



<http://www.diva-portal.org>

This is the published version of a paper published in *Journal of the Acoustical Society of America*.

Citation for the original published paper (version of record):

Elowsson, A. (2017)

Predicting the perception of performed dynamics in music audio with ensemble learning.

*Journal of the Acoustical Society of America*, 141(3): 2224-2242

<https://doi.org/10.1121/1.4978245>

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-204657>

# Predicting the perception of performed dynamics in music audio with ensemble learning

Anders Elowsson<sup>a)</sup> and Anders Friberg

*KTH Royal Institute of Technology, School of Computer Science and Communication, Speech, Music and Hearing, Stockholm, Sweden*

(Received 7 April 2016; revised 15 February 2017; accepted 17 February 2017; published online 30 March 2017)

By varying the dynamics in a musical performance, the musician can convey structure and different expressions. Spectral properties of most musical instruments change in a complex way with the performed dynamics, but dedicated audio features for modeling the parameter are lacking. In this study, feature extraction methods were developed to capture relevant attributes related to spectral characteristics and spectral fluctuations, the latter through a sectional spectral flux. Previously, ground truths ratings of performed dynamics had been collected by asking listeners to rate how soft/loud the musicians played in a set of audio files. The ratings, averaged over subjects, were used to train three different machine learning models, using the audio features developed for the study as input. The highest result was produced from an ensemble of multilayer perceptrons with an  $R^2$  of 0.84. This result seems to be close to the upper bound, given the estimated uncertainty of the ground truth data. The result is well above that of individual human listeners of the previous listening experiment, and on par with the performance achieved from the average rating of six listeners. Features were analyzed with a factorial design, which highlighted the importance of source separation in the feature extraction. © 2017 Acoustical Society of America.

[<http://dx.doi.org/10.1121/1.4978245>]

[JFL]

Pages: 2224–2242

## I. INTRODUCTION

### A. Performed dynamics

By varying the dynamics in a musical performance, the musician can accentuate or soften different parts of the musical score. This is an effective way to convey structure and expression to the listener. The listener's perception of the overall dynamics in the performance should thus be an important perceptual feature to model and predict. Dynamics is a broad concept, which comprises many different aspects of music (Berndt and Hähnel, 2010). This study is focused on *performed dynamics*. What do we mean when we refer to performed dynamics? While the term *dynamics* is used in different musical contexts, we wish to directly refer to the musical *performance*, thus distinguishing the term from other similar meanings. For example, the audio engineer can control the dynamics by varying the level of compression on individual instruments or in the main mix; and researchers can, e.g., investigate the effect of dynamical changes in intensity or tempo in music. These aspects of dynamics are not the subjects of this study, as we are solely referring to the force, or energy, of the musicians when they play their instruments. In this context, *dynamics markings* are used in traditional music notation to inform the musician about the desired performed dynamics (e.g., *piano*, *fortissimo*).

Performed dynamics is controlled differently on different instruments. For example, on the violin it is accomplished by varying bow velocity, bow pressure, and bow position, while on the piano it is ultimately the velocity of

the hammer as it reaches the string that controls the resulting performed dynamics. In most acoustic instruments, the sound level, the timbre, and the onset character (e.g., onset velocity) all change in a rather complex way with varying performed dynamics (Luce and Clark, 1967; Fastl and Zwicker, 2006; Fabiani and Friberg, 2011). Generally, however, many musical instruments, as well as the voice, will produce a spectrum with more high frequency content when they are played with a higher performed dynamics (Luce, 1975; Fabiani, 2009). These variations of acoustical parameters make it possible for humans to deduce the performed dynamics regardless of listening volume when listening to recordings of acoustic instruments (Nakamura, 1987). This means that, given a set of sound examples where the listening level is normalized, listeners can still distinguish between songs with a low or high performed dynamics.

The relationship between performed dynamics and loudness becomes evident when measuring changes in loudness within the same piece of music, or changes in loudness for isolated tones from the same instrument. In one such study, Geringer (1995) explored the perceived loudness changes in musical pieces containing a crescendo or a decrescendo. In a similar study of Baroque music (Berndt and Hähnel, 2010), dynamic transitions were connected to loudness changes in the audio recordings. Dynamic transitions have also been detected by modeling loudness changes in Chopin's Mazurkas (Kosta *et al.*, 2015). In a study of isolated notes from the clarinet, flute, piano, trumpet, and violin (Fabiani and Friberg, 2011), it was found that both timbre and sound level influenced the perception of performed dynamics with about equally large effects. Here the pitch also influenced the perceived performed dynamics in most of the

<sup>a)</sup>Electronic mail: elov@kth.se

investigated instruments. A recent study mapping MIDI (Musical Instrument Digital Interface) velocities to dynamic markings also found an interaction with pitch and note duration in the model (Chacón and Grachten, 2015). This is perhaps not surprising, as it has been found that instruments playing the same dynamic markings will change in loudness depending on pitch (Clark and Luce, 1965).

## B. Loudness, timbre, and intensity

Loudness and timbre are closely related to performed dynamics. Loudness is a psychological measurement of sound strength, which is functionally related to sound pressure level, frequency distribution, and duration (Olson, 1972). It is known from previous studies (Luce and Clark, 1967; Geringer, 1995; Berndt and Hähnel, 2010; Fabiani and Friberg, 2011), that an increase in loudness can be related to an increase in performed dynamics. Furthermore, loudness has been explicitly modeled in, e.g., time-varying loudness models, such as those by Glasberg and Moore (2002) and Chalupper and Fastl (2002). Could these methods then be used to estimate the average performed dynamics of a music audio file? Unfortunately, this is not the case. When music is recorded, mixed, and mastered, the relationship between the loudness of a performance and the loudness of that performance in the musical mixture is removed. A model of perceived performed dynamics, applicable to a set of recorded multi-instrumental music examples, must therefore be invariant with regards to the sound level (and ultimately loudness) of these recordings. Although models that solely try to detect *relative* performed dynamics (i.e., trying to detect dynamic transitions or crescendos and decrescendos) have used loudness as the independent variable (Berndt and Hähnel, 2010; Kosta *et al.*, 2015), the task of estimating the average perceived performed dynamics of a musical excerpt (ME) involves the complexity of mapping a music audio file to an *absolute* target value, regardless of the sound level of the audio file. Additionally, the added complexity of multi-instrumental music examples further blurs the relationship between loudness and performed dynamics.

Although the connection to the real sound level that each instrument generates is lost when music is recorded, the timbres of the instruments are still captured in the recording. Therefore, we are in this study focusing mainly on *timbre* and *timbre variations* related to performed dynamics. There has been a considerable amount of research about timbre, where researchers have tried to model its main characteristic features. Such studies have not been fully successful, potentially because timbre can be described as a negation; the attribute of a tone which is not pitch, loudness, or duration (Hajda *et al.*, 1997). However, in similarity-ratings of different isolated musical tones, using a three-dimensional multi-dimensional scaling solution, the obtained factors can be interpreted to be related to the attack (e.g., the rise time), the spectral characteristics, and the spectral changes of the tones (e.g., MacAdams *et al.*, 1995). Thus, relevant features for the prediction of performed dynamics should arguably describe spectral properties as well as spectral changes.

The related subject of *perceptual intensity* in music has been explored in two studies. Perceptual intensity differs from perceived performed dynamics in that the former is more directly related to the impression of energy in the music, while the latter is more directly related to the impression of the energy level of the performing musicians. In this sense, performed dynamics may be more abstract in nature, and hence more difficult to model. In a study by Zils and Pachet (2003), the perceived intensity was first rated on a scale ranging from “low energy” to “very high energy,” and a mean value of listener ratings computed for each excerpt. The study used Mpeg7 low-level audio descriptors (Herrera *et al.*, 1999) in combination with features from the Extractor Discovery System to reach an  $R^2$  of 79.2 (expressed as a correlation of 0.89 between annotated intensity and predicted intensity from a combination of features). Sandvold and Herrera (2005) classified musical intensity into five classes, ranging from ethereal, through soft, moderate, and energetic, to wild. The authors used basic features such as order statistics of the distributions of sound levels along time, as well as spectral characteristics such as spectral centroid and spectral skewness. The reported classification accuracy was 62.7%.

## C. Machine learning and feature extraction in MIR

There are various methods to model perceptual aspects of music audio. A common approach in music information retrieval (MIR) is to first extract some characteristics (features) from the audio file with signal processing techniques, and then to infer how the features relate to annotated ground truth targets with machine learning methods. When dealing with complex tasks or small datasets (such as the dataset used in this study), the training instances will not be able to fully specify the underlying mapping between input and ground truth (Krizhevsky *et al.*, 2012). In these cases, it is beneficial to either incorporate prior knowledge into the configuration of the machine learning algorithm, e.g., by using weight sharing in artificial neural networks, or, as is also common in MIR, to use a feature extraction process.

### 1. Feature extraction

Features in MIR are often extracted from a spectrogram of the audio file by sampling the bin magnitudes, or by trying to detect spectral characteristics or changes in the spectrum over time, e.g., using Mel-Frequency Cepstrum Coefficients (MFCCs) (Logan, 2000). For an up-to-date review of feature extraction techniques in music we refer the reader to the study by Alías *et al.* (2016).

An important consideration for the feature extraction is the type and extent of prior knowledge to incorporate. If a lot of training data are available, it is possible to infer more complex relationships between the data and the annotations, and therefore not necessary to incorporate as much prior knowledge. In these cases, it is common to use bin magnitudes from the spectrograms directly. If little training data are available, it is necessary to make more assumptions about what characteristics of the data that are relevant for the task. This process is generally referred to as feature engineering, and it is a relevant factor in the success of many

models that rely on machine learning (Domingos, 2012). In this study, prior knowledge (and assumptions) about the features that human listeners associate with performed dynamics will be applied. Given our assumption that timbre is important, the features will be related to timbral characteristics.

A downside of feature engineering (besides being time consuming to develop), is that it may remove relevant information in the process, thus reducing the performance of the subsequent learning algorithm. *We will therefore try to retain as much information as possible, by applying a large variation of settings in the feature extraction.*

## 2. Inference with machine learning

After a set of relevant features have been extracted, machine learning is usually applied to learn from examples, and infer associations between the features and the target values. This is standard practice in MIR, and has been used to, e.g., perform *mood regression* with a recurrent neural network (Weninger et al., 2014) or to perform *genre classification* with standard statistical pattern recognition classifiers (Tzanetakis and Cook, 2002).

A successful machine learning model, a model with high generalization capabilities, minimizes the error on the training set (to prevent *underfitting*), while also minimizing the gap between the training error and the error on the test set (to prevent *overfitting*) (Goodfellow et al., 2016). These two goals can be hard to achieve simultaneously due to the *bias-variance* tradeoff, the fact that minimizing underfitting (reducing bias) may lead to higher variance (overfitting). For example, underfitting may be reduced by creating a model with a sufficiently high complexity. On the other hand, a model with a high complexity may make erroneous assumptions, and therefore overfit the training data. One way to control the complexity of a model is to balance the number of input features, and in the case of neural networks (NNs), to also regulate the size of the hidden layers of the network. An effective technique to improve generalization is ensemble learning, where multiple models are trained, and the average of their predictions is used as a global prediction for the test set. To achieve good results with ensemble learning, models of the ensemble should make diverse predictions, because the average predictions from these models can then be expected to provide a better prediction than randomly choosing one of them (Sollich and Krogh, 1996; Polikar, 2006). A common way of achieving this is to use bootstrap aggregating (bagging) to train, e.g., multiple NNs from different feature subsets (Hansen and Salamon, 1990; Sollich and Krogh, 1996; Polikar, 2006).

In summary, constructing methods with high generalization capabilities is a central factor for achieving good results with machine learning. This can be achieved by incorporating appropriate assumptions about the data, controlling the complexity of the models, and by using ensemble learning.

## 3. Global and local models

Another challenge in MIR is to handle the time-domain in tasks where the annotations for the prediction is a *single*

class or value for the *whole* song, as in this study. One strategy to handle this, called *multiple-instance learning* (Maron and Lozano-Pérez, 1998), is to assign the global target locally to each time frame, then make a prediction for each frame, and finally compute, e.g., an average of the predictions. This strategy is more suitable for tasks such as genre detection and artist recognition (Mandel and Ellis, 2008), where each frame can be expected to carry information that is in accordance with the global prediction. For tasks that do not have this property, such as vocal detection based on a global binary annotation, an iterative procedure of thresholding the local predictions during training to refine the annotations has been tried (Schlüter, 2016). The experiments underlined the importance of frame-wise annotations for the task. Arguably, performed dynamics is functionally somewhere in between these two tasks with regards to how well a global annotation translates to accurate local annotations. Timbre of most music is fairly stable over time (Orio, 2006), but, e.g., MEs with sudden orchestral hits followed by intermediate silence will be rated as having a high performed dynamics globally, but the majority of local frames will not support this rating. If a method can be developed to extract features from the most relevant parts of each excerpt, it would not be necessary to apply frame-wise estimates. Such a method will be developed in this study, as it can be useful, both for performed dynamics and other tasks in MIR.

