				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 + aperiodic power	0.36	0.15	0.32	0.28	0.05
YIN + median	0.04	0.01	0.04	*0.03*	0.00
YIN + 10% quantile	0.11	0.04	0.09	0.08	0.03
pYIN + median	0.04	0.17	0.07	0.09	0.12
pYIN + 10% quantile	0.04	0.10	0.01	0.05	0.04

is distinctly worse than the YIN algorithm with median aggregation.

Interestingly, clarinet tones are the main challenge for pYIN, whereas for all other approaches, the error rates for clarinet are the lowest and cello tones seem to be the most difficult ones. In general, the pitch of clarinet tones is easier to estimate because these tones have a relatively low intensity of the even partials which might prevent octave errors. For trumpet and cello tones, often the frequency of the second partial is confused with the fundamental frequency. Again, pitches of cello tones which are accompanied by string instruments are especially difficult to estimate.

For the best method with auditory model—the classification approach using spectral features and either linear SVM or RF—group-based feature selection is performed (as introduced in Section 4.5.3. 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 two 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 $d_{\text{right}}[k]$ are picked, whereas for SVM, $p_{\text{mean}}[k]$ and $P_1^{\text{mean}}[k]$ are chosen. In the backward variant, the SVM just needs the following nine feature groups to achieve the same error rate as with all features: $c[k]$, $p_{\text{mean}}[k]$, $p_{\text{max}}[k]$, $b[k]$, $d_{\text{left}}[k]$, $d_{\text{right}}[k]$, $P_4^{\text{mean}}[k]$, $P_8^{\text{mean}}[k]$, and $P_9^{\text{mean}}[k]$. 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 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 $\left(R_a^2 = 0.12\right)$ 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 is expected 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 [66, 67] 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 [68]. 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% [69]. 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 [67]. In fact, also humans can perceive absolute pitch. 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 [70].

In conclusion, comparing the results of our study to human data constitutes a big challenge. We can assume that a normal-hearing listener might be able to perceive relative pitches almost perfectly w.r.t. the tolerance of $\frac{1}{2}$ semitone at least in the monophonic case. This estimation approximately corresponds to the result of the classification method with DFT features which yields a mean error rate of 2% in our study. The human ability for the perception of polyphonic music has not yet been adequately researched to make a comparison. Hence, in future studies, extensive listening tests are necessary.

6.1.3 Instrument recognition

The error rates for instrument recognition 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

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 × 29 = 1189	4 × 29 = 116	35 × 29 = 1015	41 × 2 = 82	41 × 28 = 1148
RF: error rate	0.08	0.10	0.07	0.09	0.08
SVM: number of features	41 × 29 = 1189	5 × 29 = 145	23 × 29 = 667	41 × 2 = 82	41 × 9 = 369
SVM: error rate	0.07	0.10	0.07	0.09	0.07

Table 10 Evaluation over all instruments and all Plackett-Burman Designs (factors as in Table 5)

	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*

The error rate is the target variable— a: pitch estimation and SVM (auditory model + DFT), b: instrument recognition and SVM (auditory model features)—(*italics* = significant at 10% level)

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 good 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 and 1047 Hz, corresponding to the channels 6 to 24 which have best frequencies between 167 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 Fig. 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 first, the third, the fourth, and the seventh MFCC coefficients.

Table 10b shows the evaluation of the experimental design for predominant instrument recognition. Here, the goodness of fit is moderate $\left(R_a^2 = 0.21\right)$ 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 tasks are very similar, there exists one important difference which causes diverging results of the two disciplines. Dissimilar ratings are subjective measures which rely on judgements of humans, whereas instrument recognition is a well-defined task [51]. Nevertheless, also, some studies have conducted experiments about the human ability to distinguish music instruments (see [71] for a tabular overview). The most comprehensive experiment is reported in [72], a listening experiment with music experts. The subjects had to distinguish isolated notes of

Table 11 Mean error rates of instrument recognition methods (italics indicates the best result for the overall performance)

	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	$41 \times 21 = 861$	$2 \times 21 = 42$	$12 \times 21 = 420$	$41 \times 3 = 123$	$41 \times 17 = 697$
RF error rate	0.019	0.034	0.011	0.058	0.016
SVM number of features	$41 \times 21 = 861$	$2 \times 21 = 42$	$12 \times 21 = 420$	$41 \times 3 = 123$	$41 \times 8 = 328$
SVM error rate	0.011	0.030	0.017	0.045	0.015

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 recognition 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 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 caution due to the overall relative poor results of automatic onset detection. However, for a strong hearing loss such as HD1, it 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 worse than the normal hearing model. 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 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 [73].

