Munoz-Montoro *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2019) 2019:23
https://doi.org/10.1186/s13636-019-0168-6

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Online/offline score informed music signal decomposition: application to minus one

Antonio Jesús Munoz-Montoro*, Julio José Carabias-Orti, Pedro Vera-Candeas, Francisco Jesús Canadas-Quesada and Nicolás Ruiz-Reyes

Abstract

In this paper, we propose a score-informed source separation framework based on non-negative matrix factorization (NMF) and dynamic time warping (DTW) that suits for both offline and online systems. The proposed framework is composed of three stages: training, alignment, and separation. In the training stage, the score is encoded as a sequence of individual occurrences and unique combinations of notes denoted as score units. Then, we proposed a NMF-based signal model where the basis functions for each score unit are represented as a weighted combination of spectral patterns for each note and instrument in the score obtained from a trained a priori over-completed dictionary. In the alignment stage, the time-varying gains are estimated at frame level by computing the projection of each score unit basis function over the captured audio signal. Then, under the assumption that only a score unit is active at a time, we propose an online DTW scheme to synchronize the score information with the performance. Finally, in the separation stage, the obtained gains are refined using local low-rank NMF and the separated sources are obtained using a soft-filter strategy. The framework has been evaluated and compared with other state-of-the-art methods for single channel source separation of small ensembles and large orchestra ensembles obtaining reliable results in terms of SDR and SIR. Finally, our method has been evaluated in the specific task of acoustic minus one, and some demos are presented.

Keywords: Score-informed source separation, Non-negative matrix factorization, Dynamic time warping, Online source separation, Minus one

1 Introduction

Sound source separation (SS) seeks to segregate constituent sound sources from an audio signal mixture. Once separated, the sources can be processed separately and reassembled eventually for many purposes such as denoising, remastering, or desoloing.

During the last decade, there has been a growing demand for streaming music content, to such an extent that currently there are several entertainment platforms such as MyOpera Player[1] and *Medici.TV*[2], which broadcast live classical music with enriched features. These classical music services could be improved using SS techniques to develop applications such as 3D rendering, acoustic emphasis, acoustic scenarios recreation, and minus one.

Many approaches have been addressed in the last two decades in order to achieve this separation. The most commonly used consists of decomposing a time-frequency representation of the mixture signal using methods such as non-negative matrix factorization (NMF), independent component analysis (ICA), or probabilistic latent component analysis (PLCA). Among these factorization techniques, NMF has been widely used for music audio signals, as it allows to describe the signal as a non-substractive combination of sound objects (or "atoms") over time. However, without further information, the quality of the separation using the aforementioned statistical methods is limited. One solution

*Correspondence: jmontoro@ujaen.es
Department of Telecommunication Engineering, University of Jaen, 23700, Linares Jaen, Spain

[1] https://www.myoperaplayer.com/
[2] https://www.medici.tv/

is to exploit the spectro-temporal properties of the sources. For example, spectral harmonicity and temporal continuity can be assumed for several musical instruments while percussive instruments are characterized by short bursts of broadband energy [1]. Speech source spectrogram can be modeled using a source-filter model [2]. Other approaches also used spatial localization of the sources [3–5]. Besides, when training material is available, it is possible to learn the spectro-temporal patterns and the methods are referred to as supervised [6].

Best separation results are obtained when information about the specific sources in the mixture is provided a priori. For example, in [7], information about the spatial location for each source in the mixture is known a priori. In [8], the authors combined prior information from an aligned score with a panning and time-frequency-based method for SS of synthetic music signals. In fact, due to the widespread availability of musical scores, mostly in MIDI format, an increasing number of score-informed SS approaches have been conducted lately. Ganseman et al. [9] uses score synthesis to initialize a signal decomposition monaural SS system using a PLCA model. Similarly, in [10], the musical score is used to initialize a parametric NMF of the mixture spectrum. Ewert and Muller [11, 12] use the score to constraint the basis functions and the time-varying gains assuming harmonicity and allowing some misalignment using a tolerance window over the score note activity. The approach in [6] uses score information to adapt the model parameters using an NMF - based approach. In Miron et al. [13], score information are used to initialize the time-varying gains in a multichannel NMF model for orchestra music SS. More recently, novel deep learning (DL) strategies have been developed combining deep neural networks (DNN) with score information to estimate soft masks for specific instrument classes [14–16].

In the mentioned approaches, the score of the pieces must be previously aligned to the recording; this synchronization is usually performed beforehand and is typically obtained using a twofold procedure: (1) feature extraction from audio and score and (2) temporal alignment [17]. In the former, the features extracted from the audio signal characterize some specific information about the musical content. Different representations of the audio frame have been used such as the output of a short-time Fourier transform (STFT) [18], chroma vectors [19, 20], or multi-pitch analysis information [21, 22]. On the latter, the alignment is performed by finding the best match between the feature sequence and the score. In fact, classical offline systems rely on cost measures between events in the score and in the performance. Two methods well known in speech recognition have been extensively used in the literature: statistical approaches (e.g., hidden Markov model (HMM)) [22–26] and dynamic time warping (DTW) [19, 27–29].

Although several online audio-to-score approaches have been developed in the literature [22, 28, 29], only the works in [6, 22, 30] combined score alignment with SS in an online fashion. In [22] and in its extension [6], the alignment is performed using a hidden Markov process model, where each audio frame is associated with a 2-D state of score position and tempo. The observation model is defined as the multi-pitch likelihood of each frame, i.e., the likelihood of seeing the audio frame given the pitches at the aligned score position. Then, particle filtering is employed to infer the score position and tempo of each audio frame by the time it is captured. Regarding the SS, [22] uses a soft masking strategy based on a multipitch estimator, whereas the method in [6] used a frame-level NMF with a trained dictionary which is updated during the factorization. On the other hand, in our preliminary work [30], we proposed to use the source separation procedure presented in [13] along with the real-time implementation of the online alignment method from [31] presented in [32]. The signal model used in [30] was the same than in [13] but restricted to single-channel signals. However, with this original signal model, the separation performance was unreliable when dealing with large ensembles datasets.

In this paper, we propose a score-informed SS framework, which suits for both online and offline applications, based on NMF and DTW. Similar to [6, 30], we use a priori learned dictionary composed of spectral templates (a.k.a basis functions) for the instruments presented in the score. However, in this work, the score information is encoded within the signal model under the assumption that a music signal can be described as a sequence of unique occurrences of individual and concurrent notes from several instruments (here denoted as score units). Then, the cost function for the alignment procedure can be obtained by computing the projection for each score unit over the frame level spectrum and the minimum cost path is estimated using the online DTW framework proposed in [29] and based on the original work presented by Dixon et al. [28]. To account for possible misalignments, the time-varying gains of the optimal score units and its neighbors are refined using a local low-rank NMF scheme. Finally, a soft-filter strategy is performed to obtain the separated sources. The proposed framework has been evaluated for both online and offline SS tasks and compared with other state-of-the-art methods showing reliable results, specially in terms of SDR and SIR.

Unlike a previous work presented by the authors in [29, 30], where synthetic signals generated from the MIDI score are used to learn spectral patterns for the alignment process, here, we use isolated notes of real solo instrument recordings to learn a dictionary of spectral patterns. Using this dictionary of isolated notes allows to model the relative amplitude contribution of each instrument in order to obtain a velocity model for each note and instrument. Additionally, in contrast to [6, 30], we propose a new signal model where the score information is encoded in the form of score units. Consequently, the novelty of this work lies in developing a method for single-channel and multi-timbral (i.e., with multiple instruments) signal SS, which uses the score information encoded within the signal model. Finally, an extension of the method has been developed in order to be applied to a practical music scenario, concretely minus one application.

The structure of the rest of the article is as follows. In Section 2, we briefly review the NMF and DTW approaches. The proposed framework is presented in Section 3. A minus one application strategy based on our framework is presented in Section 4. In Section 5, the evaluation setup is presented and the proposed method is tested and compared with other reference systems. Finally, we summarize the work and discuss the future perspectives in Section 6.

2 Background

2.1 NMF for source separation

NMF [33] is an unsupervised factorization technique used for linear representation of two-dimensional non-negative data that has been successfully applied to the decomposition of audio spectrograms [34–38]. In the context of audio signal processing, the popularity of this technique is related to its ability to obtain parts-based representation of the most representative objects (e.g., notes and chords) by imposing non-negative constraints that allow only additive, not subtractive, combinations of the input data unlike many other linear representations such as ICA [39] and principal component analysis (PCA) [40].

Given an input audio signal $\mathbf{x}(t)$ which is constituted by the mixture of J sources and whose magnitude (or power) spectrogram $\mathbf{X} \in \mathbb{R}_+^{F\times T}$ is composed of $f = 1, ..., F$ frequency bins, $t = 1, ..., T$ time frames, and a linear combination of L components (see Table 1 for the notations we adopted in this paper), NMF finds an approximate factorization $\hat{\mathbf{X}} \in \mathbb{R}_+^{F\times T}$ as follows:

Table 1 Nomenclature, symbols, and notation

M	Matrices in bold upper case
$\mathbf{M}_i(j)$	Entry vector M_i from a matrix M (bold upper case)
$\odot$	The Hadamard (element-wise) multiplication
$\circ$	The convolution operation
^{B}	Over a matrix indicates that it is a binary matrix
$\hat{}$	Over a matrix indicates that it is an estimated matrix

$$\mathbf{X}(f,t) \approx \hat{\mathbf{X}}(f,t) = \sum_l \mathbf{W}_l(f)\mathbf{G}_l(t) \tag{1}$$

being $\mathbf{W} \in \mathbb{R}_+^{F\times L}$ the basis matrix whose columns are meaningful elements called basis functions (or spectral patterns) and $\mathbf{G} \in \mathbb{R}_+^{L\times T}$ the activation matrix that shows the temporal activity for each l individual basis function. Note that the number of components (a.k.a rank) L is generally chosen such that $FL + LT \ll FT$ in order to reduce the dimensions of the data associated to the input spectrogram. In the context of musical sound separation, we assume that each pair of spectral pattern and activations describes a sound from a single instrument. Then, each source spectrogram can be obtained by grouping all the sounds belonging to the same source.