## D. Previous study of performed dynamics

Previously, we have studied performed dynamics as part of a larger investigation of the concept of perceptual features; both regarding basic perception and for computational modeling in MIR applications (Friberg et al., 2011; Friberg et al., 2014). The ground truth annotations consisted of listener ratings of overall performed dynamics for a variety of music examples. The same listener ratings and datasets are also used in the present study, as described in Sec. II. In the previous study, 25 audio features were used for the prediction (Friberg et al., 2014). The features were calculated using the MIRTtoolbox (Lartillot and Toivainen, 2007), (number of features in parenthesis): MFCCs (13), zero crossings (1), brightness (3), spectral centroid (1), spectral spread, skewness, kurtosis, and flatness (4), spectral roll-off (2), and spectral flux (SF) (1). Ratings of perceived dynamics were predicted using linear regression (LR), partial least-square regression, and support vector regression (SVR), on a dataset consisting of 100 popular music clips and a dataset of 110 film music clips (See Sec. II A). For both datasets, the best result was obtained using SVR, with an  $R^2$  of 0.58 for the popular music dataset and 0.74 for the film music dataset, using 10-fold cross-validation. The rated performed dynamics was thus in general modestly well predicted.

## E. Purpose of the present study

The main purpose of the present study is to build a computational model that can predict the overall performed dynamics in a music audio file. From previous research presented above, it is clear that several acoustical parameters vary with performed dynamics. The characteristics of this

variation have not yet been fully determined, and it also differs between instruments. Considering that the present study aims at modeling performed dynamics for a mix of different instruments, it is difficult to formulate a specific hypothesis regarding the expected relationship between acoustical parameters and the perception of performed dynamics. Therefore, we will extract a broad range of features, and then analyze how different transformations and settings affect prediction accuracy and correlations between features and listener ratings. By doing so, it becomes possible to relate various aspects of the audio to the perception of performed dynamics. This may give insight into what specific characteristics in, e.g., spectral changes that are the most relevant. The feature analysis of this study (Sec. VII) is therefore rather extensive. The previous study for the prediction of performed dynamics only used one SF feature, although various aspects of spectral changes should be important variables. It will therefore be especially interesting to give a more in-depth analysis of this feature extraction technique.

Particularly challenging when using small datasets is to minimize underfitting while also creating models that generalize well, as outlined in Sec. IC. For many MIR-tasks, these challenges are interconnected with the feature extraction, as the large amount of data in the audio signal across time may compel researchers to discard important information during the signal processing stage. One purpose of the study is therefore to explore how these challenges can be managed. We will develop a *sectional* feature extraction of the SF, engineered to capture performed dynamics while balancing the focus to both local and global characteristics of the target. Relevant information will be retained during the signal processing stage by using multiple settings for each transformation, computing features by applying all combinations of settings. This process produces a large feature set, which facilitates ensemble learning where individual models use only a subset of the features. Bootstrapping the feature set in such a way decorrelates the predictions of the models, which improves generalization by satisfying the conditions for a successful ensemble learning specified in Sec. IC2.

Finally, two different ways to estimate the accuracy of the best model in relation to the ground truth annotations are presented in Sec. VIB. This should be useful also for future studies, as the subject has been widely discussed in recent MIR conferences.<sup>1</sup>

## II. DATASETS

### A. Music examples

Two different datasets of music audio recordings were used in the study, for which perceptual ratings of performed dynamics had been collected in two previous experiments (Friberg *et al.*, 2014). The first dataset contains 100 audio examples of popular music (average length 30 s) that were originally produced in the MIDI format and then converted to audio. The second dataset was provided by Eerola and Vuoskoski (2011) and consists of 110 audio examples of film music (average length 15 s), selected for investigating the communication of emotional expression. Both datasets were almost exclusively polyphonic, containing a variety of

musical styles and instrumentations. In the present study, all the examples in the datasets were normalized according to the loudness standard specification ITU-R BS.1770 (ITU, 2006). For the first set this was done before the listeners rated the MEs. The loudness normalization is useful as the overall sound level of musical mixtures varies based on factors not directly related to the performed dynamics in the audio (see Sec. IB for a discussion about this). The processing implicitly enables the developed models to predict performed dynamics without being influenced by the sound level of the analyzed musical mixtures.

### B. Perceptual ratings of performed dynamics

The overall perceived performed dynamics (as well as eight other perceptual features) was rated for both datasets on a quasi-continuous scale by two groups of 20 and 21 listeners, respectively, in two previous experiments (Friberg *et al.*, 2014). In each experiment, the performed dynamics (along with several other perceptual features) of each musical example was rated on a scale ranging from soft (1) to loud (10). The listeners gave one global rating for each ME, and the ground truth was then computed as the average rating of all listeners for each music example. The listeners generally had some musical knowledge, such as playing an instrument on an amateur level or being a music teacher, but none of them were professional musicians. The music examples were presented over high quality loudspeakers, with a calibrated sound pressure level at the listening position. The resulting reliability of the mean estimate across listeners was high [Cronbach's alpha (CA) = 0.94–0.95 for both groups]. For more details about the datasets and procedure we refer to the studies by Friberg *et al.* (2014), Friberg *et al.* (2011), and Friberg and Hedblad (2011).

### C. Final dataset

To get a bigger and more varied dataset, the audio examples from the two datasets were pooled into one dataset consisting of 210 MEs. Pooling two different datasets annotated by different people effectively decorrelates the noise in the annotations, which is good for getting accurate models. The effect is that the developed model will be less likely to model any individual preferences of the annotators. As previously mentioned, the reliability of the ground truth average was estimated by the standardized CA (Cronbach, 1951; Falk and Savalei, 2011). This measure determines the extent to which a set of items (corresponding to listeners in this study) have been measuring the same concept. If a listener does not understand the concept that is being rated, they will decrease the reliability of the final average rating, resulting in a lower CA. The reliability of the ratings will influence how well the two datasets can be used for the same model after pooling. A procedure of removing items that decrease CA can be used to increase the reliability of a construct (Santos, 1999). Therefore, in order to use a reliable estimate, ratings from subjects that decreased the CA were removed in an iterative procedure. The change in CA was calculated for the listeners depending on if they were included or not. For each iteration, if there were any listeners that decreased the value, the listener

that decreased it the most was removed. This procedure resulted in two listeners being removed from the first dataset (increasing CA from 0.937 to 0.940) and seven listeners being removed from the dataset of film music (increasing CA from 0.951 to 0.957). The practice could be especially useful for pooled ratings from different listeners on different datasets. In this case, internal consistency could also increase consistency between datasets. At the same time, the risk that the predicted concept drifts when listeners are removed is reduced, as the procedure is performed on two separate datasets.

### III. FEATURES

What kind of audio features can be used to model the perception of performed dynamics? As discussed in Sec. I, the overall spectrum and the spectral changes are important parameters that vary in relation to the performed dynamics (disregarding sound level). Given that the MEs are polyphonic, with several simultaneous instruments, it is not possible to use any specific instrument model. Instead, a broad range of audio processing transformations was used, and features computed by varying the settings of the transformations. Two groups of features were extracted. The first group was spectral features related to the overall spectrum of the audio. The second group was features computed from the SF, capturing spectral changes mainly related to the onset characteristics in the audio.

As information in percussive and harmonic sounds may be related to the perception of performed dynamics in different ways, it seems reasonable to separate these sounds with source separation before the feature extraction. Source separation has previously been used as a pre-processing step for, e.g., the estimation of perceived speed and tempo in audio (Elowsson and Friberg, 2013; Elowsson and Friberg, 2015). An overview of the feature extraction is shown in Fig. 1.

Harmonic/percussive separation was applied first in the processing chain, as described in Sec. III A. This was followed by the extraction of spectral features described in Sec. III B, and the SF-based feature extraction described in Sec. III C. The features were used to build a machine learning model to predict performed dynamics, as described in Sec. IV, with results presented in Sec. VI. In Sec. VII, an analysis of the 2105 extracted features is provided.

#### A. Source separation

Before calculating features, harmonic/percussive separation was performed on the audio file based on the method

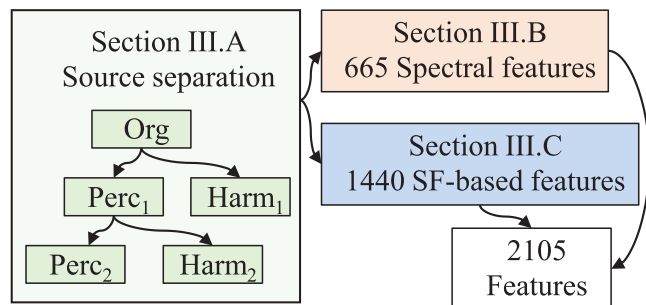


FIG. 1. (Color online) Overview of the feature extraction process.

proposed by FitzGerald (2010). In summary, by applying median filtering on a spectrogram separately across both the time and frequency direction, harmonic and percussive sounds are detected. The resulting harmonic and percussive spectrograms ( $H$  and  $P$ ) are used to create soft masks through Wiener filtering. The  $i$ th frequency of the  $n$ th frame of the harmonic mask ( $M_H$ ) is given by

$$M_{H_{i,n}} = \frac{H_{i,n}^2}{H_{i,n}^2 + P_{i,n}^2}. \quad (1)$$

To compute the percussive mask, the relationship between  $H$  and  $P$  is reversed. Subsequently, the harmonic and percussive audio files are generated by taking the Hadamard product of the masks ( $M_H$  or  $M_P$ ) and the complex valued original spectrogram (FitzGerald, 2010). The resulting complex spectrograms can then finally be inverted back to the time domain by applying the inverse short-time Fourier transform (STFT). In this study we repeated the procedure twice, with the second iteration applied to the percussive audio waveform from the first iteration to remove any remaining harmonic information in the percussive waveform.

For the first iteration, the STFT was used to compute the spectrogram, with a window size of 4096 samples (the audio file was sampled at 44 100 samples/s) and a hop size 1024 samples (about 23 ms). Some frequencies were not filtered with exactly the median value, but the order statistic instead varied with frequency. The exact values and details are specified by Elowsson and Friberg (2015), which give a more detailed overview of this source separation procedure.

In the second iteration, the percussive waveform from the first iteration was filtered again. This time the constant-Q transform (CQT) was used, which produces a spectrogram with logarithmically spaced frequency bins (Schörkhuber and Klapuri, 2010). The frequency resolution was set to 60 bins per octave and each frame was median filtered across the frequency direction with a window size of 40 bins. The resulting spectrograms were then inverted back to the time domain after filtering, resulting in a percussive waveform without harmonic traces, as well as a waveform consisting of these harmonic traces.

The result of the harmonic/percussive separation is five different waveforms. The waveforms extracted in the second filtering step will be denoted as  $Harm_2$  and  $Perc_2$ , whereas the waveforms from the first step will be denoted  $Harm_1$  and  $Perc_1$ , and the unfiltered waveform will be denoted  $Org$ . In this study, features are extracted from each of the waveforms. The relevance of using the different waveforms is analyzed in Sec. VII, and further discussed in Sec. VIII. Note that the harmonic part that is the result of the second filtering step on the percussive waveform will mostly consist of audio from instruments that change pitch suddenly. This has been observed previously and used to extract vocals in a similar setup (FitzGerald, 2012).

#### B. Spectral features

Spectral features were extracted from all five waveforms. In summary, the band-wise spectrum was extracted

from the STFT and a root-mean-square (RMS) value of each band was computed for each ME. The sound level of these bands was then used to compute features.

First the STFT was computed from the source separated audio waveforms (sample frequency 44.1 kHz), with a window size of 1024 samples and a hop size of 441 samples. With this window size, each frequency bin covers about 43 Hz of the spectrum. The magnitudes in the frequency bins of each frame were transformed into bands by multiplication with overlapping triangular filters, uniformly distributed on the log-frequency scale (up to 14 kHz). This is a common technique; see, for example, the study by Heittola *et al.* (2009) for a similar approach. The result is one RMS-value for each band of the complete ME. The frequencies were divided into 2, 3, 4, 6, and 9 bands, respectively (see Fig. 2), which resulted in a vector of 24 RMS-values for each waveform. Features were extracted from the RMS-values as described below. An overview of the three feature types and the distribution of bands across the log-frequency spectrum is shown in Fig. 2.