Also for instrument recognition, the results of the hearing dummies are worse than the result of the normal hearing model as can be seen in Table 15. In contrast to pitch estimation, this time HD2 performs better than HD3, since here, higher channels are the most relevant ones as we have already seen in Fig. 8.

Comparison to human perception In contrast to the case of normal hearing, here the comparison to existing listening experiments is even more challenging since the

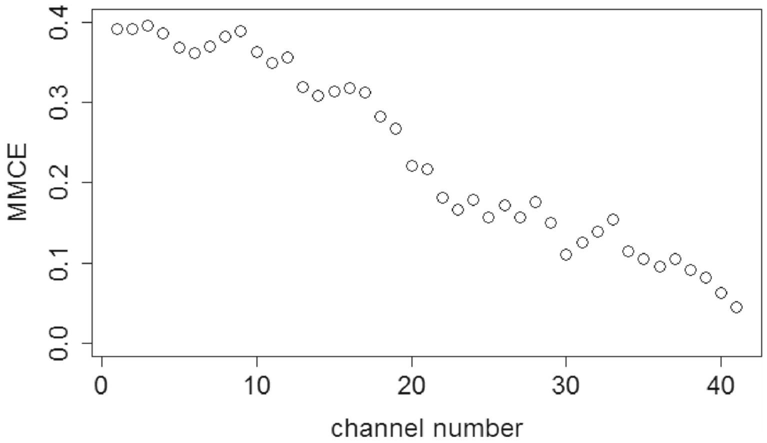

Fig. 8 Mean misclassification error (MMCE) for predominant instrument recognition using the features of just one channel and the linear SVM

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

recognition rates of hearing-impaired listeners strongly depends on their individual hearing deficits. In existing listening experiments, hearing-impaired people are only roughly grouped into two or three categories due to the typically low number of test persons. Hence, there exist no studies which consider the ability of music perception for hearing-impaired listeners which are affected by the exact deficits as the modeled hearing dummies. Nevertheless, the performance results of this study are consistent to the results of listening experiments.

It is reported that rhythm perception is not much affected by moderate hearing impairments [74], to which group HD2 and HD3 belong. For fundamental frequency discrimination, it is well established that hearing loss adversely affects the perception ability [75]. The results of pitch estimation in this study confirm this effect for all hearing dummies. In [3], the ability to discriminate musical instruments is tested with respect to the degree of hearing loss. Listeners with a high hearing loss can discriminate instruments worse than NH listeners. In this study, the instrument recognition performance is worse for all hearing dummies compared to the results of the model with normal hearing. However, there are two differences in our experiments compared to the listening experiments in [3]: First, no amplification is considered, and second, polyphonic music is considered which is particularly challenging for hearing-impaired listeners.

Apart from the degree of hearing loss, the human ability to perceive music also depends on the level of musical experience. For example, human pitch perception is supported by expectations which notes are likely to occur, and trained musicians perform better on pitch tasks than nonmusicians [67]. Such cognitive factors are not considered in the presented simulation model.

Table 14 Mean error rates of pitch detection methods for hearing dummies (HD) compared to the normal hearing (NH) model (*italics* indicates the best result for each dummy)

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

7 Conclusions

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. For instrument recognition, the performance of the proposed approach is even 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 that these results convey information about the level of difficulty for tone onset recognition. Also, for a human listener, tone onsets of a trumpet are easier to identify than onsets of a cello. Another strategy is just analyzing 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 performance 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.