The model parameters in Eq. (1) are obtained by minimizing a cost function $D(\mathbf{X}|\hat{\mathbf{X}})$ defined in Eq. (2), using a so called factorization procedure,

$$D(\mathbf{X}|\hat{\mathbf{X}}) = \sum_{f,t} d(\mathbf{X}(f,t)|\hat{\mathbf{X}}(f,t)) \tag{2}$$

where $d(a|b)$ is a function of two scalar variables. The most popular cost functions are the Euclidean distance (EUC), the generalized Kullback-Leibler divergence (KL), and the Itakura-Saito divergence (IS) [41]. The β-divergence [42] (see Eq. (3)) is another commonly used cost function that encompasses the three previously mentioned cost functions in its definition, i.e., EUC (β = 2), KL (β = 1), and IS (β = 0), and is defined as follows:

$$D_\beta(\mathbf{X}|\hat{\mathbf{X}}) = \begin{cases} \sum_{f,t} \frac{1}{\beta(\beta-1)} \left(\mathbf{X}(f,t)^\beta + (\beta-1)\hat{\mathbf{X}}(f,t)^\beta - \beta \mathbf{X}(f,t)\hat{\mathbf{X}}(f,t)^{\beta-1} \right) & \beta \in (0,1) \cup (1,2] \\ \sum_{f,t} \mathbf{X}(f,t) \log \frac{\mathbf{X}(f,t)}{\hat{\mathbf{X}}(f,t)} - \mathbf{X}(f,t) + \hat{\mathbf{X}}(f,t) & \beta = 1 \\ \sum_{f,t} \frac{\mathbf{X}(f,t)}{\hat{\mathbf{X}}(f,t)} + \log \frac{\mathbf{X}(f,t)}{\hat{\mathbf{X}}(f,t)} - 1 & \beta = 0 \end{cases} \tag{3}$$

In order to obtain the model parameters that minimize the cost function and ensure the non-negativity of the bases and the activations, several approaches were developed. In the original formulation of NMF [33], $D(\mathbf{X}|\hat{\mathbf{X}})$ was minimized using an iterative approach based on the gradient descend algorithm. The multiplicative update rules are obtained by applying diagonal rescaling to the step size of the gradient descent algorithm. The multiplicative update rule for each scalar parameter Z is given by,

$$Z \leftarrow Z \odot \left(\frac{\nabla_Z^- D(\mathbf{X}|\hat{\mathbf{X}})}{\nabla_Z^+ D(\mathbf{X}|\hat{\mathbf{X}})} \right) \tag{4}$$

where $\nabla_Z^- D$ and $\nabla_Z^+ D$ are the negative and positive terms of the partial derivative of the cost function $\nabla_Z D$.

Unfortunately, without further constraints, the expressiveness (i.e., the disjointness) of the basis functions is limited. This property can be usually maximized by considering the sparsity of the representations (except when the energy time-frequency distribution of the sources completely overlap). Sparsity is the property of a signal that relates to the amount of non-zero coefficients in a given representation. Several criteria for sparsity have been proposed in the literature [43, 44]. Other approaches use restrictions on the spectro-temporal structure such as harmonicity of the basis functions [45] or temporal continuity of the gains [43, 46]. While such extensions typically lead to a significant gain in separation quality over classic NMF, they do not fully solve the problem.

Alternatively, when dealing with music signals and when the score information is available, certain priors can be imposed to the signal model parameters in order to favor sparsity. A review of the typical score-informed constraints in the literature will be presented in Section 2.3.

2.2 Deep learning approaches for source separation

Recently, DL approaches have outperformed NMF in audio source separation challenges [47]. In contrast to the NMF methods, DL methods are less computationally expensive [48] at the separation stage, as estimating the sources involves a single feed forward pass through the network rather than an iterative procedure. State-of-the-art DL methods typically estimate a soft mask for each specific source in the time-frequency domain, even though there are approaches that operate directly on time-domain signals and use a DNN to learn a suitable representation from it (see e.g., [16, 49]).

Using a soft-masking strategy, the time-frequency representation $\hat{\mathbf{S}}_i \in \mathbb{R}^{F \times T}$ for each source i in the mixture can be expressed as:

$$\hat{\mathbf{S}}_i(f,t) = \mathbf{M}_i(f,t)\mathbf{X}(f,t) \tag{5}$$

where $\mathbf{M}_i \in \mathbb{R}^{F \times T}$ represents the soft mask for each source i within the time-frequency mixture signal $\mathbf{X} \in \mathbb{R}^{F \times T}$. Typical time-frequency signal representations are the short-time Fourier transform (STFT), constant-Q, or mel spectrogram.

In the case of single-channel source separation, the quality of the separation relies on modeling the spectral structure of sources. In fact, DL methods can be classified in two categories: (1) methods that aim to predict the soft mask $\mathbf{M}_i$ based on the mixture input $\mathbf{X}$ [48, 50, 51] and (2) methods that aim to predict the source signal spectrum $\hat{\mathbf{S}}_i$ directly from the mixture input [52]. Consequently, supervised learning process learns the relation between the input mixture time-frequency representation and the target out, which could be either the oracle mask or the clean signal spectrum.

Several DL architectures have been used for the SS task including the use of standard methods such as convolutional [15] and recurrent [51] layers which have the advantage of modeling a larger time context. Novel DL SS systems propose specialized models which propose building an NMF logic into an autoenconder [53] or cluster components over large time spans [54].

2.3 Score-informed constraints

The number of digital scores freely available in the Internet is continuously growing. In fact, most of the classical music pieces are available without copyright in repositories such as IMSLP[3], Mutopia[4], or Classical Archives[5]. The most common format is MIDI, which includes information about the actual instruments in the score and the onset/duration of the played notes for each instrument. Recently, some other formats such as musicXML, MEI, or Lylipond have emerged to keep the attributes of the original paper sheet.

Traditionally, NMF methods using score information enforced the sparsity by imposing certain constraints on the basis functions $\mathbf{S}$ and/or the time-varying gains $\mathbf{A}$. For example, the musical structure of the score can be exploited to penalize activations of notes or combinations of notes which are not present in the score [34, 55]. Additionally, if the score is pre-aligned with the interpretation, it is possible to set to zero those gains associated to the basis functions of non-active notes. In fact, once an entry in $\mathbf{S}$ or $\mathbf{A}$ is initialized to zero, it will remain zero-valued during the subsequent multiplicative update steps [33]. Unfortunately, in music, imposing sparsity constraint only over the gains does not ensure the dissociation of the basis functions, since concurrent notes in the score are often harmonically related.

[3] https://imslp.org/
[4] http://www.mutopiaproject.org/
[5] https://www.classicalarchives.com/

To overcome this problem, it is usual to apply constraints simultaneously on both basis functions and activations. In the case of the basis functions, the most common approach is to assume that the templates in **S** posses a harmonic structure. In general, a harmonic sound is one whose energy in a time-frequency representation is distributed around integer multiples of the so-called fundamental frequency (a.k.a harmonics). To enforce this harmonicity, it is common to set to zero all the frequency bins of **S** between harmonics [34, 56]. Note that perfect tuning is very uncommon (i.e., the exact frequencies are not known). Therefore, a range of frequencies around the perfectly tuned F0 is kept non-zero valued in the basis functions. Some methods in the literature proposed to use a semitone frequency resolution by integrating all the bins belonging to the same semitone to avoid possible problems related to F0 deviations [57]. As a drawback, a low frequency resolution does not allow to separate pitches from overlapping harmonics from different instruments. Using a higher frequency resolution could mitigate this problem, but then, a F0 tracking is required [13].

Score information has been also used to improve the separation results of DL methods. For instance, in [14], the authors used weak label information from the score by introducing a score-unit-based dropout variant that discards those combinations of notes which not present in the score. In [15], the authors used pre-aligned score information to refine the output soft mask of their CNN-based network. Finally, in [16] label-conditioned information is used to inform the network about the active and inactive instruments in the mixture.

2.4 Dynamic time warping for audio to score alignment

In the previous section, we assumed that we had a temporal alignment between the score and performance times. In fact, manual alignment is very tedious since the musician usually interprets each piece in a personal way by introducing variations in the tempo, dynamics, and/or articulation. To automate this process, there are several methods for estimating a temporal alignment between score and audio representations, a task also referred to as score-audio synchronization.

Up to date, DTW-based methods have demonstrated to provide the best alignment results in the MIREX Real-time Audio to Score Alignment[6] task, and therefore, as will be explained in the next section, we have chosen this approach as the basis of the alignment procedure in our system.