- The sound level was computed from each RMS-value ( $A$ ), by taking  $10 \log_{10}(A)$ . With 24 bands and 5 waveforms, this resulted in 120 features. Observe that the prior normalization of the audio file (Sec. II A) has the implicit effect that the sound level is computed *in relation* to the sound level of the other bands.
- The difference in sound level between each band and the corresponding band of the other waveforms was computed. For example, the difference of the sound level of

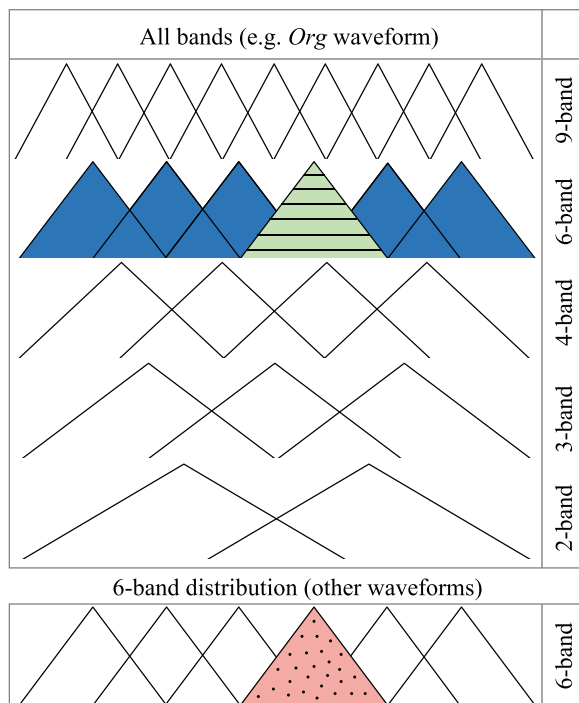


FIG. 2. (Color online) The band-wise distribution (shown across the log-frequency spectrum) of the spectra. The spectra are divided into 2, 3, 4, 6, and 9 bands, respectively. The figure shows which bands are used to compute features for the fourth band in the 6-band distribution, which is the striped triangle-band (green online). The features are computed as the sound level of the RMS-value of this band as well as the difference in sound level of this band and the shaded bands (blue online) of the same waveform and the dotted bands (red online) of the other waveforms.

the fourth band in the 6-band distribution of the original waveform and the sound level of the fourth band in the 6-band distribution of the percussive waveform. This corresponds to the difference in sound level between the striped triangle-band (green online) and the dotted triangle-band (red online) in Fig. 2. With 24 separate bands and 10 unique waveform-pairs, this resulted in 240 features.

- Finally, features were computed by taking the difference in sound level between all bands within the same band-wise distribution (and not between different band-wise distributions) for each waveform. Features of this type were, e.g., computed as the difference in sound level of the fourth band in the 6-band distribution of the original waveform and the other five bands in the 6-band distribution of the original waveform. This corresponds to the difference in sound level between the striped triangle-band (green online) and the shaded triangle-bands (blue online) in Fig. 2. With 61 unique band-pairs (within the band-wise distributions) and 5 waveforms, this resulted in 305 features.

Note that the features described in the two last bullet points consists of linear combinations of the features in the first bullet point. This is just a way to guide the machine learning, by computing some representations that could be useful in the predictions. It also makes it possible to show how the sound level in different frequency bands interact to affect the perception of dynamics, by measuring the correlations between these features and listener ratings. This is done in Sec. VII. The computations resulted in a total of 665 spectral features.

### C. SF-based features

The SF-based features were computed from the CQT (Schörkhuber *et al.*, 2014) of each of the five waveforms. We used 60 bins per octave, a frequency range of about 37 Hz–14.5 kHz, and set the parameter that controls the time-frequency resolution trade-off in the lower frequencies to  $\gamma = 11.6$ . The resulting magnitude spectrogram had 518 frequency bins and the hop size was 256 samples/frame.

There are a few different settings in the transformations for computing the SF that commonly vary between authors, such as the step size and if the computation is performed on the magnitude spectrum or the decibel spectrum of the audio. In this study, some transformations were given multiple settings, using 1–6 different settings for each operation. All possible combinations of the settings were then computed. The feature extraction can thus be understood as a tree-structure, with the transformations as *nodes*, the different settings as *children* of these nodes, and the final features as *leaves*.

By introducing many different (but relevant) nonlinear transformations it is possible to find settings for the SF that are appropriate to model performed dynamics, and it is also possible to combine them into a model with good prediction accuracy, as discussed in Secs. IC and IE. If only one setting is used for each transformation, the risk of discarding relevant information in the feature calculation process increases. A key notion here is to cover relevant and *contrasting* factors of

variations in the audio. For example, some MEs do not contain percussive instruments, which results in less information in the percussive waveforms, and subsequently rather low correlations with listener ratings for these features. But the performance of the final machine learning model may still be increased by including the percussive waveform, if these features capture properties uniquely related to the percussive instruments that can complement features from the other waveforms. However, settings that give lower correlations without covering any relevant factors of variations should still be avoided, as these settings just give extra noise in the features, and lower results overall.

We extracted a total of 1440 SF-based features. A flow-chart of the feature calculation for the SF-based features is shown in Fig. 3, and in Secs. III C 1–III C 10, we describe the operations performed. The abbreviations in parentheses for the headings of Secs. III C 1–III C 10 are the same as those used in the circles of Fig. 3.

### 1. Computing the log of the magnitude spectrogram (SL)

In the first step, the input magnitude spectrogram was either unchanged or transformed to sound level. With this transformation, the magnitude spectrum was converted to the decibel spectrum. The transformation is common (Elowsson and Friberg, 2015), as the decibel scale is more closely related to the perception of loudness than the raw magnitude values (Olson, 1972). If conversion to sound level is not applied, the SF will however be more sensitive to changes in magnitudes for the louder sounds, and this may be perceptually desirable to model performed dynamics.

The sound level was restricted to a range of 50 dB. This was done by normalizing the maximum sound level of the spectrogram to 0 dB and setting any bins below  $-50$  dB relative to the maximum sound level to  $-50$  dB. Note that the MEs were normalized previously according to the ITU-R BS.1770 loudness standard specification (as described in Sec. II). If this had not been done, the subsequent SF computations (when the

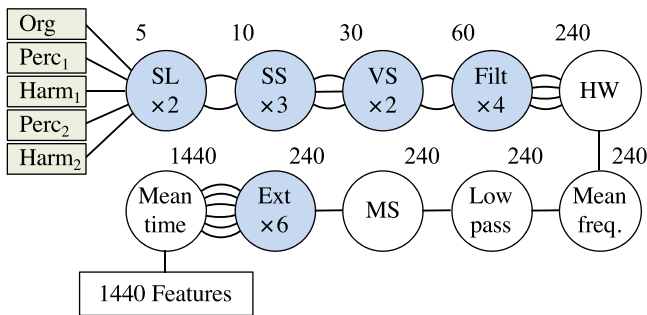


FIG. 3. (Color online) Overview of the transforms used to compute the SF-based features. The input is the magnitude spectrograms from any of the five waveforms. Each circle corresponds to a transformation performed on the signal. The number in the shaded circles (blue online) corresponds to the number of different settings for these transformations. Each setting splits the feature extraction into multiple “nodes” for the next processing steps, as indicated by the black lines. The numbers above each circle correspond to the total number of nodes in the tree structure at the beginning of that processing step. The number of nodes at each processing step grows as all combinations of the different settings are applied, and finally results in 1440 features.

magnitude spectrum is kept) would have been affected by the specific magnitude level of the original waveform.

### 2. Computing the SF with varying step sizes (SS)

Spectral fluctuations are often tracked in MIR, e.g., to detect onsets (Dixon, 2006). As outlined previously, the onset characteristics are an important property of timbre, and therefore the focus for the SF-based features. Let  $L_{i,j}$  represent the input to the SF function at the  $i$ th frequency bin of the  $j$ th frame. The SF for each bin is given by

$$SF_{i,j} = L_{i,j} - L_{i,j-s}, \quad (2)$$

where the variable  $s$  is the step size. We used three different step sizes of 1, 2, and 4 frames (5.8, 11.6, and 23.2 ms).

### 3. Applying vibrato suppression (VS)

Vibrato suppression was used to discard changes in the SF that occur due to small pitch fluctuations (e.g., vibrato), while retaining changes that occur due to onsets. It was applied as an alternative setting during the SF calculation, using the technique of max filtering described by Elowsson and Friberg (2013). In summary, the SF-step from Eq. (2) was changed by including adjacent bins and calculating the maximum value before applying the subtraction

$$SF_{i,j} = L_{i,j} - \max(L_{i-1,j-s}, L_{i,j-s}, L_{i+1,j-s}). \quad (3)$$

The usefulness of a log-frequency resolution for suppressing vibrato consistently over the entire frequency spectrum is the main reason for using the CQT for the SF-based features instead of the STFT.

### 4. Frequency filtering (Filt)

Previous studies indicate that both spectral characteristics and spectral changes are important properties of timbre (MacAdams *et al.*, 1995). Therefore, frequency filtering (Filt) was applied within the SF-based feature extraction to provide information about spectral changes in different frequency bands. Three different filters were applied, and as a fourth alternative, filtering was omitted. The filters  $F_{1,2,3}$ , constructed from a Hann window with a width of 780 frequency bins, are shown in Fig. 4. The centers of the Hann windows were equally distributed over the log-frequency scale and located at bins 130, 260, and 390, corresponding to frequencies of approximately 160 Hz, 740 Hz, and 3.3 kHz. The filtered spectral flux, SF, was computed by applying the Hadamard product between a filter  $F_x$  and each time frame  $SF_i$ , of the SF matrix

$$\dot{S}F_i = SF_i \cdot F_x.$$

### 5. Half-wave rectification (HW)

Half-wave rectification has been proposed many times for the SF computation, see, e.g., Dixon (2006). Half-wave rectification enables the subsequent one-dimensional SF-



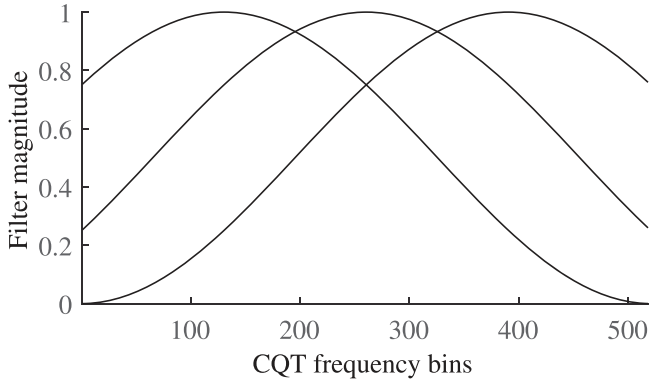


FIG. 4. The filters that were used to change the importance of different frequency bins in the SF. The centers of the Hann windows are at bins 130, 260, and 390, corresponding to frequencies of approximately 160 Hz, 740 Hz, and 3.3 kHz.

curve to account for increases in energy in some frequency bands, regardless of any decrease in energy in other bands. The half-wave rectified response for each frequency bin  $i$  and each frame  $j$  is given by restricting the response to non-negative numbers, after adding the threshold  $lim$  to the corresponding SF-bin

$$SF_{HW_{i,j}} = \max(S\dot{F}_{i,j} + lim, 0). \quad (4)$$

We set  $lim$  to 0, after first experimenting with small negative values.

### 6. Computing the mean across frequency (Mean freq.)

After the half-wave rectification step, the two-dimensional matrix  $SF_{HW}$  was converted to a one-dimensional vector over time, by computing the mean across frequency. This operation is commonly applied to facilitate further processing on the one-dimensional vector (see, e.g., Elowsson and Friberg, 2015).

### 7. Low pass filtering (Low pass)

The one-dimensional  $SF_{HW}$  curve was low pass filtered over time with a zero-phase second order Butterworth filter, with a cutoff frequency of 2.56 Hz. This filtering results in a smoother SF-curve, as shown in Fig. 5. The effect of the

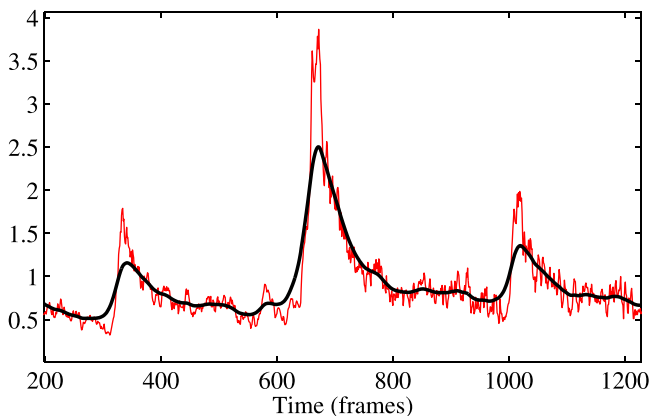


FIG. 5. (Color online) Low pass filtering of the SF curve.

smoothing is that the subsequent processing in Sec. III C 9 is less susceptible to short, noisy peaks. This type of processing is common to get well-defined peaks in the SF-curve (Elowsson *et al.*, 2013; Elowsson and Friberg, 2015).