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

Table 15 Mean error rates of instrument recognition methods for hearing dummies (HD) compared to the normal hearing (NH) model (*italics* indicates the best result for each dummy)

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*

performs best with an error rate of 7%. However, further improvements might be possible since the YIN algorithm, the state-of-the-art method which does not employ an auditory model, performs even better with an error rate of only 3%.

For the classification approach, 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 equally acceptable error rates as the full model. 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 ($P_{\mathrm{pl}}^{\mathrm{mean}}[k]$) seem to be more meaningful than the features corresponding to the maximum amplitude ($P_{\mathrm{pl}}^{\mathrm{max}}[k]$). However, most of the ($P_{\mathrm{pl}}^{\mathrm{mean}}[k]$) 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 auditory channels. For the RF, 12 auditory channels are sufficient to achieve the best error rate. Since the standard 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. However, this hypothesis needs to be verified in future studies focusing solely on instrument recognition with larger taxonomies of instruments.

For all three considered hearing dummies, the error rates increase for all classification tasks. The degree of performance loss 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 future work, we want to verify the results of the simulation experiments by a listening test. Such test is also necessary for the investigation how to combine the individual error measures into one overall measure for music intelligibility. This measure could be applied for assessing and optimizing hearing instruments for music with several parameters to adjust.

Acknowledgements
This 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).

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Statistics, TU Dortmund University, 44221 Dortmund, Germany. [2]Institute of Communication Acoustics, Ruhr-Universität Bochum, Building ID 2/231, 44780 Bochum, Germany.

Received: 9 June 2016 Accepted: 13 February 2017
Published online: 02 March 2017

References

1. HJ McDermott, Music perception with cochlear implants: a review. Trends Amplification. **8**(2), 49–82 (2004). doi:10.1177/108471380400800203
2. KE Gfeller, C Olszewski, C Turner, B Gantz, J Oleson, Music perception with cochlear implants and residual hearing. Audiology and Neurotology. **11**(Suppl. 1), 12–15 (2006). doi:10.1159/000095608
3. S Emiroglu, B Kollmeier, Timbre discrimination in normal-hearing and hearing-impaired listeners under different noise conditions. Brain Res. **1220**, 199–207 (2008). doi:10.1016/j.brainres.2007.08.067
4. K Fitz, M Burk, M McKinney, in *Proceedings of Meetings on Acoustics*. Multidimensional perceptual scaling of musical timbre by hearing-impaired listeners, vol. 6 (Acoustical Society of America, 2009). doi:10.1121/1.3186749
5. T Jürgens, SD Ewert, B Kollmeier, T Brand, Prediction of consonant recognition in quiet for listeners with normal and impaired hearing using an auditory model. J Acoust Soc Am. **135**(3), 1506–1517 (2014). doi:10.1121/1.4864293
6. M Karbasi, D Kolossa, in *Proc. Annual Meeting of the German Acoustical Society (DAGA)*. A microscopic approach to speech intelligibility prediction using auditory models (German Acoustical Society, Berlin, 2015)
7. M Chasin, FA Russo, Hearing aids and music. Trends Amplification. **8**(2), 35–47 (2004)
8. K Fitz, M McKinney, in *Proceedings of Meetings on Acoustics*. Music through hearing aids: perception and modeling, vol. 9 (Acoustical Society of America, 2015). doi:10.1121/1.3436580
9. HK Maganti, M Matassoni, Auditory processing-based features for improving speech recognition in adverse acoustic conditions. EURASIP J Audio Speech Music Process. **2014**(1), 1–9 (2014). doi:10.1186/1687-4722-2014-21
10. A Klapuri, Multipitch analysis of polyphonic music and speech signals using an auditory model. Audio Speech Lang Process IEEE Trans. **16**(2), 255–266 (2008). doi:10.1109/TASL.2007.908129
11. P McLeod, Fast, accurate pitch detection tools for music analysis. PhD Thesis (2009)
12. MG Heinz, X Zhang, IC Bruce, LH Carney, Auditory nerve model for predicting performance limits of normal and impaired listeners. Acoust Res Lett Online. **2**(3), 91–96 (2001). doi:10.1121/1.1387155
13. MS Zilany, IC Bruce, Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery. J Acoust Soc Am. **120**(3), 1446–1466 (2006). doi:10.1121/1.2225512
14. ML Jepsen, T Dau, Characterizing auditory processing and perception in individual listeners with sensorineural hearing loss. J Acoust Soc Am. **129**(1), 262–281 (2011). doi:10.1121/1.3518768
15. ML Jepsen, T Dau, O Ghitza, Refining a model of hearing impairment using speech psychophysics. J Acoust Soc Am. **135**(4), 179–185 (2014). doi:10.1121/1.4869256
16. R Meddis, W Lecluyse, CM Tan, MR Panda, R Ferry, Beyond the audiogram: identifying and modeling patterns of hearing deficits, 631–640 (2010). doi:10.1007/978-1-4419-5686-6_57