Let us define $\mathbf{U} = (u_1, \ldots, u_\tau, \ldots, u_{T_m})$ and $\mathbf{V} = (v_1, \ldots, v_t, \ldots, v_{T_r})$ as two vectors of features that represent the time series to be aligned, where τ and t are the indexes in the time series. The first step in the DTW algorithm is the estimation of a local distance matrix (cost matrix) $\mathbf{D}(\tau, t) = \Psi(u_\tau, v_t)$, where $\mathbf{D} \in \mathbb{R}_+^{T_m \times T_r}$ represents the match cost between every two points in the time series. The function Ψ could be any cost function that returns cost 0 for a perfect match, and a positive value otherwise. In the field of audio to score alignment, this cost function is typically computed from the score synthesized audio signal and the recorded audio from the performance, for example by computing the cosine distance between their STFT magnitudes [58] or the Euclidean distance between their chroma vectors [28]. Secondly, the warping matrix $\mathbf{C} \in \mathbb{R}_+^{T_m \times T_r}$ is filled recursively as follows:

$$\mathbf{C}(\tau, t) = \min \left\{ \begin{array}{l} \mathbf{C}(\tau-1, t-1) + \sigma_{1,1}\mathbf{D}(\tau, t) \\ \mathbf{C}(\tau-2, t-1) + \sigma_{2,1}\mathbf{D}(\tau, t) \\ \vdots \\ \mathbf{C}(\tau-\alpha_r, t-1) + \sigma_{\alpha_r,1}\mathbf{D}(\tau, t) \\ \mathbf{C}(\tau-1, t-2) + \sigma_{1,2}\mathbf{D}(\tau, t) \\ \vdots \\ \mathbf{C}(\tau-1, t-\alpha_t) + \sigma_{1,\alpha_t}\mathbf{D}(\tau, t) \end{array} \right\} \quad (6)$$

where the step size at each dimension has a range from 1 to α_τ and 1 to α_t, respectively. α_τ and α_t are the maximum step size at each dimension. Parameter σ controls the bias toward diagonal steps. **C** is the accumulated cost of the minimum cost path up to (τ, t) and $\mathbf{C}(\tau, 1) = \mathbf{D}(\tau, 1), \forall \tau$.

Finally, the minimum cost path $w = w_1, ..., w_l, ..., w_L$, where each w_l is an ordered pair (τ_l, t_l) meaning that instant τ_l must be aligned with t_l , is obtained by tracing the recursion backwards from $\mathbf{C}(\tau_L, T_r)$, where $\tau_L = \arg\min_\tau \mathbf{C}(\tau, T_r)$. Since the audio query is usually a small fragment of the whole composition, the first and the last elements of the path can be at any point along the τ axis. Globally, the path has to satisfy the following three conditions:

1. Boundary condition: $w_1 = (\tau, 1)$ and $w_L = (\tau, T_r)$
2. Monotonicity condition: $\tau_{l+1} \geq \tau_l$ and $t_{l+1} \geq t_l$
3. Step size condition: $\tau_{l+1} \leq \tau_l + \alpha_\tau$ and $t_{l+1} \leq t_l + \alpha_t$

In the next section, we present a signal factorization-based system that uses DTW during the factorization to perform both synchronization and SS in a joint manner.

3 Proposed method

In this paper, we propose a score-informed audio SS framework that is suitable for both online and offline

[6] The Music Information Retrieval Evaluation eXchange (MIREX) is an annual evaluation campaign for MIR algorithms. Real-time Audio-to-Score Alignment (a.k.a. Score Following) is one of the evaluation tasks. http://www.music-ir.org/mirex

systems. In particular, we propose a signal decomposition-based model that enables the alignment between the score and the performance jointly with the SS of the different type of instruments in the mixture.

The block diagram of the proposed method is displayed in Fig. 1. First, in the training stage, we initilize the basis functions using a set of pre-trained spectral patterns for the instruments in the score and estimate the relative amplitude between instruments from the synthetic input score. Second, in the alignment stage, we used a NMF-based signal decomposition approach to estimate the model parameters by minimizing the signal reconstruction error. Then, an online DTW scheme on the resulting cost matrix is used to estimate the alignment between the score and the performance. Third, at frame level, a low rank factorization stage is used to refine the model estimated parameters accounting to the most probable active notes in the score at this frame. Finally, a generalized Wiener filtering strategy is used to obtain the sources reconstruction.

3.1 Signal model

In this work, score information and elementary spectral patterns for the target sources (i.e., the instruments presented in the score) are combined within the signal model. First, a piano-roll matrix $\mathbf{H}^B \in \mathbb{N}^{P\times J\times T_m}$ is inferred from the input MIDI score, where $p \in[1...P]$ represents the notes in the MIDI scale, $j \in[1...J]$ are the instruments in the score, and $\tau \in[1...T_m]$ is the MIDI time (in frames). Then, the score information is encoded into score units using an approach similar to [29]. In particular, score units $k \in[1...K]$ are defined from each unique occurrence of individual or concurrent notes in the score. Note that, for most of the scores, the number of score units will be much lower than the number of notes (generalizing $K \leq P$). Under this representation, the score can be seen as a sequence of states, where each state is defined by a single unit. Note that multiple states can be represented by the same unit, whereas the units are unique. Therefore, $K \leq M$, where M is the number of states.

The proposed score model is defined as follows:

$$\mathbf{H}^B_{p,j}(\tau) = \sum_k \mathbf{E}^B_{p,j,k} \underbrace{\sum_m \mathbf{Q}^B_{k,m}\, \mathbf{P}^B_m(\tau)}_{\mathbf{A}^B_k(\tau)} \qquad (7)$$

where the piano-roll matrix $\mathbf{H}^B$ is decomposed as a notes-to-unit matrix $\mathbf{E}^B \in \mathbb{N}^{P\times J\times K}$ that encodes the active instruments j and the notes p at each score unit k and a unit activity matrix $\mathbf{A}^B \in \mathbb{N}^{K\times T_m}$ that represents the active score units at each MIDI frame τ. Note that the superscript ^{B} over a matrix indicates that it is a binary matrix. Besides, the unit activity can be subsequently decomposed into a units-to-states matrix $\mathbf{Q}^B \in \mathbb{N}^{K\times M}$ which encodes the active unit k at each state m and a states-time matrix $\mathbf{P}^B \in \mathbb{N}^{M\times T_m}$ that represents the activity of each state m at each MIDI time frame τ. The presented score model in Eq. (7) is displayed in Fig. 2. Note that, for the definition of the score model parameters, only the MIDI events note-on and note-off messages are used from the MIDI file. These messages together specify the start and end time of notes played at a given pitch of a given instrument. Therefore, timing information such as beat resolution or tempo changes are not used in our definition.

The proposed signal model combines spectral patterns for each instrument together with the score model once it is synchronized with the audio time ($\tau \rightarrow t$). Therefore, the signal model presented in Eq. (8) consists of factorizing the time-frequency spectrogram of the audio mixed

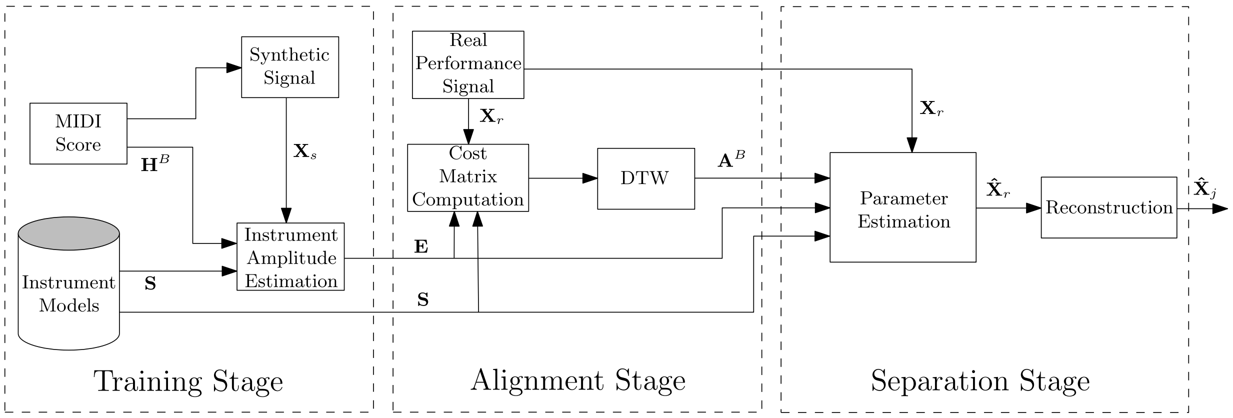
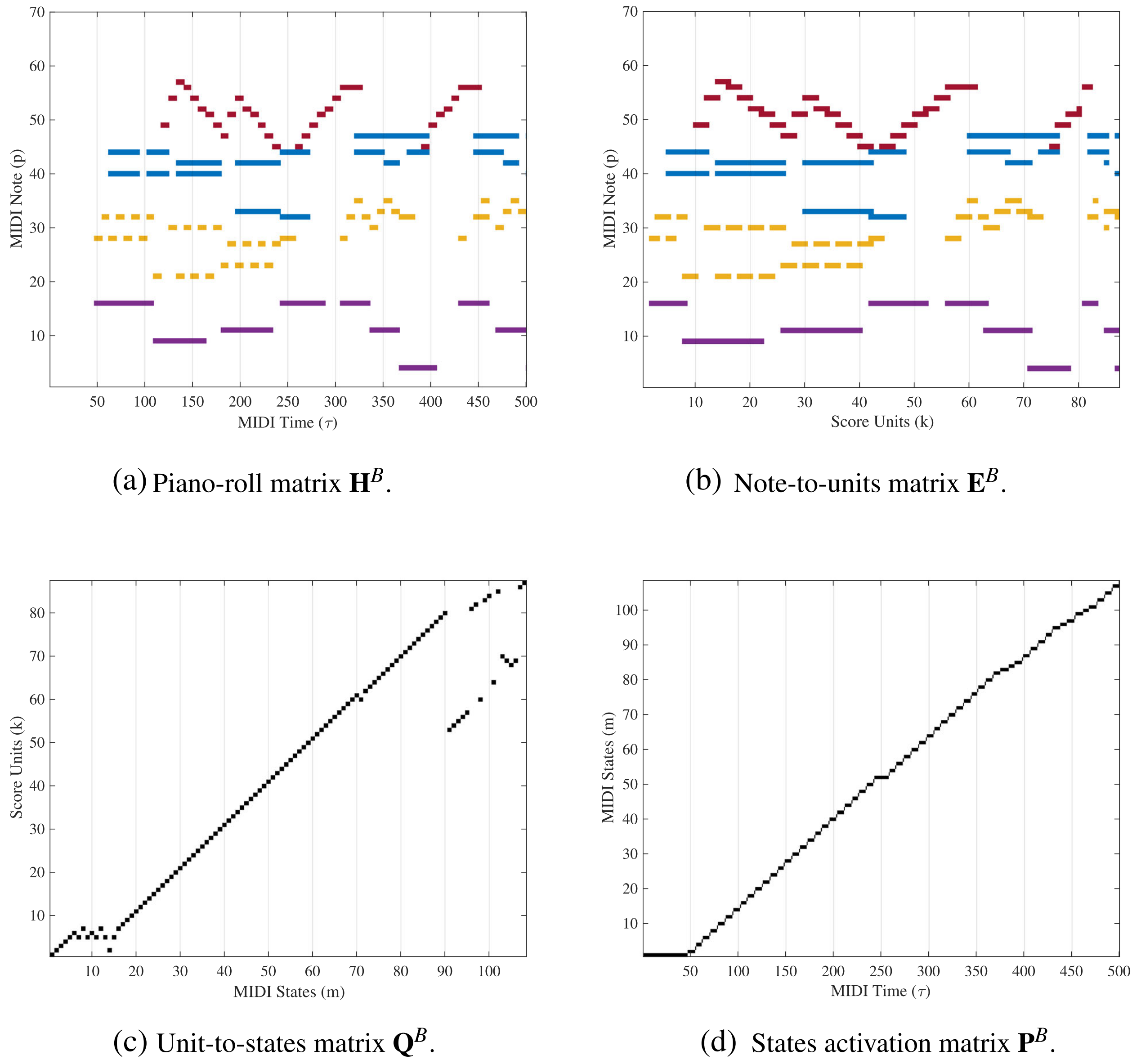