### 8. Subtracting the mean (MS)

As the next step, the mean was subtracted from the low pass filtered curve ( $SF_S$ ).

$$SF_{MS} = SF_S - \overline{SF}_S. \quad (5)$$

### 9. Extending the analyzed region (Ext)

Positions in the SF-curve that are far away from any peaks will generally consist of stationary sounds or silence. For the SF-based features, these stationary or silent parts of the audio should arguably be discarded. This was done by restricting what parts of  $SF_S$  to average over when computing the mean across time, by only taking the parts where the corresponding regions in  $SF_{SM}$  was above zero. The processing chain can therefore be understood as a *sectional* SF, as it specifically targets regions around the onsets. Before computing the average, the targeted regions were however extended with 0, 25, 75, and 175 frames (corresponding to 0, 0.15, 0.44, and 1.02 s). We extended each region both before the positions where  $SF_{SM}$  rises above 0, and after the positions where  $SF_{SM}$  falls below 0, as shown in Fig. 6. These positions can be thought of as the beginning and the end of any non-stationary parts in the music.

After this operation, the mean of the  $SF_S$  will not only focus on the “onset”-sections of the audio (corresponding to an  $SF_{SM}$  above 0), but also on energy fluctuations just before and after these parts. The big variations in settings was a way to minimize the risk of discarding important information, as we were not sure in the beginning of this study as to which parts of  $SF_S$  that are the most relevant. At the same time, we were reasonably certain that removing parts of silence in the audio before computing the mean SF removes noise in the features; it is not likely that a song starting with

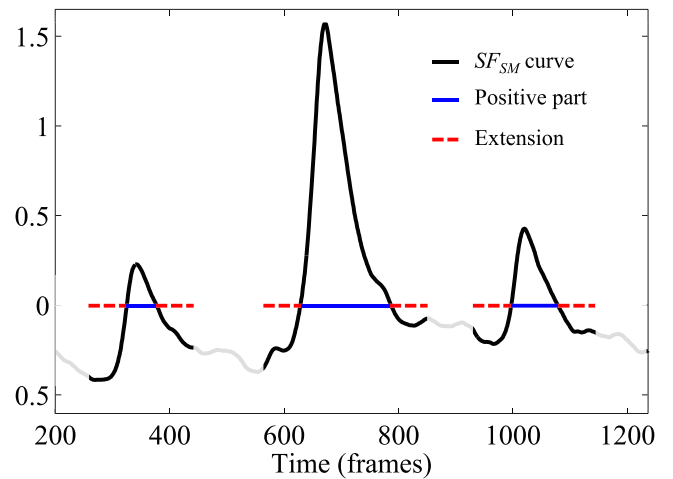


FIG. 6. (Color online) An illustration of the extension step used for deciding what parts of the SFs curve that the *Mean time* operation should be applied to. This is done by extending the regions of interest in the  $SF_{SM}$  curve beyond the positive regions.

a few seconds of silence would be rated much different than when that silence is removed.

Recognizing that sections of the ME before and after onsets may contain different kinds of information related to performed dynamics, the extension was also applied separately to either side (using 75 frames in this case, corresponding to 0.44 s). This resulted in a total of six different settings for the extension. Note that the example in Figs. 4–6 is from an orchestral piece. For music tracks with drums, the extension will quite often make the whole  $SF_S$  curve used. During the development of the feature extraction we tested to omit the *MS* and *Ext* steps, but that resulted in a few percentage points lower correlations with listener ratings for most features.

### 10. Computing the mean across time (Mean time)

Finally, the mean across time was computed from the  $SF_S$ -curve to generate the feature values. This operation was restricted to specific parts of the ME, as described in Sec. III C 9. Note that it is important that the *Mean time* operation is calculated from the extended  $SF_S$  curve and not the  $SF_{SM}$  curve. This is crucial, as the mean subtraction removes relevant information (e.g., the amplitudes of the peaks) in the SF. The last few processing steps (Secs. III C 8–III C 10) are a straight-forward solution for handling the time domain in MIR feature extraction; a problem previously discussed in Sec. I C 3.

### 11. Settings

Several of the operations in Secs. III C 1–III C 10 had different settings, and all combinations of settings were used. In Table I the various settings are summarized. By varying the settings as shown in Table I and Fig. 3, a total of  $5 \times 2 \times 3 \times 2 \times 4 \times 6 = 1440$  features were extracted. The 665 spectral features and 1440 SF-based features resulted in a total of 2105 features.

## IV. MACHINE LEARNING METHODS

As only 210 MEs were available, it was necessary to use methods that generalize well despite a limited number of training instances. Three different machine learning models were employed to predict dynamics; LR, bootstrap aggregated decision trees (BRTs), and an NN in the form of a multilayer perceptron (MLP).

TABLE I. The different settings when computing the SF-based features, as outlined in Fig. 3 and described in Secs. III C 1–III C 10.

Wave	SL	SS	VS	Filt	Ext
Org	True	1	True	False	0
Harm <sub>1</sub>	False	2	False	Low	25
Perc <sub>1</sub>		4		Mid	75
Harm <sub>2</sub>				High	175
Perc <sub>2</sub>					Start
					End

## A. Ensemble learning

Ensemble learning was employed (see Sec. I C 2) by creating multiple instances of a model with the same parameters and averaging their predictions. For all three machine learning methods, an ensemble of 500 models was created. Features were assigned to each model randomly, while ensuring that all features were used an equal number of times. With this setup, each model gets a subset of the features from the larger pool of all features, a technique generally referred to as bootstrap aggregating. As it was ensured that all features were used an equal amount of times, the setup can also be described as using several repetitions of the *random subspace method* (Ho, 1998). By assigning subsampled features to the models, the predictions of the different models will vary, and this can result in good generalization capabilities (Polikar, 2006). The random initialization of NNs further decorrelates the errors of their outputs (Hansen and Salamon, 1990), which together with their non-linearities and relatively lower generalization capabilities (but higher predictive power) should make them extra amenable to ensemble learning.

## B. Configuration of the machine learning methods

For all methods, we tested a few different parameter settings (manual grid search over the parameter space), to determine optimal parameter values, e.g., the number of features in each model.

### 1. Ensemble of LR models (ELR)

An ensemble of LR models was employed, which relied on 40 features in each of the 500 models. Given the total of 2105 features, each feature was used in approximately 10 models.

### 2. Ensemble of BRTs (EBRT)

The second method was to use bootstrap aggregated (bagged) regression trees (Breiman, 1996) in an ensemble. For this method, 20 features were used in each of the 500 models, after first trying a few different feature sizes. Given the total of 2105 features, each feature was used approximately 5 times. Each model had 20 regression trees, with the minimum number of observations for each leaf set to 1. Note that BRT is in itself an ensemble method, thus this method is an ensemble of ensembles.

### 3. Ensemble of MLPs (EMLP)

MLPs were used as a representative of feedforward NNs, in an EMLP consisting of 500 models. The generalization capabilities of MLPs vary with, e.g., the number of hidden neurons, the number of input features for each network, and the number of epochs for training. After performing a manual grid search to determine satisfying parameter values, the following structure of each MLP was used:

- Each model was assigned 40 features. This means that with 500 models and 2105 features in total, each feature was used in approximately 10 models.

- The network was trained with the Levenberg-Marquadt optimization (Marquardt, 1963).
- Each network was trained for ten epochs. Early stopping was not used, mainly because the small size of the dataset makes validation performance unreliable for determining when to stop training. Furthermore, as the MLPs were used in an ensemble, an appropriate epoch to halt training cannot be properly determined during the training of a single MLP. Maximizing the test performance of individual MLPs does not necessarily lead to a maximized performance of the ensemble.
- The networks had three hidden layers, with 6 neurons in each layer, giving an architecture (including input and output layer) of {40, 6, 6, 6, 1}. This means that the MLPs were relatively deep, although the number of parameters was still small.
- Each input feature was normalized by the minimum and maximum value to fit the range between  $-1$  and  $+1$ .
- The non-linearities in the first hidden layer were hyperbolic tangent (tanh) units, and the non-linearities for the following two hidden layers were rectified linear units (ReLU). The output layer had simple linear activation functions.

The unconventional architecture with a mixture of tanh units and ReLUs gave the best results in a manual grid search of different combinations of non-linearities, and has also been used by Elowsson (2016). Here we will outline some of the possible advantages of this architecture. The success of ReLUs is often attributed to their propensity to alleviate the problem of vanishing gradients, and to introduce sparse representations in the network (Glorot et al., 2011). For vanishing gradients, the effect should be the most prominent when, e.g., sigmoid or tanh units are placed in the later layers, as gradients flow backwards through the network. With tanh units in the first layer, only gradients for one layer of weight and bias values will be affected. At the same time, the network will be allowed to make use of the smoother non-linearities. Concerning sparsity induced by ReLUs, it could be argued that sparsity is more favorable in most tasks for later layers of the network, when the representations are more high level.

Bayesian regularization (Foresee and Hagan, 1997; MacKay, 1992) was also tried for the MLP, but this did not improve the results. This is perhaps not surprising, as it has been observed that under-regularized models should be used to maximize the variance-reducing effects of ensemble learning (Sollich and Krogh, 1996).

## V. EVALUATION PROCEDURE

The small size of the dataset makes the training and testing of the models difficult. When small datasets are directly split into a training set and a test set, the results will depend heavily on the composition of the few examples in test set (Goodfellow et al., 2016). Thus, results with such a methodology will vary significantly if a new random split of the dataset is done and the training and testing repeated. To alleviate this problem, a 40-fold cross-validation was used instead, building the model from scratch for each new partition of the dataset as shown in Fig. 7. The cross-validation procedure was also repeated 50 times (making a random

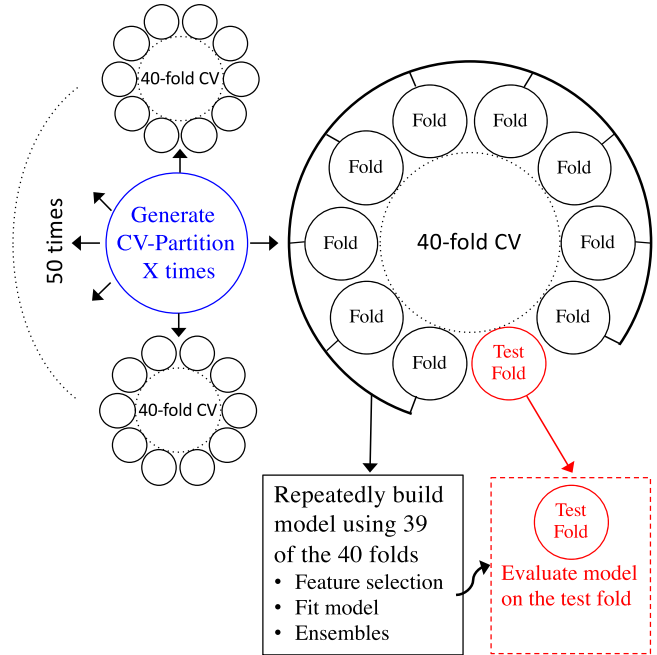


FIG. 7. (Color online) The procedure used in this study for evaluation. Cross-validation was used multiple times, and for each partition and each fold the complete system was trained and tested.

partition of the dataset each time) to ensure consistent results. The design ensures that test folds are disjoint, and the performance of each model was then computed as the average of all test runs. When setting some of the parameters (e.g., the size of the network), we did so by repeating the complete experiment. In this case, we only tested a few combinations of parameters.

## VI. RESULTS

### A. Main results

Prediction accuracy for all models was measured by computing the coefficient of determination,  $R^2$ . This measure corresponds to the squared correlation between the ground truth annotations and the predicted values. The results are presented in Table II. To compute 95% confidence intervals, the 50 test runs were sampled (with replacement)  $10^6$  times and the distributions of mean  $R^2$ s were calculated. Thus, the confidence intervals show the reliability of the computed  $R^2$  based on its variation over the different test runs.