17. MR Panda, W Lecluyse, CM Tan, T Jürgens, R Meddis, Hearing dummies: individualized computer models of hearing impairment. Int J Audiol. **53**(10), 699–709 (2014). doi:10.3109/14992027.2014.917206
18. R Meddis, Auditory-nerve first-spike latency and auditory absolute threshold: a computer model. J Acoust Soc Am. **119**(1), 406–417 (2006). doi:10.1121/1.2139628
19. J Salamon, E Gómez, Melody extraction from polyphonic music signals using pitch contour characteristics. Audio Speech Lang Process IEEE Trans. **20**(6), 1759–1770 (2012). doi:10.1109/TASL.2012.2188515
20. J Schluter, S Bock, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Improved musical onset detection with convolutional neural networks (IEEE, 2014), pp. 6979–6983. doi:10.1109/ICASSP.2014.6854953
21. JP Bello, L Daudet, S Abdallah, C Duxbury, M Davies, MB Sandler, A tutorial on onset detection in music signals. IEEE Trans Speech Audio Process. **13**(5), 1035–1047 (2005). doi:10.1109/TSA.2005.851998
22. N Bauer, K Friedrichs, B Bischl, C Weihs, in *Analysis of Large and Complex Data*, ed. by FX Adalbert, HAK Wilhelm. Fast model based optimization of tone onset detection by instance sampling (Springer, Bremen, 2016)
23. A Klapuri, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. Sound onset detection by applying psychoacoustic knowledge, vol. 6 (IEEE, 1999), pp. 3089–3092. doi:10.1109/ICASSP.1999.757494
24. A Holzapfel, Y Stylianou, AC Gedik, B Bozkurt, Three dimensions of pitched instrument onset detection. IEEE Trans Audio Speech Lang Process. **18**(6), 1517–1527 (2010). doi:10.1109/TASL.2009.2036298
25. N Bauer, K Friedrichs, D Kirchhoff, J Schiffner, C Weihs, in *Data Analysis, Machine Learning and Knowledge Discovery*, ed. by M Spiliopoulou, L Schmidt-Thieme, and R Janning. Tone onset detection using an auditory model, vol. Part VI (Springer, Hildesheim, 2014), pp. 315–324. doi:10.1007/978-3-319-01595-8_34
26. De Cheveigné, H Kawahara, Yin, a fundamental frequency estimator for speech and music. J Acoust Soc Am. **111**(4), 1917–1930 (2002). doi:10.1121/1.1458024
27. M Mauch, S Dixon, in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pyin: A fundamental frequency estimator using probabilistic threshold distributions (IEEE, 2014), pp. 659–663
28. Z Duan, B Pardo, C Zhang, Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions. IEEE Trans Audio Speech Lang Process. **18**(8), 2121–2133 (2010). doi:10.1109/TASL.2010.2042119
29. A Klapuri, in *6th Sound and Music Computing Conference, Porto, Portugal*. A classification approach to multipitch analysis (Sound and Music Computing research community, Porto, 2009)
30. R Meddis, MJ Hewitt, Virtual pitch and phase sensitivity of a computer model of the auditory periphery. i: Pitch identification. J Acoust Soc Am. **89**(6), 2866–2882 (1991). doi:10.1121/1.400725
31. R Meddis, L O'Mard, A unitary model of pitch perception. J Acoust Soc Am. **102**(3), 1811–1820 (1997). doi:10.1121/1.420088
32. M Goto, A real-time music-scene-description system: predominant-F0 estimation for detecting melody and bass lines in real-world audio signals. Speech Commun. **43**(4), 311–329 (2004). doi:10.1016/j.specom.2004.07.001
33. T Sandrock, Multi-label feature selection with application to musical instrument recognition. PhD thesis (2013)
34. KD Martin, YE Kim, Musical instrument identification: a pattern-recognition approach. J Acoust Soc Am. **104**(3), 1768–1768 (1998). doi:10.1121/1.424083
35. K Patil, M Elhilali, Biomimetic spectro-temporal features for music instrument recognition in isolated notes and solo phrases. EURASIP J Audio Speech Music Process. **2015**(1), 1–13 (2015). doi:10.1186/s13636-015-0070-9
36. A Wieczorkowska, E Kubera, A Kubik-Komar, Analysis of recognition of a musical instrument in sound mixes using support vector machines. Fundamenta Informaticae. **107**(1), 85–104 (2011)
37. JJ Bosch, J Janer, F Fuhrmann, P Herrera, in *ISMIR*. A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals (FEUP Edições, Porto, 2012), pp. 559–564
38. EA Lopez-Poveda, R Meddis, A human nonlinear cochlear filterbank. J Acoust Soc Am. **110**(6), 3107–3118 (2001). doi:10.1121/1.1416197