Fig. 1 Block diagram of the proposed system



(a) Piano-roll matrix $\mathbf{H}^B$.

(b) Note-to-units matrix $\mathbf{E}^B$.

(c) Unit-to-states matrix $\mathbf{Q}^B$.

(d) States activation matrix $\mathbf{P}^B$.

Fig. 2 Example of score matrix decomposition. **a** Piano-roll matrix $\mathbf{H}^B$. **b** Note-to-units matrix $\mathbf{E}^B$. **c** Unit-to-states matrix $\mathbf{Q}^B$. **d** States activation matrix $\mathbf{P}^B$. The instruments (dimension *j*) in **a** and **b** are displayed using different colors

signal into the product of three nonnegative matrices $\mathbf{S} \in \mathbb{R}_+^{F\times P\times J}$, $\mathbf{E} \in \mathbb{R}_+^{P\times J\times K}$, and $\mathbf{A} \in \mathbb{R}_+^{K\times T_r}$ as

$$\mathbf{X}(f,t) \approx \hat{\mathbf{X}}(f,t) = \sum_{p,j} \mathbf{S}_{p,j}(f)\, \overbrace{(\sum_k \mathbf{E}_{p,j,k}\mathbf{A}_k(t))}^{\mathbf{H}_{p,j}(t)} \qquad (8)$$

where $\mathbf{X} \in \mathbb{R}_+^{F\times T_r}$ is the mixture signal spectrogram with $f \in [1...F]$ frequency bins and $t \in [1...T_r]$ frames, $\hat{\mathbf{X}} \in \mathbb{R}_+^{F\times T_r}$ is the estimation of this magnitude spectrogram, $\mathbf{S}$ is the basis matrix, $\mathbf{E}$ is the matrix that represents the relative amplitude contribution of each instrument and note at each score unit, and $\mathbf{A}$ is the time-varying gains matrix. Notice that this proposed signal model is an extension of the baseline model presented in [30],

$$\hat{\mathbf{X}}(f,t) = \sum_{p,j} \mathbf{S}_{p,j}(f)\mathbf{H}_{p,j}(t) \qquad (9)$$

where $\mathbf{H} \in \mathbb{R}_{+}^{P\times J\times T_r}$ is the matrix which holds the gains of the basis $\mathbf{S}$ corresponding to the note p and instrument j at frame t.

Please note that in this model, we do not consider the possible improvisations or errors (e.g., note deletion, insertion, and substitution) during the performance but rather assume that the musician will follow the written score. In fact, large deviations may cause an underperformance of the proposed method for both alignment and separation procedures.

The proposed score-informed signal model is illustrated in Fig. 3.

3.2 Training stage

In this stage, we learn the spectral basis function $\mathbf{S}$ and the relative amplitude of the notes-to-unit matrix $\mathbf{E}$ in Eq. (8). First, the spectral basis function $\mathbf{S}$ for each instrument j in the score is selected from a pre-trained dictionary of spectral patterns for all the notes and instruments from the Real World Computing (RWC) Musical Instrument Sound Database [59, 60]. In this work, we have used an approach similar to [57], where the instrument-dependent bases are learned in advance and fixed during the separation process. This approach has been shown to perform well when the conditions of the music scene do not differ too much between the training and the test data [61].

Second, the relative amplitude of the notes-to-unit matrix $\mathbf{E}$ in Eq. (8) is initialized to account for the velocity of each note from a particular instrument. To this aim, we synthesize the MIDI score signal $\mathbf{X}_s \in \mathbb{R}_{+}^{F\times T_m}$ and propose the following signal decomposition model:

$$\mathbf{X}_s(f,\tau) \approx \hat{\mathbf{X}}_s(f,\tau) = \sum_{k,p,j} \mathbf{S}_{p,j}(f)\mathbf{E}_{p,j,k}\mathbf{A}_k(\tau) \qquad (10)$$

where $\tau \in [1...T_m]$ represents the time (in frames) of the score, the spectral patterns $\mathbf{S}$ are known a priori, and the unit activity matrix $\mathbf{A} \in \mathbb{R}_{+}^{K\times T_m}$ is directly initialized to the binary piano-roll matrix $\mathbf{A}^B$ from the score model in Eq. (7). Note that spectral patterns $\mathbf{S}$ are kept fixed, whereas the optimal value for the notes-to-unit matrix $\mathbf{E}$ and the unit activity matrix $\mathbf{A}$ are obtained using the gradient descend algorithm to minimize the β-divergence cost function $D_\beta(\mathbf{X}_s, \hat{\mathbf{X}}_s)$. In particular, the update rules to obtain the parameter $\mathbf{E}$ and $\mathbf{A}$ are defined in Eqs. (11) and (12), respectively and computed iteratively until the cost function converges.

$$\mathbf{E}_{p,j,k} \leftarrow \mathbf{E}_{p,j,k} \odot \left(\frac{\sum_{f,\tau} \mathbf{S}_{p,j}(f)[\hat{\mathbf{X}}_s(f,\tau)^{\beta-2} \odot \mathbf{X}_s(f,\tau)]\,\mathbf{A}_k(\tau)}{\sum_{f,\tau} \mathbf{S}_{p,j}(f)\hat{\mathbf{X}}_s(f,\tau)^{\beta-1}\mathbf{A}_k(\tau)} \right) \qquad (11)$$

$$\mathbf{A}_k(\tau) \leftarrow \mathbf{A}_k(\tau) \odot \left(\frac{\sum_{f,p,j} \mathbf{S}_{p,j}(f)[\hat{\mathbf{X}}_s(f,\tau)^{\beta-2} \odot \mathbf{X}_s(f,\tau)]\,\mathbf{E}_{p,j,k}}{\sum_{f,p,j} \mathbf{S}_{p,j}(f)\hat{\mathbf{X}}_s(f,\tau)^{\beta-1}\mathbf{E}_{p,j,k}} \right) \qquad (12)$$

In addition, scaling the parameters is necessary to ensure that $\mathbf{E}$ models only the relative amplitude between instruments. The scaling procedure is presented as follows:

$$e_k = \sqrt{\sum_{p,j} \mathbf{E}_{p,j,k}^2} \quad , \quad \mathbf{E}_{p,j,k} = \frac{\mathbf{E}_{p,j,k}}{e_k} \quad , \quad \mathbf{A}_k(\tau) = e_k\mathbf{A}_k(\tau) \qquad (13)$$

As Fig. 1 outlines, after this training process, both the spectral basis function $\mathbf{S}$ and the relative amplitude of the notes-to-unit matrix $\mathbf{E}$ are initialized for the separation stage.

3.3 Alignment stage

In this stage, the aim is to synchronize the audio signal of the musical performance with its corresponding score to guide the separation process. The proposed method can be carried out in an offline or online manner depending on the application.