The EMLP was best at predicting performed dynamics. One possible explanation for the success of this method in comparison with the EBRT is the mid-level representations that can be formed in the layer-wise structure of the MLPs. The ability to model non-linear relationships is probably the

TABLE II. The coefficient of determination ( $R^2$ ) of the predictions in relation to ground truth annotations. The EMLP performed the best, with an  $R^2$  of 0.840 for the complete set. Confidence intervals are very narrow due to the high number of test runs.

Results/Method	ELR	EBRT	EMLP
<b>Average <math>R^2</math></b>	0.803	0.814	0.840
<b>95% CI</b>	0.802–0.803	0.814–0.814	0.839–0.840

reason for the higher performance of the EBRT and the EMLP in comparison with the ELR.

What were the effects of using ensemble learning? The number of models in the ensemble of the different methods in relation to explained variance ( $R^2$ ) is shown in Fig. 8. Performance improved with more models for all methods, although the improvement seems to level off after around 50 models have been added. The lowest result with a single model was achieved by the MLP, with an  $R^2$  of 0.690 on average over the 50 repetitions, while the corresponding results were about 0.76 for the other two methods. Five models were required in the EMLP to reach the same performance as an equal number of models for the EBRT. This is not all that surprising, as the MLPs should be able to infer more complex relationships between features and annotations, while lacking generalization capabilities for smaller datasets. The complexity enables the system to handle the more complicated relationships at the cost of a larger variance, which then is reduced for each MLP added to the ensemble.

The predictions from one random partition of the EMLP are shown in Fig. 9. The deviations from the ground truth targets are evenly distributed across the range, with a maximal deviation of about one unit. To get a better overview of the distance between ground truth annotations and fitted values, the error distribution is shown in Fig. 10. These results are the average over all 50 random partitions of the EMLP. More than 90% (0.905) of the MEs were within one point from the correct answer on the scale, the mean distance being 0.43 points.

### B. Results in relation to the performance of the human raters

How does the performance of the EMLP compare to the performance of the individual human raters? This was measured in two different ways.

A rough estimate of how a similar model would translate to other datasets was computed by taking the ratio between the correlation ( $r = \sqrt{R^2}$ ) of the model and the mean CA of the two datasets. As CA can be interpreted as representing the expected correlation between the annotations and any reproduced annotations with new subjects,  $r/CA$  can be interpreted as the correlation of the model in relation to the maximum possible correlation. This interpretation assumes that it is not possible (nor desirable) to model any variance in the ratings that is due to noise. It is thus a way to take the uncertainty of the ground truth data into consideration when evaluating the performance of the model. A value close to 1 indicates that the model is able to explain as much of the variance as possible, whereas a value well below 1 indicates that it should be possible to build a better model, given the data. This ratio has been used previously in a study of subjective well-being by Headey and Wearing (1989). The result when inserting the previously reported values is

$$\frac{r}{CA} = \frac{\sqrt{0.83973}}{(0.95721 + 0.93960)/2} = 0.96622. \quad (6)$$

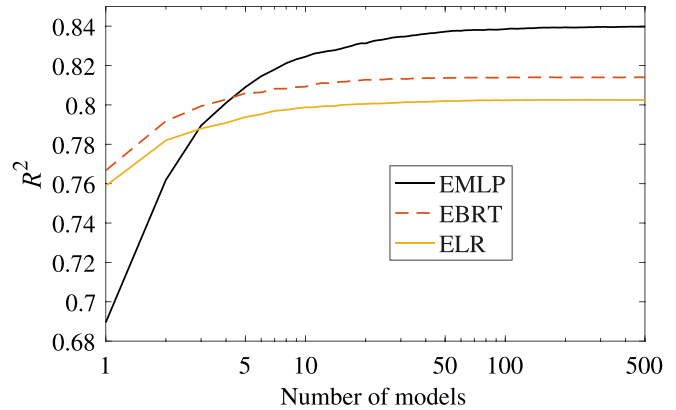


FIG. 8. (Color online) The performance ( $R^2$ ) in relation to the number of models in the ensembles. Results are shown as the mean of all 50 test runs, which gives very small confidence intervals (as indicated in Table II). The x-axis is set to the log-scale. When one model is used, the single BRT and LR model perform better, but the ensemble of MLPs gives higher performance when adding more models.

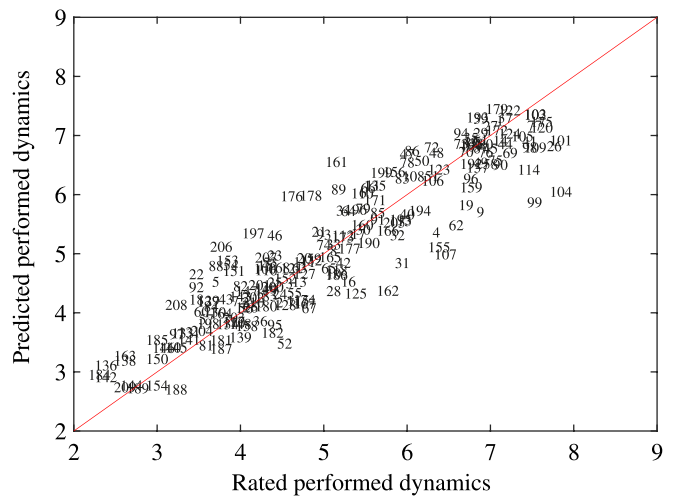


FIG. 9. (Color online) The fitted predictions of the EMLP for one test run (y-axis) in relation to the ground truth annotations (x-axis). The results are from a random partition with an  $R^2$  of 0.842. The MEs are enumerated (to give an overview for future researchers on the same datasets), with the 110 film clips added as numbers 101–210.

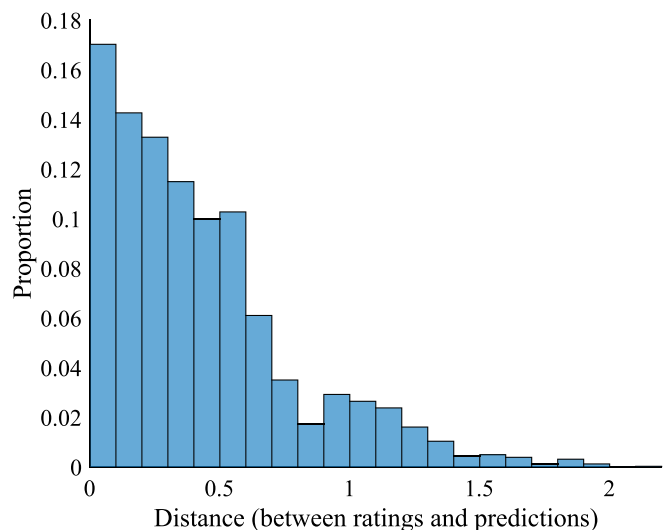


FIG. 10. (Color online) A histogram of the distance between ground truth performed dynamics and predicted performed dynamics. The method used was an EMLP, and results are presented as the average of all 50 test runs.

This indicates that it should be hard to build a much more accurate model.

We also compared the performance of the model with the performance of the human listeners more directly, by evaluating the performance of the individual raters as well as the performance from using the mean of several raters. The performance of  $n$  raters was computed by randomly sampling (with replacement)  $n$  raters, computing their mean rating, and then calculating the  $R^2$  in relation to the mean rating of the full ground truth from either 18 or 14 raters for the two datasets (after removing raters as described in Sec. II C). The procedure was repeated  $10^4$  times before the average  $R^2$  was computed, and the number of raters,  $n$ , was varied between 1–10. The result in comparison to the EMLP is shown in Fig. 11. The EMLP is approximating the ground truth as well as the mean rating of about six human listeners.

## VII. FEATURE ANALYSIS

In this section, the extracted features are analyzed. The correlations between features and ground truth ratings are presented in Sec. VII A. The features are further analyzed in Sec. VII B to establish how important the different types of features are for a high performance in the EMLP model. In Sec. VII C, the findings are used to propose a *reduced feature set*, useful for future models.

### A. Correlations

Correlations were computed using Pearson’s correlation coefficient between features and ground truth ratings of the complete dataset. The purpose is to provide insights into how the features and feature groups relate to performed dynamics.

#### 1. Spectral features

The spectral features were divided into three groups for the analysis, based on the bullet points from Sec. III B. The first group is the features that were extracted by computing the

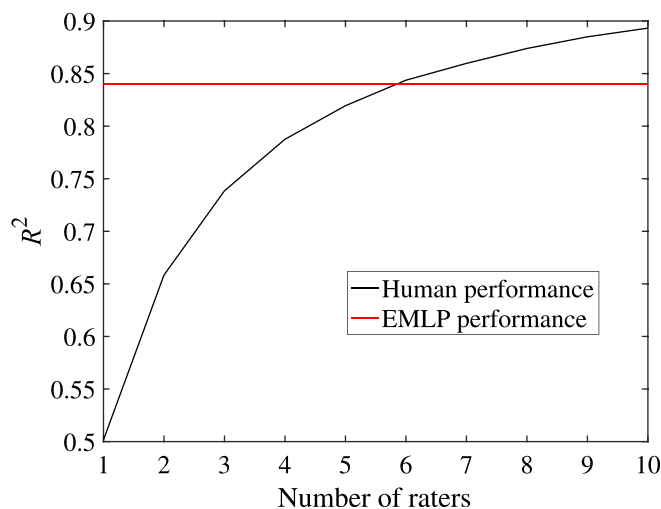


FIG. 11. (Color online) The  $R^2$  between the ground truth and averaged ratings from various numbers of human raters, as well as the  $R^2$  of the EMLP model. The EMLP gets a higher  $R^2$  than individual raters, and it takes about six raters to produce an average prediction that has the same  $R^2$  as the prediction of the EMLP.

TABLE III. The highest correlated features in the group that were computed as the sound level of each individual band. Bands with center frequencies around 5.5 kHz have a positive and rather high correlation with performed dynamics.

Waveform	Center freq. (X-band distr.)	$r$
Org	5.5 kHz (6)	0.663
Org	3.8 kHz (9)	0.652
Harm <sub>1</sub>	7.3 kHz (9)	0.647
Org	7.3 kHz (9)	0.643
Perc <sub>1</sub>	3.8 kHz (9)	0.612
Harm <sub>1</sub>	5.5 kHz (6)	0.598
Perc <sub>2</sub>	3.8 kHz (9)	0.597

sound level of each individual band. Table III shows the highest correlated features from this group. Bands with center frequencies in the range of 4–7 kHz gave the highest correlations.

The second group is the features that were extracted by computing the difference in sound level in different waveforms within the same bands. Table IV shows the highest correlated features in this group. This feature group had the smallest correlations among the three groups. A higher sound level in percussive waveforms corresponded to a higher performed dynamics.

Finally, the third group consists of the features that were extracted by computing the difference in sound level between all bands within the same band-wise distribution and waveform. Table V shows the highest correlated features in this group. This group had the highest correlations among the spectral features. The difference in sound level between low–mid frequencies and high frequencies gave high negative correlations. This shows that a high sound level in the high frequencies and a low sound level in the low–mid frequencies is related to higher perceived performed dynamics.

#### 2. SF-based features

All SF-based features had a positive correlation with listener ratings. In Table VI we give the setting of the SF-based feature (Sec. III C) from each of the five waveforms that had the highest correlation with performed dynamics.

Overall, features extracted from the original waveform had the highest correlations with performed dynamics, but features from the harmonic waveform also had high

TABLE IV. The strongest correlated features that were computed as the difference in sound level of frequency bands in different waveforms. The negative correlations when features from the percussive waveforms were subtracted indicates that more percussive content corresponds to a higher performed dynamics.

Waveform	Center freq. (X-band distr.)	$r$
Harm <sub>1</sub> –Perc <sub>2</sub>	1.6 kHz (2)	–0.557
Perc <sub>1</sub> –Harm <sub>1</sub>	1.6 kHz (2)	0.533
Harm <sub>1</sub> –Perc <sub>2</sub>	340 Hz (6)	–0.528
Harm <sub>1</sub> –Perc <sub>2</sub>	2.7 kHz (3)	–0.527
Org–Perc <sub>1</sub>	1.6 kHz (2)	–0.524
Harm <sub>1</sub> –Perc <sub>2</sub>	260 Hz (9)	–0.523
Org–Perc <sub>1</sub>	520 Hz (3)	–0.520

TABLE V. The strongest correlations for features that were computed from the difference in sound level between all bands within the same band-wise distribution and waveform.