39. T Jürgens, NR Clark, W Lecluyse, R Meddis, Exploration of a physiologically-inspired hearing-aid algorithm using a computer model mimicking impaired hearing. Int J Audiol. **55**, 346–357 (2016). doi:10.3109/14992027.2015.1135352
40. S Böck, F Krebs, M Schedl, in *ISMIR*. Evaluating the online capabilities of onset detection methods (FEUP Edições, Porto, 2012), pp. 49–54
41. C Rosao, R Ribeiro, DM De Matos, in *ISMIR*. Influence of peak selection methods on onset detection (FEUP Edições, Porto, 2012), pp. 517–522
42. J Vos, R Rasch, The perceptual onset of musical tones. Percept Psychophys. **29**(4), 323–335 (1981)
43. High Performance Computer-Cluster LiDOng (2016). http://www.dowir.de/dowirforum. Accessed 2016
44. DR Jones, M Schonlau, WJ Welch, Efficient global optimization of expensive black-box functions. J Glob Optim. **13**(4), 455–492 (1998). doi:10.1023/A:1008306431147
45. B Bischl, S Wessing, N Bauer, K Friedrichs, C Weihs, in *Learning and Intelligent Optimization*. MOI-MBO: multiobjective infill for parallel model-based optimization (Springer, Gainesville, 2014), pp. 173–186. doi:10.1007/978-3-319-09584-4_17
46. K Friedrichs, C Weihs, in *Classification and Data Mining*. Auralization of auditory models (Springer, Florence, 2013), pp. 225–232. doi:10.1007/978-3-642-28894-4_27
47. C Weihs, K Friedrichs, B Bischl, in *Second Bilateral German-Polish Symposium on Data Analysis and Its Applications (GPSDAA)*. Statistics for hearing aids: Auralization (Uniwersytet Ekonomiczny, Cracow, 2012), pp. 183–196
48. R Plomp, The ear as a frequency analyzer. J Acoust Soc Am. **36**(9), 1628–1636 (1964). doi:10.1121/1.1919256
49. JG Bernstein, AJ Oxenham, Pitch discrimination of diotic and dichotic tone complexes: harmonic resolvability or harmonic number J Acoust Soc Am. **113**(6), 3323–3334 (2003). doi:10.1121/1.1572146
50. K Friedrichs, C Weihs, Comparing timbre estimation using auditory models with and without hearing loss. Technical Report 51/2012 (2012). doi:10.17877/DE290R-10355
51. K Siedenburg, I Fujinaga, S McAdams, A comparison of approaches to timbre descriptors in music information retrieval and music psychology. J New Music Res. **45**(1), 27–41 (2016)
52. O Lartillot, P Toiviainen, in *Proc. of the 10th Int. Conference on Digital Audio Effects (DAFx-07)*. A MATLAB toolbox for musical feature extraction from audio (DAFx, Bordeaux, 2007), pp. 1–8
53. L Breiman, Bagging predictors. Mach Learn. **24**(2), 123–140 (1996). doi:10.1007/BF00058655
54. L Breiman, Random forests. Mach Learn J. **45**(1), 5–32 (2001). doi:10.1023/A:1010933404324
55. V Vapnik, *Statistical Learning Theory*. (John Wiley and Sons, USA, 1998)
56. R Kohavi, GH John, Wrappers for feature subset selection. Artif Intell. **97**(1), 273–324 (1997). doi:10.1016/S0004-3702(97)00043-X
57. JH Jensen, MG Christensen, SH Jensen, in *Proc. European Signal Processing Conf*. A framework for analysis of music similarity measures (European Association for Signal Processing, Poznan, 2007), pp. 926–930
58. RL Plackett, JP Burman, The design of optimum multifactorial experiments. Biometrika, 305–325 (1946). doi:10.2307/2332195
59. L Fahrmeir, T Kneib, S Lang, *Regression: Modelle, Methoden und Anwendungen*. (Springer, Berlin Heidelberg, 2007)
60. P Yin, X Fan, Estimating R2 shrinkage in multiple regression: a comparison of different analytical methods. J Exp Educ. **69**(2), 203–224 (2001)
61. B Bischl, M Lang, J Richter, J Bossek, L Judt, T Kuehn, E Studerus, L Kotthoff, Mlr: Machine Learning in R.R package version 2.5. (2016). https://github.com/mlr-org/mlr. Accessed 2016
62. A Liaw, M Wiener, Classification and regression by randomforest. R News. **2**(3), 18–22 (2002)
63. A Karatzoglou, A Smola, K Hornik, A Zeileis, kernlab—an S4 package for kernel methods in R. J Stat Softw. **11**(9), 1–20 (2004). doi:10.18637/jss.v011.i09
64. B Bischl, J Bossek, D Horn, M Lang, mlrMBO: Model-Based Optimization for Mlr. R package version 1.0 (2016). https://github.com/berndbischl/mlrMBO. Accessed 2016
65. B Bischl, M Lang, O Mersmann, J Rahnenführer, C Weihs, BatchJobs and BatchExperiments: abstraction mechanisms for using R in batch environments. J Stat Softw. **64**(11), 1–25 (2015). doi:10.18637/jss.v064.i11
66. CL Krumhansl, Rhythm and pitch in music cognition. Psychol Bull. **126**(1), 159 (2000)