In this work, the alignment is performed using a similar scheme than in [31]. As a novel proposal, a single spectral pattern for each score unit is first computed from the trained parameters in Section 3.2 as follows:

$$\mathbf{B}_k(f) = \sum_{p,j} \mathbf{S}_{p,j}(f)\mathbf{E}_{p,j,k} \qquad (14)$$

Then, generalizing the concurrent notes in the score as units, the score can be seen as a sequence of individual

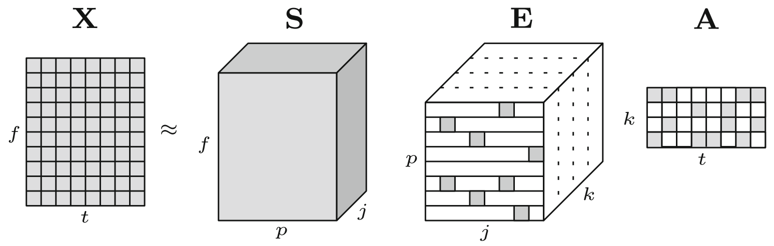


Fig. 3 Proposed signal model parameters. Zero values are displayed in white and non-zero values in gray

units. In other words, for a particular frame t in the real-world mixture signal spectrogram $\mathbf{X}_t(f)$, we can assume that only a single unit k can be active. Under this assumption, the time-varying gain $g_{k,t}$ for each score unit k and frame t can be obtained minimizing the β-divergence of the single-unit constrained signal model:

$$D_\beta(\mathbf{X}_t(f)|g_{k,t}\mathbf{B}_k(f)) = \sum_f \frac{1}{\beta(\beta-1)}\left[\mathbf{X}_t(f)^\beta + (\beta-1)(g_{k,t}\mathbf{B}_k(f))^\beta - \beta\mathbf{X}_t(f)(g_{k,t}\mathbf{B}_k(f))^{\beta-1}\right] \tag{15}$$

Therefore, the value of the gain for unit k and frame t is obtaining as the projection of its corresponding spectral pattern $\mathbf{B}_k(f)$ over the observed signal spectrogram $\mathbf{X}_t(f)$ as,

$$g_{k,t} = \frac{\sum_f \mathbf{X}_t(f)\mathbf{B}_k(f)^{(\beta-1)}}{\sum_f \mathbf{B}_k(f)^\beta} \tag{16}$$

Then, the distortion matrix for each unit k at the current frame t is defined by,

$$\boldsymbol{\Phi}_{\boldsymbol{k,t}} = D_\beta(\mathbf{X}_t(f)|g_{k,t}\mathbf{B}_k(f)) \tag{17}$$

where $D_\beta(\cdot)$ is the β-divergence function and β can take values in the range $\in$ [0,2].

As can be inferred, the distortion matrix $\boldsymbol{\Phi} \in \mathbb{R}_+^{K\times T_r}$ provides information about the similitude of each k-th unit spectral pattern with the real signal spectrum at each frame t. Using this information, the cost matrix between the score and the performance time, τ and t, respectively, can be obtained as follows:

$$\mathbf{D}(\tau,t) = \sum_k \mathbf{A}_k^B(\tau)\Phi_{k,t} \tag{18}$$

where $\mathbf{A}^B$ is the unit activity matrix in Eq. (7) extracted from the score information and Φ is the distortion matrix in Eq. (17) computed from the real-world mixture signal.

Note that the cost matrix is computed as the projection of each score unit spectral pattern over each captured audio frame. This computation is a non-iterative process, which enables fast runtimes. However, for the case of long musical compositions, this implementation requires to estimate the cost function for each score unit at each frame making the system unsuitable for real-time applications. To mitigate these computational requirements, a temporal window from the optimum unit in the previous frame can be used, estimating the cost function only for the score units within that threshold.

As can be seen, the proposed framework operates at frame level, and therefore, depending on the application, the alignment can be performed offline (i.e., once the whole input signal has been processed) or online with a fixed latency of one frame. In this work, the synchronization between score and performance is done using DTW (see Section 2.4). For each frame t in the performance, a warping matrix $\mathbf{C} \in \mathbb{R}_+^{T_m\times T_r}$ is computed from the cost matrix $\mathbf{D} \in \mathbb{R}_+^{T_m\times T_r}$ using Eq. (6). In this work, the step size α_τ and α_t range from 1 to 4 which in terms of music performance means that a performer can play four times faster or slower than the reference speed of interpretation ($\alpha_\tau = 1, \alpha_t = 1$). Besides, the control parameter $\sigma_{\alpha_\tau,\alpha_t}$ has been set to one for all the combinations of (α_τ, α_t) which in turn bias the path towards the diagonal.

In the offline scheme, the optimum path is computed using backtracking after the warping matrix for the whole real performance input signal is computed (see detail information in [31]). Consequently, the alignment decision for every state cannot be known until the whole piece is processed.

Alternatively, in the online scheme, the signal is partially unknown, and therefore, the global path constraints cannot be directly computed. In this work, the online alignment system has a fixed latency of just one frame, and thus, backtracking is not allowed, that is, the decision is made directly from the information at each frame t. Here, we have used the online DTW in [29], where the optimal path vector $\mathbf{W} \in \mathbb{R}_+^{T_m}$ for frame t is computed as,

$$\mathbf{W}_t(\tau) = \begin{cases} 1 & \text{if } \tau = \arg\min_\tau \mathbf{C}_t(\tau) \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

Subsequently, the coefficients of the time-varying gains matrix at frame t can be computed from the optimal path matrix $\mathbf{W}$ as,

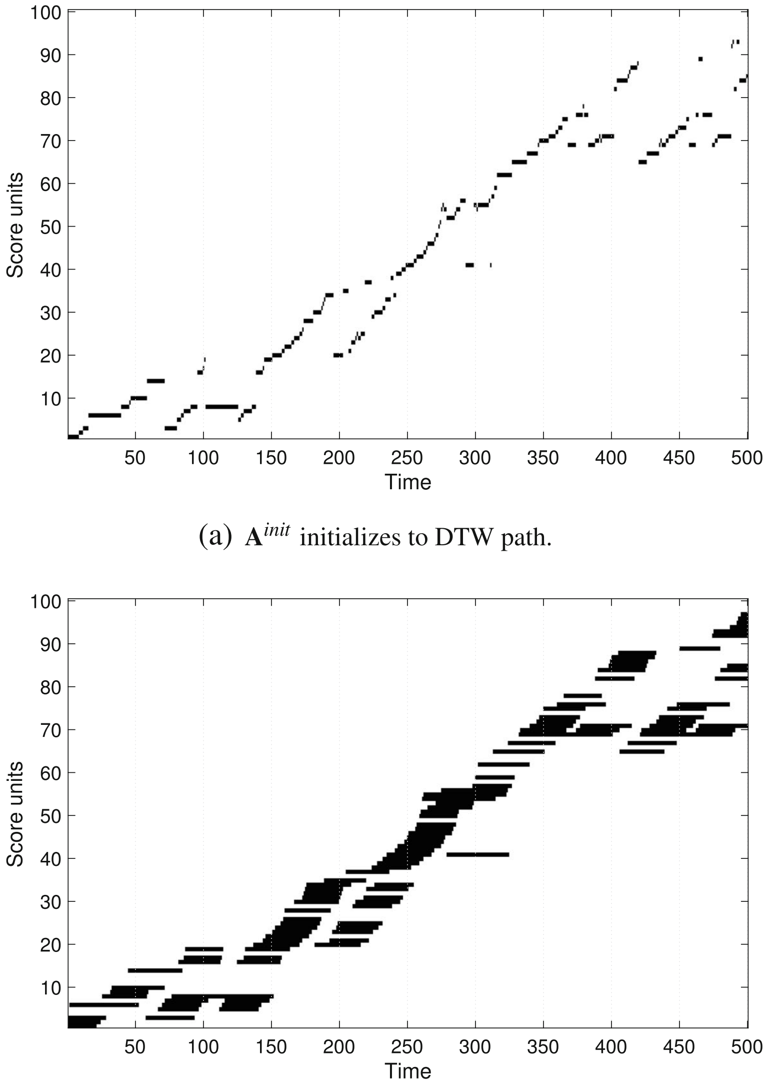
$$\mathbf{A}_{k,t}^{\text{init}} = \sum_\tau \mathbf{A}_k^B(\tau)\mathbf{W}_t(\tau) \tag{20}$$

As can be inferred, the matching precision is going to have influence in the separation stage. In fact, attending to the MIREX Score Following results presented in [29], the best alignment performance is obtained using a tolerance window of 1 s. In other words, the estimated alignment could deviate up to a maximum of 1 s from the reported note onset. To account for this misalignment, the gains matrix $\mathbf{A}^{\text{init}} \in \mathbb{N}^{K\times T_r}$ is extended using a temporal window around the frame t selected by the alignment,

$$\mathbf{A}_k(t) = \mathbf{A}_k^{\text{init}}(t) \circ \mathbf{1}_\lambda, \tag{21}$$

where $\mathbf{1}_\lambda$ is a ones window with length $\lambda = 1$ s and symbol $\circ$ stands for convolution.

Figure 4 illustrates an example of this widen procedure for the initialized gains for the second passage from A. Bruckner Symphony no. 8, II movement (detailed in Section 5.1). The estimated optimal path is displayed in subplot (Fig. 4a). In subplot (Fig. 4b), a temporal window of 1 s is applied extending each score unit in several time frames. Note that this extended gains allows more than one active unit per frame.