Waveform	Band centers (X-band distr.)	$r$
Harm <sub>1</sub>	520 Hz–7.3 kHz (9)	-0.69
Harm <sub>1</sub>	260 Hz–7.3 kHz (9)	0.665
Harm <sub>1</sub>	1.0 kHz–7.3 kHz (9)	-0.664
Harm <sub>1</sub>	860 Hz–5.5 kHz (6)	-0.660
Org	1.0 kHz–3.8 kHz (4)	-0.655
Org	860 Hz–5.5 kHz (6)	-0.654
Org	340 Hz–5.5 kHz (6)	-0.651

correlations. The best explaining features from the original waveform consistently had *SL* and *VS* set to *false*, whereas the best explaining features from the harmonic waveform consistently had *SL* and *VS* set to *true*. The extension of 75 frames (0.44 s) in both directions was the most common among the highest correlated features, but 25 frames (0.15 s) and 175 frames (1.02 s) also gave high correlations. All three step sizes in the SF were equally common. Filtering to keep mid and high frequencies generally gave higher correlations than when including lower frequencies.

To give a more in-depth analysis of the effect of using different settings in the SF computation, the already computed features were used as *experimental runs* at *factorial points* similar to a *full factorial experiment* (see, e.g., Fisher, 1949), with each processing step that had more than one setting representing a *factor*, and each setting representing a *level*. Given the variation in settings described in Table I, this analysis for the SF-based features can be denoted as a  $2^2 \times 3 \times 4 \times 5 \times 6$ -factorial design. In contrast to a traditional factorial experiment, the response for each *experimental condition* became a vector, as the same features had been computed for all 210 MEs. To get a single response variable for each experimental condition, the correlations between feature vectors and ground truth annotations were computed. It was then possible to compute the impact of each setting (similar to the *main effect* in factorial experiments) by taking the average across the response variables from a specific setting, and then to compare with features that had the other settings in the same processing step. This was done across all processing steps.

To compute approximate 95% confidence intervals on the main effect, we used Monte Carlo bootstrapping, resampling the 210 MEs with replacement and calculating the

TABLE VI. The settings for the highest correlated SF-based feature from each of the different waveforms. The different settings are described in Sec. III C, and  $r$  is the correlation coefficient between the feature and the ground truth ratings of performed dynamics for the 210 MEs. Rank is given by ranking all the features from the highest correlated to the lowest correlated.

Waveform	SL	SS	VS	Filt	Ext	$r$	Rank
Org	False	1	False	Mid	75	<b>0.861</b>	<b>1</b>
Harm <sub>1</sub>	True	2	True	Mid	75	<b>0.851</b>	<b>9</b>
Perc <sub>1</sub>	True	4	False	Mid	75	<b>0.664</b>	<b>486</b>
Harm <sub>2</sub>	False	2	False	—	75	<b>0.730</b>	<b>314</b>
Perc <sub>2</sub>	True	2	False	—	25	<b>0.537</b>	<b>994</b>

correlation  $10^5$  times. The effect is that the response variable will vary slightly, depending on the randomized composition of the dataset in each resampling. The average response variable for a specific setting was calculated within each bootstrap, which resulted in  $10^5$  average responses for each setting across all bootstraps. The vectors of responses for all settings were sorted, and the values at 5% at 95% selected as confidence intervals. The confidence intervals thus approximate how the correlations could be expected to vary in other datasets with different compositions but similar types of recordings. In Figs. 12 and 13, the results of the factorial experiment concerning correlations for the SF-based features is shown, and the computed confidence interval for each setting is indicated with a black line.

Varying the source (waveform) had quite a big effect on the correlations for the other settings, so we decided to use only the *Org* waveform when testing all other processing steps. After this, we tested the effect of using the different waveforms separately. The variation in response variable (correlation) for the settings (groupings of features) is shown in Fig. 12, in relation to the average response (correlation) of all the 288 SF-based features from the original waveform, which was 0.692. For the settings, the effects of the vibrato suppression (*VS*) and frequency filtering (*Filt*) were the strongest. Applying vibrato suppression and filtering to keep the bass frequencies reduced correlations with about 0.09 on average over the different features. For vibrato suppression, these results indicate that there is information related to performed dynamics in the smaller pitch shifts that this processing step suppresses. Shorter step sizes seem to give higher

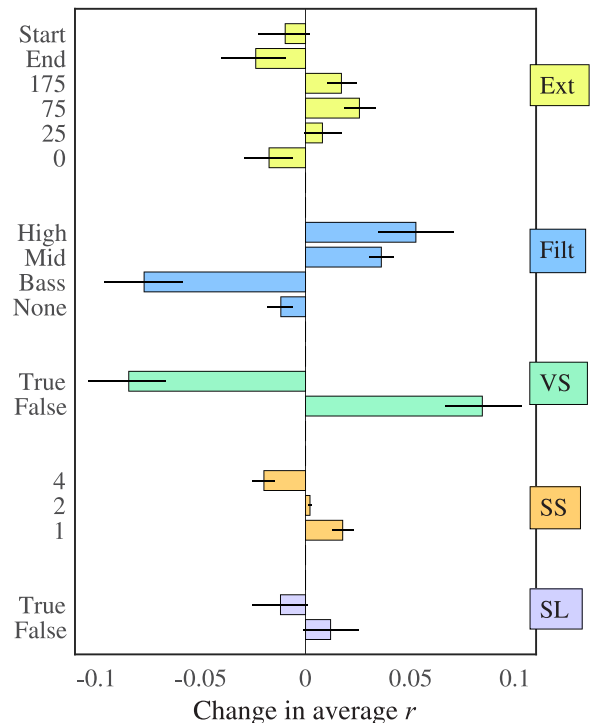


FIG. 12. (Color online) The main effect for different settings when computing the SF-based features. Bars indicate change in average correlation for features computed with a specific setting and black lines indicate approximated 95% confidence intervals. Using vibrato suppression and focusing on the bass frequencies reduced correlations significantly.

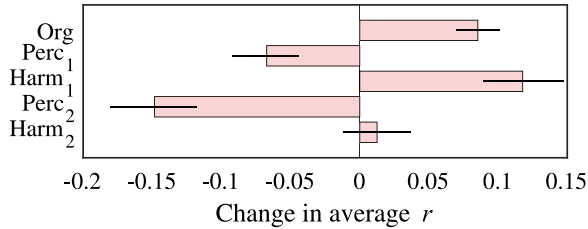


FIG. 13. (Color online) The main effect for different sources when computing the SF-based features. Bars indicate change in average correlation for features computed from a specific waveform and black lines indicate approximated 95% confidence intervals. Features computed from harmonic source material had higher correlations with ground ratings.

correlations and using a two-sided extension of 75 frames (0.44 s) gave the highest correlations on average for the extension step.

The variations in correlation for features when using different waveforms are shown in Fig. 13, in relation to the average correlation of all SF-based features (0.607). The *Org* and *Harm<sub>1</sub>* waveforms gave the highest correlations, whereas the features from the two percussive waveforms had the lowest correlations. Given that the waveform that consists only of harmonic traces filtered out from the percussive waveform (*Harm<sub>2</sub>*) had a higher correlation than any of the percussive waveforms, it is evident that more harmonic material in the audio source increased correlations.

## B. Performance

The most important aspect of the computed features is the potential they offer for predicting performed dynamics in the actual models. The impact on performance of the different feature calculation settings was therefore tested with the EMLP described in Sec. IV B 3, using the evaluation procedure described in Sec. V. The already computed features were used as experimental runs for factorial points in a factorial design, similar to when previously testing correlations for the SF-based features. Features were divided into subsets based on their corresponding setting in each processing step. This means that all processing steps produced subsets that spanned the complete feature space. For testing performance, features in the subsets were not correlated individually with ground truth ratings to compute a response variable as previously, but they were instead used to train the specific machine learning method. One single response variable was thus produced for each subset, by training the EMLP with the features in the subset, and then computing the  $R^2$  of the predictions for the MEs in relation to ground truth ratings. As we will show, the performance analysis provided additional information about the relevance of the different processing steps.

Each MLP of the ensemble was assigned an input of 40 randomly selected features from the subsets, just as in the main experiment of the study. However, one feature subset only had 35 features in total (the 2-band distribution for spectral features), and in this case these 35 features were used as input for each MLP of the ensemble. Due to the computational cost of training the EMLP, the evaluation was limited to 20 test runs (down from 50 in the main

experiment), each with a randomized partitioning of the dataset for the cross-validation. The number of MLPs used in each ensemble was reduced to 50 (down from 500), due to the limited performance gain when using more networks (see Fig. 8). To compute 95% confidence intervals, the 20 test runs were sampled (with replacement)  $10^6$  times and the distributions of mean  $R^2$ s were calculated. Thus, the confidence intervals show the reliability of the computed  $R^2$ s based on their variations over the different test runs. To be able to calculate the change in  $R^2$  in relation to when training on the complete feature set, the experiment was also run for 50 test runs on each of the two feature sets (the spectral features and the SF-based features).

## 1. Spectral features

The spectral features were divided into three different levels. The *Type* divides the features into those that were computed as the sound level of the different bands (*Only sound level*), the difference in sound level for all bands between different waveforms (*Diff - HP*), and the difference in sound level within each band distribution (*Diff - WB*). For the other two levels, the features were divided into those computed from different waveforms (*HP*), and finally those computed from the different band-wise distributions (*BD*) (see Fig. 2). The analysis was thus performed for a  $3 \times 5^2$ -factorial design. The performance of the EMLP trained on the specified subsets, subtracted from the performance of the EMLP trained on all the spectral features ( $R^2 = 0.728$ ), is shown in Fig. 14.

It is clear from Fig. 14 that the highest performance was achieved when training the EMLP with features computed only from the sound level in each band (*No Diff*). This gives each MLP more freedom to map how the relationships

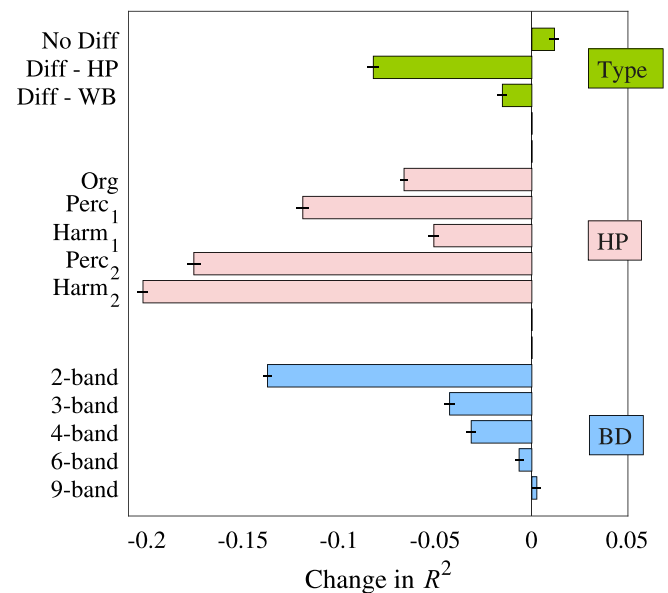


FIG. 14. (Color online) Change in the coefficient of determination ( $R^2$ ) for the EMLP when using different subsets of the spectral features for training, in relation to when all spectral features were used. Results and 95% confidence intervals were computed by running the complete experiment 20 times. A positive change in  $R^2$  was achieved when using only the raw sound level of each band to compute features (*No Diff*).

between sound levels in different bands translate to the ground truth ratings of performed dynamics. The performance only decreased slightly when features computed from the difference in sound level within each band-distribution were used (*Diff - WB*). This indicates that this transformation (directly comparing how the sound levels in different bands relate to each other) is still rather relevant to the task. However, when the features used were those extracted as the difference in sound level between the different waveforms (*Diff - HP*), the performance decreased more. These types of features also had the lowest correlations, and it seems like such representations are not as relevant to the perception of performed dynamics.

Generally, “raw” features led to the best performance. For example, when the frequency resolution was decreased by using features from fewer and broader bands (e.g., *2-band*), the performance suffered considerably. It is also clear that the source separation increased performance. When the EMLP only had access to features from one of the waveforms, performance was lower. The harmonic content in the audio seems to be the most relevant for the spectral features, as indicated by a relatively high result when only features from the *Harm*<sub>1</sub> waveform were used.

## 2. SF-based features

The already computed SF-based features were divided into the same subsets as in Sec. VII A 2 when computing correlations, and the same technique for analyzing performance was used as in Sec. VII B 1. The performance of the EMLP trained with these subsets, subtracted from the performance of the EMLP trained with all SF-based features ( $R^2 = 0.824$ ), is shown in Fig. 15.