67. JH McDermott, AJ Oxenham, Music perception, pitch, and the auditory system. Curr Opin Neurobiol. **18**(4), 452–463 (2008)
68. C Wier, W Jesteadt, D Green, Frequency discrimination as a function of frequency and sensation level. J Acoust Soc Am. **61**(1), 178–184 (1977). doi:10.1121/1.381251
69. EM Burns, WD Ward, Categorical perception–phenomenon or epiphenomenon: evidence from experiments in the perception of melodic musical intervals. J Acoust Soc Am. **63**(2), 456–68 (1978)
70. DJ Levitin, SE Rogers, Absolute pitch: perception, coding, and controversies. Trends Cogn Sci. **9**(1), 26–33 (2005)
71. JC Brown, O Houix, S McAdams, Feature dependence in the automatic identification of musical woodwind instruments. J Acoust Soc Am. **109**(3), 1064–1072 (2001)
72. KD Martin, Sound-source recognition: a theory and computational model. PhD thesis (1999)
73. AJ Oxenham, Pitch perception and auditory stream segregation: implications for hearing loss and cochlear implants. Trends Amplification. **12**(4), 316–331 (2008)
74. V Looi, H McDermott, C McKay, L Hickson, Music perception of cochlear implant users compared with that of hearing aid users. Ear Hearing. **29**(3), 421–434 (2008). doi:10.1097/AUD.0b013e31816a0d0b
75. BC Moore, BR Glasberg, The effect of hearing loss on the resolution of partials and fundamental frequency discrimination. J Acoust Soc Am. **130**(5), 2891–2901 (2011). doi:10.1121/1.3640852