(a) $\mathbf{A}^{init}$ initializes to DTW path.

(b) $\mathbf{A}$ initializes to DTW extended path.

Fig. 4 A initialization for a music signal from the test database in Section 5.1 (A. Bruckner's Symphony no. 8, II movement, bars 1-61). **a** Aligned piano-roll. **b** Extended aligned piano-roll

As can be observed in Fig. 1, after this alignment process, the time-varying gains matrix $\mathbf{A} \in \mathbb{N}^{K \times T_r}$ is initialized for the following stage.

3.4 Separation stage

At this point, all the parameters of the signal model in Eq. (8) (see Section 3.1) are initialized. In particular, the basis functions $\mathbf{S}$ are obtained from a computed a priori dictionary for all the instruments/notes in the score, and the score information is encoded into the notes-to-unit matrix $\mathbf{E}$ and initialized from the synthetic signal obtained from the score in Section 3.2. Finally, the time-varying gains $\mathbf{A}$ are initialized from the alignment procedure in Section 3.3.

In this section, separation is performed in two steps. First, the time-frequency spectrogram for each sound source is estimated using a signal factorization procedure. Second, a Wiener soft-masking strategy is used to obtain the time-domain separated source signals.

3.4.1 Signal model parameter estimation

Here, we present a NMF approach to estimate the spectrogram (magnitude or power) for each source in the mixture. To this end, we applied the gradient descend algorithm to minimize the cost function between the observed and the estimated signal spectrogram $D_\beta(\mathbf{X}(f,t)|(\sum_{k,p,j} \mathbf{S}_{p,j}(f)\mathbf{E}_{p,j,k}\mathbf{A}_k(t)))$ (see Eq. (2)).

The multiplicative update rules for each parameter of the signal model in Eq. (8) are defined using Eq. (5) as follows:

$$\mathbf{E}_{p,j,k} \leftarrow \mathbf{E}_{p,j,k} \odot \left(\frac{\sum_{f,t} \mathbf{S}_{p,j}(f) [\, \hat{\mathbf{X}}(f,t)^{\beta-2} \odot \mathbf{X}(f,t)] \, \mathbf{A}_k(t)}{\sum_{f,t} \mathbf{S}_{p,j}(f) \hat{\mathbf{X}}(f,t)^{\beta-1} \mathbf{A}_k(t)} \right) \tag{22}$$

$$\mathbf{A}_k(t) \leftarrow \mathbf{A}_k(t) \odot \left(\frac{\sum_{f,p,j} \mathbf{S}_{p,j}(f) [\, \hat{\mathbf{X}}(f,t)^{\beta-2} \odot \mathbf{X}(f,t)] \, \mathbf{E}_{p,j,k}}{\sum_{f,p,j} \mathbf{S}_{p,j}(f) \hat{\mathbf{X}}(f,t)^{\beta-1} \mathbf{E}_{p,j,k}} \right) \tag{23}$$

Moreover, scaling the parameters is necessary to ensure that **E** models only the relative amplitude between instruments. The scaling procedure is presented as follows:

$$e_k = \sqrt{\sum_{p,j} \mathbf{E}^2_{p,j,k}} \quad , \quad \mathbf{E}_{p,j,k} = \frac{\mathbf{E}_{p,j,k}}{e_k} \quad , \quad \mathbf{A}_k(t) = e_k \mathbf{A}_k(t) \tag{24}$$

Note that the relative amplitude between the different instruments that compose the mixture signal usually remains constant throughout the music performance, and therefore, **E** is common for all the audio frames. In other words, the notes-to-unit matrix can be updated only for the offline scheme (i.e., when the whole signal is processed). Alternatively, when the signal is partially unknown (i.e., the online scheme), the factorization consists of updating only the parameter **A**, whereas **E** is a fixed parameter.

To allow a real-time implementation of the online scheme, we propose to use a low-rank NMF method similar to [62]. In particular, the method in [62] consist of decomposing an input image as a set of subimages represented using a reduced set of components for each subimage. In our approach, we perform the factorization at frame level, and thus, the number of active score units is sparse. Therefore, we propose to compute the signal factorization at frame t using only the subset of score units k' active after the initialization of **A** in the alignment stage as displayed in the example of the Fig. 4b.

The whole offline and online signal factorization procedures are summarized in Algorithm 1 and Algorithm 2, respectively.

Algorithm 1 Offline signal factorization scheme

```
Initialize $\mathbf{S}$ from the trained a priori dictionary and keep it fixed.
Initialize $\mathbf{E}$ from the input score (see Section 3.2).
Initialize $\mathbf{A}$ from the offline alignment procedure in Section 3.3.
Compute the signal model using Eq. (8).
while not convergence and iter $\leq$ no. of iters do
    Update $\mathbf{A}$ according to Eq. (22).
    Recompute the signal model using Eq. (8).
    Update $\mathbf{E}$ according to Eq. (21).
    Scale $\mathbf{E}$ to $l1$-norm and compensate by rescaling $\mathbf{A}$ using Eq. (23).
    Recompute the signal model using Eq. (8).
end while
```

3.4.2 *Reconstruction*

Once the model parameters have been optimized, we perform the SS using generalized Wiener filtering. In fact, SS consists of estimating the complex amplitude at each time-frequency cell for each source. Generalized Wiener masks are highly used in the SS literature [6, 56]. The Wiener filter method computes the relative energy contribution of each source with respect to the energy of the mixed signal $\mathbf{x}(t)$. The Wiener soft mask α_j for each time-frequency bin (f,t) is defined as,

$$\alpha_j(f,t) = \frac{|\hat{\mathbf{X}}_j(f,t)|^2}{\sum_j |\hat{\mathbf{X}}_j(f,t)|^2} \tag{25}$$

where α_j represents the relative energy contribution of each source and $\hat{\mathbf{X}}_j \in \mathbb{R}_+^{F \times T_r}$ is the magnitude spectrogram per instrument, computed as,

$$\hat{\mathbf{X}}_j(f,t) = \sum_{p,k} \mathbf{S}_{p,j}(f) \mathbf{E}_{p,j,k} \mathbf{A}_k(t) \tag{26}$$

The sum of all the estimated source power spectrograms $|\hat{\mathbf{X}}_j(f,t)|^2$ is the power spectrogram of the mixed signal $|\hat{\mathbf{X}}(f,t)|^2$. Then, to obtain the estimated source magnitude spectrogram $\hat{\mathbf{X}}_j$, Eq. (27) is used.

$$\hat{\mathbf{X}}_j(f,t) = \sqrt{\alpha_j(f,t)} \cdot \mathbf{X}(f,t) \tag{27}$$

Algorithm 2 Online signal factorization scheme

```
Initialize $\mathbf{S}$ from the trained a priori dictionary and keep it fixed.
Initialize $\mathbf{E}$ from the input score (see Section 3.2).
for each time frame $t$ do
    Initialize $\mathbf{A}$ from the online alignment procedure in Section 3.3.
    Select the subset of active score units $k'$ from $\mathbf{A}$.
    Compute the signal model for the input signal spectrum at frame $t$ using Eq. (8).
    while not convergence and iter $\leq$ no. of iters do
        Update $\mathbf{A}$ according to Eq. (22).
        Recompute the signal model using Eq. (8).
    end while
end for
```

Finally, using the phase spectrogram of the input mixture and computing the inverse overlap-add STFT of the estimated magnitude spectrogram $\hat{\mathbf{X}}_j$, we estimate the source $\hat{\mathbf{s}}_j(t)$.

4 Minus one application

Minus one is a music application which consists in removing a concrete instrument from an orchestra signal. This technique allows professional musicians to play an instrument accompanied by a real orchestra. In general, music is distributed in a monophonic or stereophonic audio stream where all the instruments are mixed together, thus making the material unsuitable for minus one usage. This is the case of repositories such as IMSLP which stores in these formats most of the classical music pieces. Recently, some commercial platforms which implement this technique have appeared, such as Nomadplay[7]. However, they only offer a little set of compositions and instruments to play, as accompaniments are produced in recording studio by professional musicians, which is costly and time-consuming. In this way, developing a method to generate the accompaniment for user-selected instrument and audio is required. Thus, we propose to use our SS framework for the minus one application.

The goal is to provide as output the original mixture without the selected-by-the-user instrument. This can be carried out by first performing a downmix via SS and then upmixing the rest of instruments. Unfortunately, the proposed NMF/DTW+Wiener method follows an additive model, and thus, the interference caused by the removed instrument over the minus one signal might be perceptible, especially in those time-frequency regions where the removed instrument was prominent. To mitigate this problem, we have used the psychoacoustical masking model presented in [63] that incorporates the across-frequency integration observed within the human auditory system. The perception of an audio signal is the consequence of various physiological and psychological effects. In fact, auditory system models are frequently used to calculate a spectral masking threshold [64]. In this manner, any additive signal components, whose power spectral density lies completely below the masking threshold, will not be perceived by a human listener.

In this work, we propose to estimate a masking model for the resulting minus one upmixing. Then, the perceptual interference can be minimized by removing those time-frequency bins (f, t) where the amplitude of the undesired instrument is above the time-frequency masking threshold γ_m. This masking threshold is computed for each Minus One mixture using the implementation described in [63].