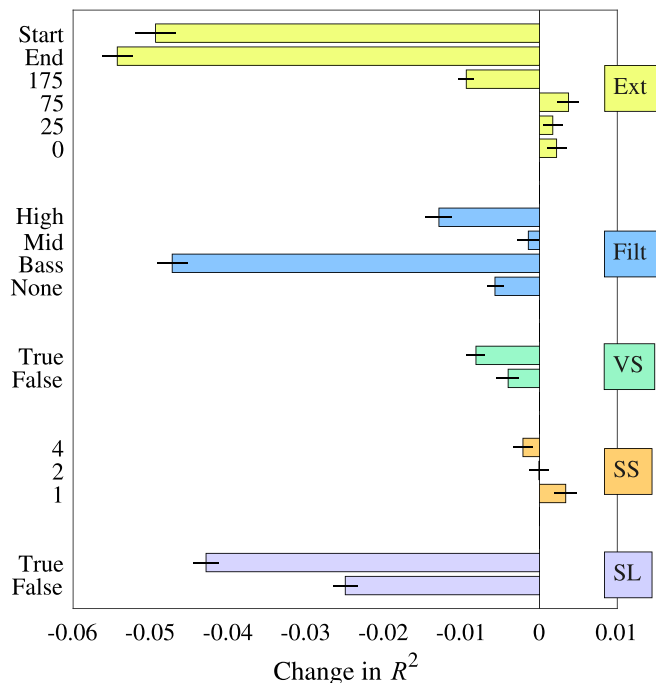


FIG. 15. (Color online) Change in the coefficient of determination ( $R^2$ ) for the EMLP when using different subsets of the SF-based features for training, in relation to when all SF-based features were used. Results and 95% confidence intervals were computed by running the complete experiment 20 times.

For some transformations, such as conversion to sound level (*SL*) and vibrato suppression (*VS*), the performance dropped in relation to the full feature set, regardless of which feature subset that was included. Why is this? When the MLPs do not have access to features with both variations of these settings, they cannot use the variation in feature values to train a model that accounts for these different factors in an appropriate way. Varying these settings thus provides the MLPs with information that they can use to learn more accurate models of performed dynamics. One example, and possible explanation for the *SL*-transformation, is that computing the SF from the log-domain spectrogram provides relevant information about spectral fluctuations for fainter sounds, whereas computing the SF of the magnitude spectrogram gives more information about fluctuations of the louder sounds. The MLPs can then learn how to combine these different features to create a model for how fluctuations at varying levels of salience interact. Both the magnitude and log-spectrogram have been used successfully for a lot of studies and tasks when computing the SF, because there is no consensus on the appropriate settings. Our results indicate that, if compute time is not a restricting factor, both variations can be employed to increase performance, by propagating the decision on how to combine them to the machine learning stage.

Using only shorter, two-sided extensions ( $Ext = 0, 25$  or  $75$ ) and only shorter step sizes during the SF ( $SS = 1$ ) increased the performance in relation to when all features were included. For features where the extension had only been applied to one side (*Start* or *End*) in the extension step, performance dropped significantly. It is interesting to see that performance was high when using features that had the extension set to 0, as this is easier to implement. Smaller step sizes in the SF seem to give better results. The features from the *mid* frequencies in the frequency filtering (*Filt*), provides the highest results, although results seem to fall slightly in comparison to when all features with different filtering settings were included.

The results when using features from the different waveforms, in relation to an EMLP trained on all the SF-based features ( $R^2 = 0.824$ ), are shown in Fig. 16. Variation across waveform had the strongest effect for the performance of the EMLP. The  $R^2$  decreased by around 0.05 when using features from the *Org*, *Perc*<sub>1</sub>, or *Harm*<sub>1</sub> waveforms, in relation to when features from all waveforms were used. For the *Perc*<sub>2</sub> and *Harm*<sub>2</sub> waveforms, the reduction in performance was even greater. These results, together with similar results for the spectral features shown in Fig. 14, highlights the importance of the source separation.

## C. Reduced feature set

The analysis in Sec. IV B indicates that the performance of the system increases when using features from a variation of settings applied to some transformations, while applying varying settings to others may not improve performance (see Figs. 14–16). Here we propose a *reduced feature set* by selecting a single setting for some transformation, while leaving all variations in settings for others. As features for



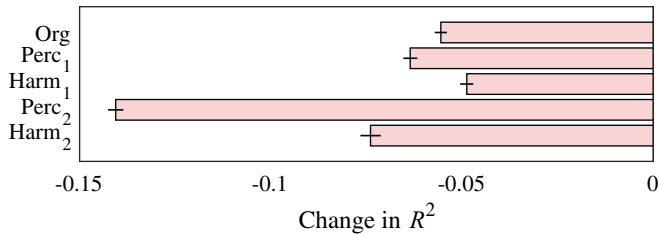


FIG. 16. (Color online) Change in the coefficient of determination ( $R^2$ ) for the EMLP when using SF-based features from different waveforms for training, in relation to when all SF-based features were used. Results and 95% confidence intervals were computed by running the complete experiment 20 times.

the reduced set have been selected based on performance, results for this feature set are not to be interpreted as a test results like those provided in Sec. VI A. Instead, the reduced feature set is used to test how the results change when less relevant features have been discarded, and as a starting point for future studies.

For the spectral features (Fig. 14), the features from the 9-band distribution were selected and only the sound level (*No Diff*) of each band was used. As the results were higher when all waveforms were used, they were all included. This resulted in a total of 45 spectral features. For the SF-based features (Figs. 15 and 16), features with a 75-frame extension and a step size of 1 frame in the SF were selected. For the other transformations, all settings were included. This resulted in 80 SF-based features, and a total of 125 features for the reduced set.

The results when training the three ensemble models with the reduced feature set and testing performance 50 times are presented in Table VII. To compute 95% confidence intervals, the 50 test runs were sampled (with replacement)  $10^6$  times and the distributions of mean  $R^2$ s were calculated. Results did not change much in relation to when the complete feature set was used, and the highest results were achieved with the EMLP. The small improvements in results for the EBTR indicate that this model cannot discard the irrelevant features as successfully as the other methods when using the complete feature set. One possible explanation is the smaller number of input features used for each BRT (20), which could reduce performance for noisier feature sets.

## VIII. CONCLUSION AND DISCUSSION

A system was developed to predict perceived performed dynamics for two small music audio datasets. Novel audio features were extracted and analyzed, and then used to train

TABLE VII. The coefficient of determination ( $R^2$ ) of the methods when trained on the reduced feature set. The EMLP performed the best—slightly better than in the main test. The EBTR had the greatest improvement in relation to the main test, but results did not vary much overall.

Results/Method	ELR	EBTR	EMLP
Mean $R^2$	0.804	0.825	0.841
95% CI	0.803–0.804	0.825–0.826	0.840–0.841

machine learning algorithms based on ground truth ratings previously collected. An ensemble of MLPs achieved an  $R^2$  of 0.840, a significant increase in comparison with a previous study on the same datasets (Friberg *et al.*, 2014). The study is therefore the first to present a successful way of predicting global absolute (i.e., not for transitions) performed dynamics from poly-instrumental music. As the MEs in the two datasets were distinctly different, this indicates that the model should generalize well to other datasets (the authors also found this to be the case when they informally tested the EMLP on a new dataset after this study was concluded). The results were well above those of the individual human listeners that rated the MEs, so applying this model to get a rating of performed dynamics would arguably give higher accuracy than asking someone to manually annotate a dataset. In fact, results indicate that it would take up to six raters to approximate the population average with higher accuracy than when using the model. By taking the uncertainty of the ratings into consideration, it was found that  $r/CA = 0.966$ , which indicates that the results are close to the upper bound, given the estimated uncertainty of the annotations.

## A. Features and feature analysis

Due to the lack of prior art for the task, novel features were developed and analyzed. The sectional SF-method of subtracting the mean from the SF-curve, finding all regions above 0, and extending these regions by a fixed number of frames, produced the features that had the highest correlation with dynamics. This processing scheme is a simple way of handling the time-domain in MIR feature extraction. It could easily be adapted to similar tasks where the dependent variable is a global annotation, but where the annotated attribute in the audio is not present during the entire ME.

In the feature analysis, the extracted features were used as experimental runs at factorial points in a factorial design, by dividing the features into different subsets. By doing this, the effect of the different settings and transformations could be determined. What does the feature analysis reveal about the relationship between audio transformations and performed dynamics?

### 1. Correlations

For the spectral features, higher sound level in the high frequencies positively correlated with performed dynamics. It seems like energy around 6 kHz has the strongest positive correlations, whereas energy around 500 Hz has the strongest negative correlations (see Tables III and V for an overview). More percussive information in the audio was also positively correlated with performed dynamics (see Table IV). These results for the spectral features have a straight-forward interpretation in relation to performed dynamics. When more force is used to produce sounds in musical instruments, the spectrum generally has more energy in higher frequencies (Luce, 1975; Fabiani, 2009). Features that cover the frequencies of the  $f_0$ s or first harmonics of common instruments therefore have strong negative correlations, while features that cover higher harmonics have strong positive correlations. At lower bass frequencies the relationship is weaker,

something that partly could be attributed to the higher variance of the long-term average spectrum in these frequencies (Elowsson and Friberg, 2017). Also, percussive sounds are generally associated with faster attack times, which is an aspect of the timbre that should result in a higher perceived performed dynamics.

For the SF-based features, Table VI and Fig. 13 show that the SF from harmonic parts of the audio gave higher correlations than the SF of the percussive parts. This indicates that the SF of harmonic instruments is a more reliable factor for measuring performed dynamics than the SF of percussive instruments. For a comprehensive interpretation of these results, it is however important to consider the composition of the dataset. About half of the MEs in the combined dataset were from film music, which generally contains less percussive instruments. Therefore, the harmonic parts of the audio overall contained more information. But the percussive waveform will generally have a lower sound level than the harmonic waveform also for regular popular music. A previous study has shown that the *Perc*<sub>2</sub> waveform is around 15 dB lower than the *Org* waveform on average (Elowsson and Friberg, 2017). Therefore, it is still reasonable to assume that the results are similar also for other datasets of, e.g., popular music.

The correlations were generally higher in the SF-based features than in the spectral features. The results however demonstrate that both spectral fluctuations and spectral balance are related to perceived performed dynamics. The factorial experiment also reveals that vibrato suppression reduces correlations across features by approximately 0.08 (see Fig. 13). This indicates that fluctuations across both time (SF) and frequency (vibrato) are related to perceived performed dynamics.

## 2. Performance

The adequacy of the different feature calculation settings was tested by measuring the performance ( $R^2$ ) of the EMLP trained on different feature subsets, as outlined in Sec. VII B. Variations in settings that transformed the signal very little (such as using different step sizes), did not contribute to an improved performance. However, some of those settings seemed to give higher results than others, so computing multiple settings also for these transformations may still be useful to determine the optimal setting. Variations in settings had a higher impact on performance for the transformations that changed the signal more drastically, such as transforming the magnitude spectrogram to the log-domain. As discussed in Sec. VII B, when settings for a transformation are varied and features computed from all variations, it becomes possible for the MLPs to infer how an attribute (e.g., vibrato, frequency), affected by that transformation, relates to the task.

When comparing the performance of the SF-based features from Fig. 15 with the correlations shown in Fig. 13, it is evident that variations in correlation only to some extent translates to variations in performance. This observation underlines that the interaction of features must be taken into consideration in the feature extraction. The observation can also serve as a

starting point for a broader discussion considering the feature extraction and its role in this and similar MIR tasks. To understand an acoustical scene, the listener must infer the characteristics of the sounding sources. This enables the listener to estimate the performed dynamics of individual instruments as well as the audio mixture. The MLPs ability to estimate performed dynamics is in a similar way related to their possibility to infer characteristics from audio features. It is therefore not surprising that the harmonic/percussive source separation was the most important transformation with regards to the performance of the model. It is reasonable to assume that the separated sources have different characteristics related to performed dynamics. Without source separation, this information will not become available to the machine learning models.

Although the correlations were higher for the features filtered to keep high frequencies, the performance was higher when using the mid-frequencies. There are possibly two different explanations. The first is that correlations were measured only for the features of the original waveform, and that the correlations in the other waveforms could have changed the results. The other explanation is that when only high frequencies are used, more information is lost. Therefore, the system is not able to deduce other interactions as effectively when making predictions.

## B. Future studies

### 1. Features

How can the feature extraction be improved for future studies? The reduced feature set proposed in Sec. VII C is a good starting-point, as it removes redundant variations in settings. The feature analysis revealed that variations in transformations that change the features in a more substantial way results in measurable improvements. One possible direction would therefore be to utilize MFCCs, as this transformation effectively decorrelates the computed features. The sectional process proposed in this study is however still important, to produce good features that gives high results. This leads us to propose that the performance of sectional MFCCs (SMFCCs) should be investigated. This feature extraction technique could be useful not only to preformed dynamics, but to many related tasks where global and local aspects of the signal must be balanced.