Denoting by $\hat{\mathbf{X}}_i \in \mathbb{R}_+^{F\times T_r}$ the magnitude spectrogram of the selected instrument signal, $\hat{\mathbf{X}}_m \in \mathbb{R}_+^{F\times T_r}$ the magnitude spectrogram of the minus one mixture, defined as

$$\hat{\mathbf{X}}_m(f,t) = \sum_{\substack{0<j<J \\ j\neq i}} \hat{\mathbf{X}}_j(f,t), \tag{28}$$

and γ_m the time-frequency masking threshold for the minus one mixture, the spectrogram of the minus one signal can be refined by:

$$\hat{\mathbf{Y}}_m(f,t) = \begin{cases} 0 & \text{if } \hat{\mathbf{X}}_i(f,t) \geq \gamma_m(f,t) \\ \hat{\mathbf{X}}_m(f,t) & \text{if } \hat{\mathbf{X}}_i(f,t) < \gamma_m(f,t) \end{cases} \tag{29}$$

Note that those bins (f, t) where $\hat{\mathbf{X}}_i$ is above the masking threshold γ_m (i.e., perceptible in the minus one mixture) are canceled in the estimated magnitude spectrogram $\hat{\mathbf{Y}}_m \in \mathbb{R}_+^{F\times T_r}$. In this way, we can minimize the interfere caused by the removed instrument.

Finally, as in Section 3.4.2, the time-domain signal for the minus one is obtained using the phase spectrogram of the input mixture and computing the inverse overlap-add STFT of the estimated magnitude spectrogram $\hat{\mathbf{Y}}_m$ in Eq. (29).

5 Experimental results and discussion

In this section, the proposed method in Section 3 is evaluated for the task of single-channel instrumental music SS using a well-known dataset of small ensembles and a more complicated large ensembles orchestra dataset. Besides, the performance of our method has been compared to other state-of-the-art algorithms to demonstrate the reliability of our proposal.

5.1 Datasets

In this work, we assess the performance of our proposed method considering two different databases. Firstly, we have used the University of Rochester Multimodal Music Performance (URMP) dataset developed by Li et al. [65]. This dataset is compounded by 44 classical chamber music pieces ranging from duets to quintets (11 duets, 12 trios, 14 quartets, and 7 quintets) and played by 14 different common instruments in orchestra. The musical score, the audio recordings of the individual tracks, the audio recordings of the assembled mixture, and the ground-truth annotation files are available for each piece. Although URMP is not an orchestra dataset, it has been used to evaluate the behavior of our proposal in ensembles with a reduced number of instruments.

Secondly, we have used the orchestra database developed by Pätynen et al. [66] and processed by Miron et al. [13]. It consists of four excerpts of approximately 3 min each one composed of symphonic music from Classical and Romantic style. The four pieces vary in terms of number of instruments sections, style, dynamics, and size of

[7] https://nomadplay-app.com/

the orchestra. The first piece is from *L. van Beethoven*'s (1770–1827) Symphony no. 7, I movement, bars 1–53, corresponding to the late Classical period. Its main features are big chords and string crescendo, what make reverberation tail of a concert hall clearly audible. The second passage is from *A. Bruckner* (1824–1896) Symphony no. 8, II movement, bars 1–61, and represents the late Romantic period characterized by large dynamics and the size of the orchestra. The third passage is from *G. Mahler*'s (1860–1911) Symphony no. 1, IV movement, bars 1–85. It represents also the Romantic period and is another example of work for large orchestras. The last piece is a soprano aria of *Donna Elvira* from the opera *Don Giovanni* by *W. A. Mozart* (1756–1791). This performance represents the Classical period and presents small orchestra characteristics, including a soloist segment. More details of the database are provided in Table 2, including measurements of the complexity of the compositions calculated from the score, such as the polyphony density (average number of simultaneous notes per frame) and the inter-onset duration (average time gap between onsets).

5.2 Experimental setup

Many signal processing applications adopt frequency logarithmic discretization. Although it is not the only way to reduce the memory requirements, using logarithmic resolution in frequency is a common approach to minimize the dimensionality and the memory footprint in matrix operations. For example, uniformly spaced subbands on the equivalent rectangular bandwidth scale are assumed in [57].

In this work, a resolution of 1/4 of a semitone in frequency is used as in [37, 67]. The time-frequency representation is obtained using 8192-point STFT and integrating the frequency bins corresponding to the same 1/4 semitone interval. The frame size and the hop size for the STFT are set to 5644 (128 ms) and 1412 (32 ms) samples, respectively, and the sampling rate is equal to 44.1 kHz.

Regarding the signal factorization scheme, we have studied the performance of the proposed method as a function of parameter β from the β-divergence cost function defined in Eq. (3). To this end, we have evaluated our algorithm over the dataset in Section 5.1 varying the value of β in the range $[0, 2]$, finding the optimal value around $\beta = 1.3$. This value is in line with other works in the state-of-the-art [13, 68, 69]. Besides, for our constrained by the score signal model, we have observed that the reconstruction error converges after 50 iterations. Therefore, we have chosen this value as the maximum number of iterations for the decomposition procedure.

5.3 Evaluation metrics

In this paper, we propose a SS system that perform both score alignment and source separation. First, to evaluate the alignment performance of the proposed method, we have used the same evaluation metrics as in the MIREX Score Following task. In this way, for each piece, an aligned rate (AR) or precision is defined as the proportion of correctly aligned notes in the score and ranges from 0 to 1. If a note onset does not deviate more than a threshold (a.k.a tolerance window) from the reference alignment, this note is considered to be correctly aligned.

Then, to evaluate the separation performance, we have used the BSS_EVAL [70] and the PEASS [71] toolboxes to evaluate the performance of our method for the task of SS. These metrics are commonly accepted in the field of SS and thus allow a fair comparison with other state-of-the-art methods. In particular, each separated signal is assumed to produce a distortion model that can be expressed as follows:

$$\hat{\mathbf{s}}_j(t) - \mathbf{s}_j(t) = \mathbf{e}_j^{\text{target}}(t) + \mathbf{e}_j^{\text{interf}}(t) + \mathbf{e}_j^{\text{artif}}(t) \qquad (30)$$

where $\hat{\mathbf{s}}_j$ is the estimated source signal for instrument j, $\mathbf{s}_j$ is the original signal of the instrument j, $\mathbf{e}^{\text{target}}$ is the error term associated with the target distortion component, $\mathbf{e}^{\text{interf}}$ is the error term due to interference of the other sources, and $\mathbf{e}^{\text{artif}}$ is the error term attributed to the numerical artifacts of the separation algorithm. The metrics provided by BSS_EVAL for each separated signal are the *source to distortion ratio* (SDR), the *source to interference ratio* (SIR), the *source to artifacts ratio* (SAR), and the *source image spatial distortion ratio* (ISR) [70]. Besides, to predict the subjective quality of estimated source signals, the PEASS toolbox [71] provides four scores: the *overall perceptual score* (OPS), the *target-related perceptual score* (TPS), the *interference-related perceptual score* (IPS), and the *artifacts-related perceptual score* (APS). These scores are obtained by making use of auditory-motivated metrics provided by the PEMO-Q auditory model [72] to assess the perceptual salience of the target distortion (qTarget),

Table 2 Characteristics of the orchestral dataset used for the evaluation of our SS system

Composer	Piece name	Dur.	Tracks	Notes	Poly. dens.	Inter-onset dur.
Beethoven	Symphony no. 7	3 min 11 s	20	3075	8.7 (4.9)	0.21 s (0.19 s)
Bruckner	Symphony no. 8	1 min 27 s	39	2789	10.6 (4.5)	0.21 s (0.08 s)
Mahler	Symphony no. 1	2 min 12 s	30	2822	5.9 (3.5)	0.24 s (0.23 s)
Mozart	Don Giovanni	3 min 47 s	10	2724	5.0 (2.4)	0.26 s (0.17 s)

For the polyphony density and the inter-onset duration, both the mean and the standard deviation (in parentheses) are included interference (qInterf), and artifacts (qArtif), computing also a global metric (qGlobal). Finally, a nonlinear mapping using neuronal networks trained with a set of different audio signals is performed in order to get the set of objective measures.

5.4 Algorithms for comparison

In order to show the benefits of proposal, we have compared the separation performance of our method with other state-of-the-art algorithms. Besides, we present two "unrealistic" baseline methods to state the extreme separation performances, here denoted as ideal separation and energy distribution. The different approaches compared here are the following:

5.4.1 Ideal separation

This method computes the optimal value of the Wiener mask at each frequency and time component assuming that the signals to be separated are known in advance. Therefore, this approach represents the upper bound for the best separation that can be reached with the used time-frequency representation.

5.4.2 Energy distribution (ED)

This procedure uses the mixture signal divided by the number of sources as input for the evaluation. This evaluation provides a starting point for the separation algorithms.

5.4.3 Miron method

We have included in the evaluation the results of the method proposed by Miron et al. in [13]. Although Miron's method proposes an offline system for score-informed audio SS for multichannel orchestral recordings, we have evaluated its method for only one channel. For single channel SS, this method uses the signal model introduced in Eq. (9). Similar to our approach, this method used a 1/4 semitone resolution and the basis functions are learned in advance from training material (using real audio samples from the RWC database). However, in Miron's method, the time-varying gains are initialized using pre-aligned score information and refined using image processing techniques.

5.4.4 Ganseman method

This method was introduced by Ganseman et al. in [9]. Ganseman proposes a method based on PLCA for SS according to source models. Here, the gains are initialized from a previously aligned MIDI score and the basis functions learned from synthetic material.