One aspect of performed dynamics that was not explicitly modeled with the features developed in this study is the relation between the spectrum and ground truth pitches ( $f_0$ s). As the magnitude distribution of the harmonics of a tone is related to the dynamics with which the instrument is played,  $f_0$ -tracking could provide useful information. This could also be explored in future studies.

### 2. Dataset and listener ratings

For future studies, it would also be beneficial to create a bigger dataset. This would enable more powerful machine learning methods, instead of only relying on methods that are known to generalize well on smaller datasets. One way of artificially expanding the dataset is to partition the MEs into overlapping sections of about 5–10 s. Predictions could

then be made for each overlapping section. This is similar to the technique of artificially expanding datasets for image processing by discarding pixels at the edges of the image (Krizhevsky *et al.*, 2012). With overlapping sections of that size (5–10 s), global annotations should still be sufficiently accurate for each frame. This represents a compromise between multiple-instance learning and global learning. Furthermore, the perceptual ratings could also be collected continuously during playback. This would result in annotations with higher accuracy for each overlapping section.

Ratings between novice and expert musicians could also be compared; they could potentially differ with regard to the perception of performed dynamics and inter-rater agreement.

## ACKNOWLEDGMENTS

This work was supported by the Swedish Research Council, Grant No. 2012-4685. Thanks to Pawel Herman for proofreading and technical input.

<sup>1</sup>As discussed during the Evaluation Town Hall of the ISMIR Conference 2016.

- Alías, F., Socoró, J. C., and Sevillano, X. (2016). “A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds,” *J. Appl. Sci.* **6**(5), 143.
- Berndt, A., and Hähnel, T. (2010). “Modelling musical dynamics,” in *Proceedings of the 5th Audio Mostly Conference: A Conference on Interaction with Sound*, 8 pp.
- Breiman, L. (1996). “Bagging predictors,” *Mach. Learn.* **24**(2), 123–140.
- Cancino Chacón, C. E., and Grachten, M. (2015). “An evaluation of score descriptors combined with non-linear models of expressive dynamics in music,” in *Discovery Science*, Lecture Notes in Computer Science Vol. 9356, edited by S. Japkowicz and S. Matwin (Springer, Cham).
- Chalupper, J., and Fastl, H. (2002). “Dynamic loudness model (DLM) for normal and hearing-impaired listeners,” *Acta Acust. Acust.* **88**(3), 378–386.
- Clark, M., and Luce, D. (1965). “Intensities of orchestral instrument scales played at prescribed dynamic markings,” *J. Audio Eng. Soc.* **13**(2), 151–157.
- Cronbach, L. J. (1951). “Coefficient alpha and the internal structure of tests,” *Psychometrika* **16**(3), 297–334.
- Dixon, S. (2006). “Onset detection revisited,” in *Proceedings of the 9th International Conference on Digital Audio Effects (DAFx-06)*, Montreal, Canada (September 18–20, 2006), pp. 133–137.
- Domingos, P. (2012). “A few useful things to know about machine learning,” *Commun. ACM* **55**(10), 78–87.
- Eerola, T., and Vuoskoski, J. K. (2011). “A comparison of the discrete and dimensional models of emotion in music,” *Psychol. Music* **39**(1), 18–49.
- Elowsson, A. (2016). “Beat tracking with a Cepstrum invariant neural network,” in *17th International Society for Music Information Retrieval Conference*, pp. 351–357.
- Elowsson, A., and Friberg, A. (2013). “Modelling perception of speed in music audio,” in *Proceedings of the Sound and Music Computing Conference*, Stockholm, Sweden, pp. 735–741.
- Elowsson, A., and Friberg, A. (2015). “Modeling the perception of tempo,” *J. Acoust. Soc. Am.* **137**(6), 3163–3177.
- Elowsson, A., and Friberg, A. (2017). “Long-term average spectrum in popular music and its relation to the level of the percussion,” presented at the 142nd Audio Engineering Society (AES) Convention, Berlin, 12 pp.
- Elowsson, A., Friberg, A., Madison, G., and Paulin, J. (2013). “Modelling the speed of music using features from harmonic/percussive separated audio,” in *Proceedings of the International Symposium on Music Information Retrieval*, pp. 481–486.
- Fabiani, M. (2009). “A method for the modification of acoustic instrument tone dynamics,” in *Proceedings of the 12th International Conference on Digital Audio Effects (DAFx-09)*, pp. 359–364.
- Fabiani, M., and Friberg, A. (2011). “Influence of pitch, loudness, and timbre on the perception of instrument dynamics,” *J. Acoust. Soc. Am.* **130**(4), 193–199.
- Falk, C. F., and Savalei, V. (2011). “The relationship between unstandardized and standardized alpha, true reliability, and the underlying measurement Model,” *J. Personal. Assess.* **93**(5), 445–453.
- Fastl, H., and Zwicker, E. (2006). *Psychoacoustics: Facts and Models*, 2nd ed. (Springer Science & Business Media, Heidelberg, Germany), 416 pp.
- Fisher, R. A. (1949). *The Design of Experiments* (Oliver and Boyd, Edinburgh, Scotland), 242 pp.
- FitzGerald, D. (2010). “Harmonic/percussive separation using median filtering,” in *Proceedings of the 13th International Conference on Digital Audio Effects (DAFX10)*, Graz, Austria, 4 pp.
- FitzGerald, D. (2012). “Vocal separation using nearest neighbours and median filtering,” in *23rd IET Irish Signals and Systems Conference*, NUI Maynooth, pp. 1–5.
- Foresee, D., and Hagan, M. (1997). “Gauss-Newton approximation to Bayesian learning,” in *Proceedings of the 1997 International Joint Conference on Neural Networks*, Vol. 3, pp. 1930–1935.
- Friberg, A., and Hedblad, A. (2011). “A comparison of perceptual ratings and computed audio features,” in *Proceedings of the SMC2011, 8th Sound and Music Computing Conference*, pp. 122–127.
- Friberg, A., Schoonderwaldt, E., and Hedblad, A. (2011). “Perceptual ratings of musical parameters,” in *Measured Interpretation: Computer-Supported Performance Analysis in the Cross Examination of Disciplines*, edited by von H. Loesch, and S. Weinzierl (Schott, Mainz, Germany), pp. 237–253.
- Friberg, A., Schoonderwaldt, E., Hedblad, A., Fabiani, M., and Elowsson, A. (2014). “Using listener-based perceptual features as intermediate representations in music information retrieval,” *J. Acoust. Soc. Am.* **136**(4), 1951–1963.
- Geringer, J. M. (1995). “Continuous loudness judgments of dynamics in recorded music excerpts,” *J. Res. Music Ed.* **43**(1), 22–35.
- Glasberg, B. R., and Moore, B. C. (2002). “A model of loudness applicable to time-varying sounds,” *J. Audio Eng. Soc.* **50**(5), 331–342.
- Glorot, X., Bordes, A., and Bengio, Y. (2011). “Deep sparse rectifier neural networks,” in *International Conference on Artificial Intelligence and Statistics*, pp. 315–323.
- Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning* (MIT Press).
- Hajda, J. M., Kendall, R. A., Carterette, E. C., and Harshberger, M. L. (1997). “Methodological issues in timbre research,” in *Perception and Cognition of Music*, edited by I. Deliège and J. Sloboda [Hove, England: Psychology Press/Erlbaum (UK) Taylor & Francis].
- Hansen, L. K., and Salamon, P. (1990). “Neural network ensembles,” *IEEE Trans. Pattern Anal. Mach. Intell.* **12**, 993–1001.
- Headey, B., and Wearing, A. (1989). “Personality, life events, and subjective well-being: Toward a dynamic equilibrium model,” *J. Personal. Soc. Psychol.* **57**(4), 731–739.
- Heittola, T., Klapuri, A., and Virtanen, T. (2009). “Musical instrument recognition in polyphonic audio using source-filter model for sound separation,” in *Proceedings of the International Symposium on Music Information Retrieval*, pp. 327–332.
- Herrera, P., Serra, X., and Peeters, G. (1999). “Audio descriptors and descriptors schemes in the context of MPEG-7,” *Proceedings of the 1999 International Computer Music Conference*, 4 pp.
- Ho, T. K. (1998). “The random subspace method for constructing decision forests,” *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(8), 832–844.
- ITU (2006). “ITU-R BS.1770-1: Algorithms to measure audio programme loudness and true-peak audio level; BS Series,” *Tech. rep.*, Int. Telecommunications Union.
- Kosta, K., Bandtlow, O. F., and Chew, E. (2015). “A change-point approach towards representing musical dynamics,” in *Mathematics and Computation in Music*, MCM 2015 Lecture Notes in Computer Science Vol. 9110, edited by T. Collins, T. Meredith, and A. Volk (Springer, Cham).
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). “Imagenet classification with deep convolutional neural networks,” *Adv. Neural Inf. Process. Syst.* **25**, 1097–1105.
- Lartillot, O., and Toiviainen, P. (2007). “A Matlab toolbox for musical feature extraction from audio,” *Proceedings of International Conference on Digital Audio Effects*, pp. 237–244.
- Logan, B. (2000). “Mel frequency Cepstral coefficients for music modeling,” *Proceedings of International Symposium on Music Information Retrieval*, 13 pp.
- Luce, D., and Clark, M. J. (1967). “Physical correlates of brass-instrument tones,” *J. Acoust. Soc. Am.* **42**(6), 1232–1243.

- Luce, D. A. (1975). "Dynamic spectrum changes of orchestral instruments," *J. Audio Eng. Soc.* **23**(7), 565–568.
- MacKay, D. J. C. (1992). "Bayesian interpolation," *Neural Comput.* **4**(3), 415–447.
- Mandel, M. I., and Ellis, D. P. (2008). "Multiple-instance learning for music information retrieval," in *Proceedings of the International Symposium on Music Information Retrieval*, pp. 577–582.
- Maron, O., and Lozano-Pérez, T. (1998). "A framework for multiple-instance learning," *Adv. Neural Inf. Process. Syst.* **11**, 570–576.
- Marquardt, D. (1963). "An algorithm for least-squares estimation of nonlinear parameters," *SIAM J. Appl. Math.* **11**(2), 431–441.
- McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., and Krimphoff, J. (1995). "Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes," *Psychol. Res.* **58**(3), 177–192.
- Nakamura, T. (1987). "The communication of dynamics between musicians and listeners through musical performance," *Percept. Psychophys.* **41**(6), 525–533.
- Olson, H. F. (1972). "The measurement of loudness," *Audio Mag.* 18–22.
- Orio, N. (2006). "Music retrieval: A tutorial and review," *Found. Trends Inf. Retrieval* **1**(1), 1–90.
- Polikar, R. (2006). "Ensemble based systems in decision making," *IEEE Circuits Syst. Mag.* **6**(3), 21–45.
- Sandvold, V., and Herrera, P. (2005). "Towards a semantic descriptor of subjective intensity in music" in *Proceedings of 5th International Conference on Music Information Retrieval*, 4 pp.
- Santos, J. R. A. (1999). "Cronbach's alpha: A tool for assessing the reliability of scales," *J. Ext.* **37**(2), 1–5.
- Schlüter, J. (2016). "Learning to pinpoint singing voice from weakly labeled examples," in *Proceedings of the 17th International Society for Music Information Retrieval Conference*.
- Schörkhuber, C., and Klapuri, A. (2010). "Constant-Q transform toolbox for music processing," in *Proceedings of the 7th Sound and Music Computing Conference*, Barcelona, Spain, pp. 322–330.
- Schörkhuber, C., Klapuri, A., Holighaus, N., and Dörfler, M. (2014). "A Matlab toolbox for efficient perfect reconstruction time-frequency transforms with log-frequency resolution," *Audio Engineering Society 53rd International Conference*, London, UK, 8 pp.
- Sollich, P., and Krogh, A. (1996). "Learning with ensembles: How overfitting can be useful," in *Advances in Neural Information Processing Systems*, edited by D. Touretzky, M. Mozer, and M. Hasselmo (MIT Press, Denver, CO), Vol. 8, pp. 190–196.
- Tzanetakis, G., and Cook, P. (2002). "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.* **10**(5), 293–302.
- Weninger, F., Eyben, F., and Schuller, B. (2014). "On-line continuous-time music mood regression with deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5412–5416.
- Zils, A., and Pachet, F. (2003). "Extracting automatically the perceived intensity of music titles," in *Proceedings of the 6th COST-G6 Conference on Digital Audio Effects (DAFX03)*, 4 pp.