5.4.5 Fritsch method

We have also incorporated results for the method proposed by Fritsch in [73]. It consists of using the IS NMF of the power spectrogram with multiplicative updates for SS. As in the Ganseman method, they have a previous stage where the score is aligned and synthesized for the learning of the basis functions dictionary.

5.4.6 Soundprism system

We evaluate the results of the online method proposed by Duan et al. in [22]. The authors proposed a system that addresses score-informed music SS that is suitable for real-time implementation. It consists of two parts: (1) a multipitch + HMM-based score follower that align each time frame of the audio performance and (2) a source separator which reconstructs the source signals informed by the score.

5.4.7 REMASSeparation system

We have also included the results of our preliminary online SS work [30]. In this work, we proposed to use the real-time alignment implementation based on [31] and presented in [32] together with the source separation procedure described in [13]. Note that the signal model used was the same as in [13] but restricted to single-channel signals. Unlike Miron's method, here, the time-varying gains are initialized using the online alignment described in [31].

5.4.8 Deep learning approaches

In order to compare the performance of the signal decomposition-based methods with the novel DL approaches, we have reviewed three state-of-the-art DL approaches for classical music SS. In particular, we have used two extensions of the well-known Wave-U-Net DL model presented in [16]: (1) Exp-Wave-U-Net is a slight modification of the original model presented in [47] that allows separation of a dynamic number of sources (original model in [47] was designed only for 2 or 4 sources). (2) CExp-Wave-U-Net [16], in this variant, the authors proposed to use instrument labels to inform the network about the instruments presented in the mixture. Note that both, Exp-Wave-U-Net and CExp-Wave-U-Net methods, have been evaluated over the URMP dataset in [16] and the source code is also available for reproducible research. Finally, the third method in [15] is a score-informed convolutional neural network (CNN) based method which estimates the optimal soft-mask for each instrument and refines them using the pre-aligned score. Here, we have used the author implementation, and we have extended the training data (from RWC instrument sound database) to account for all the instruments from the URMP dataset. As in [15], the generated training data is obtained using sample-based synthesis with samples from the RWC instrument sound database. The method synthesizes original scores at different tempos and dynamics, considering local timing deviations and using different timbres to generate a wide variety of renditions of given pieces.

5.4.9 Variants of the proposed model

We are also going to present results of several variants of our own model. In that sense, a set of configurations have been considered when comparing the models in order to know the influence of adapting different parameters. As mentioned in Section 3.1, our signal model is decomposed in three parameters: **S**, **E**, and **A**.

In our tests, the matrix **S** is trained in the previous stage and is fixed for the separation process. The gains matrix **A** is always adapted to the test signal in the separation process, but we have experimented with different initializations: (a) a random matrix (non-informed), (b) the output path of the offline DTW, and (c) the output path of the online DTW.

In the case of the notes-to-unit matrix **E**, we have the following scenarios:

- scoreFree: **E** is initialized as a binary matrix that relates each set of active notes with their corresponding score unit inferred from the MIDI score. Therefore, no assumption on the relative amplitude between notes/instruments is made in the initialization of this parameter. Instead, these relative amplitudes are learned during the update of parameter **E** in the separation stage.
- scoreSynthFree: Opposite to the ScoreFree setup, in this configuration, **E** is initialized with the relative amplitudes between notes/instruments belonging to the same score unit. This information is learned in advance from the MIDI score using a synthesizer that accounts for perceptual parameters annotated in the MIDI file such as the velocity (i.e., how fast/hard each note is pressed/released), channel volume, pan, modulation, or effects. Moreover, parameter **E** is also updated during the factorization to adapt the information to the actual performance.
- scoreSynthFix: In this configuration, **E** is initialized from the synthesized MIDI score as in the scoreSynthFree setup. However, different than in scoreSynthFree, this parameter is kept fixed during the separation stage.

Notice that updating parameter **E** during the factorization in the separation stage allows to adapt the velocity of the instruments in the mixture to the actual performance. Consequently, the scoreSynthFree variant allows to mitigate velocity deviations from the MIDI score, whereas in the case of the scoreSynthFix, the performance will underperform for large deviations.

Additionally, we have evaluated an oracle variant denoted as ground-truth (GT) annotation which uses the manually annotated by musicologist score as a hard prior initialization for the times varying gains **A**. This variant give us a measure to evaluate the reliability of the score alignment procedure in Section 3.3 over the evaluated dataset.

All of these variants are summarized in Table 3.

In the spirit of reproducible research, the code of this experimental study is available online[8].

5.5 Evaluation of small ensembles single-channel signals

This section first presents the results of the evaluation of the score alignment where a comparison with other reference methods is carried out for the URMP dataset. Then, the SS results for the signal decomposition-based models presented in Section 5.4 are shown. Finally, a comparison of our proposal with novel DL approaches is performed for this small ensembles dataset.

5.5.1 Score alignment results

To analyze the performance of our alignment methods (offline and online), a comparison has been carried out with four reference methods: (a) Carabias's offline [31], (b) Carabias's online [31], (c) Ellis' offline [74], and (d) Soundprism [22]. Figure 5 shows the alignment results in terms of precision values of the analyzed methods as a function of the onset deviation threshold. The threshold value varies from 50 to 2000 ms. As can be observed, all the offline approaches reach similar results. Note that the number of instruments in this ensembles are limited (2 to 5 instruments), so the uncertainty here is lower than in a full orchestra scenario. Regarding the online methods, our online approach obtains a more precise alignment than the Carabias' online and the Soundprism approaches. The best performance of our model with respect to Carabias' proposal is because the spectral patterns for each score unit are computed using the parameter **S** (see Eq. 14), which is learned in advance using isolated notes of real solo instrument recordings, while Carabias uses the synthesized MIDI signal to learn these spectral patterns.

According to Fig. 5, we can see that our methods require a tolerance window of at least 1 s to converge to the optimum alignment. Using a lower threshold could be beneficial to minimize the interference between instruments. However, it can also provoke the lost of the onsets/offsets causing a poor separation performance. As a compromise between reducing the intereference between instruments and avoiding the lost the onset/offset information, in this work, we chose a temporal window of 1 s to extend the gains matrix $\mathbf{A}^{\text{init}}$ around each frame t (as described in Section 3.3).

[8] https://github.com/AntonioJMM/OISS_Minus-One.github.io

Table 3 Algorithm notation

Abbr.	Description	Trained parameters	Separation stage	
			Free parameters	*Fixed parameters*
scoreFree	**E** is initialized from the MIDI score and set as a free parameter in the separation stage. **A** is initialized to random values.	**S**	**E**, **A**	**S**
scoreFree-DTWoff	**E** is initialized from the MIDI score and set as a free parameter in the separation stage. **A** is initialized to the output path of the offline DTW.	**S**	**E**, **A**	**S**
scoreSynthFree	**E** is initialized from the synthesized MIDI score and set as a free parameter in the separation stage. **A** is initialized to random values.	**S**, **E**	**E**, **A**	**S**
scoreSynthFree-DTWoff	**E** is initialized from the synthesized MIDI score and set as a free parameter in the separation stage. **A** is initialized to the output path of the offline DTW.	**S**, **E**	**E**, **A**	**S**
scoreSynthFix	**E** is initialized from the synthesized MIDI score and kept fixed in the separation stage. **A** is initialized to random values.	**S**, **E**	**A**	**S**, **E**
scoreSynthFix-DTWoff	**E** is initialized from the synthesized MIDI score and kept fixed in the separation stage. **A** is initialized to the output path of the offline DTW.	**S**, **E**	**A**	**S**, **E**
scoreSynthFix-DTWon	**E** is initialized from the synthesized MIDI score and kept fixed in the separation stage. **A** is initialized to the output path of the online DTW.	**S**, **E**	**A**	**S**, **E**
GT	**E** is initialized from the synthesized MIDI score and set as a free parameter in the separation stage. **A** is initialized using a hard prior from the ground-truth annotated score.	**S**, **E**	**E**, **A**	**S**

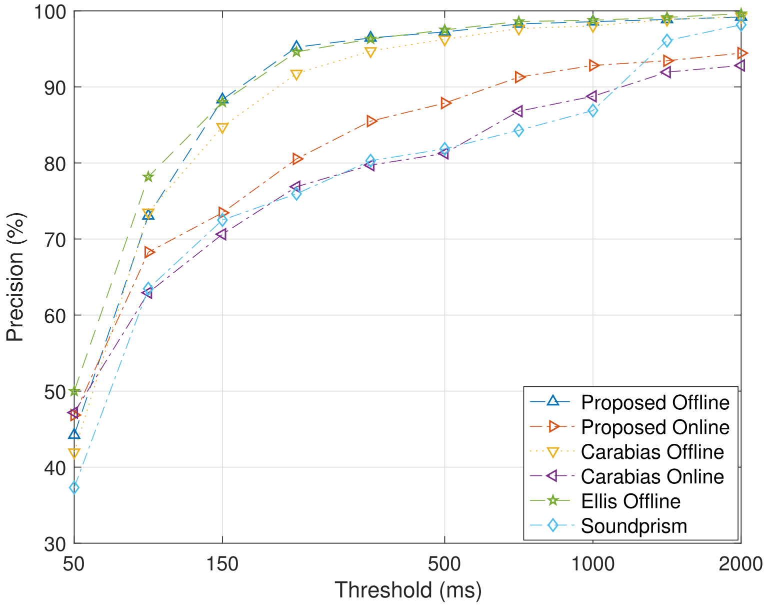


Fig. 5 Alignment evaluation over the URMP dataset in terms of precision values in function of the tolerance window

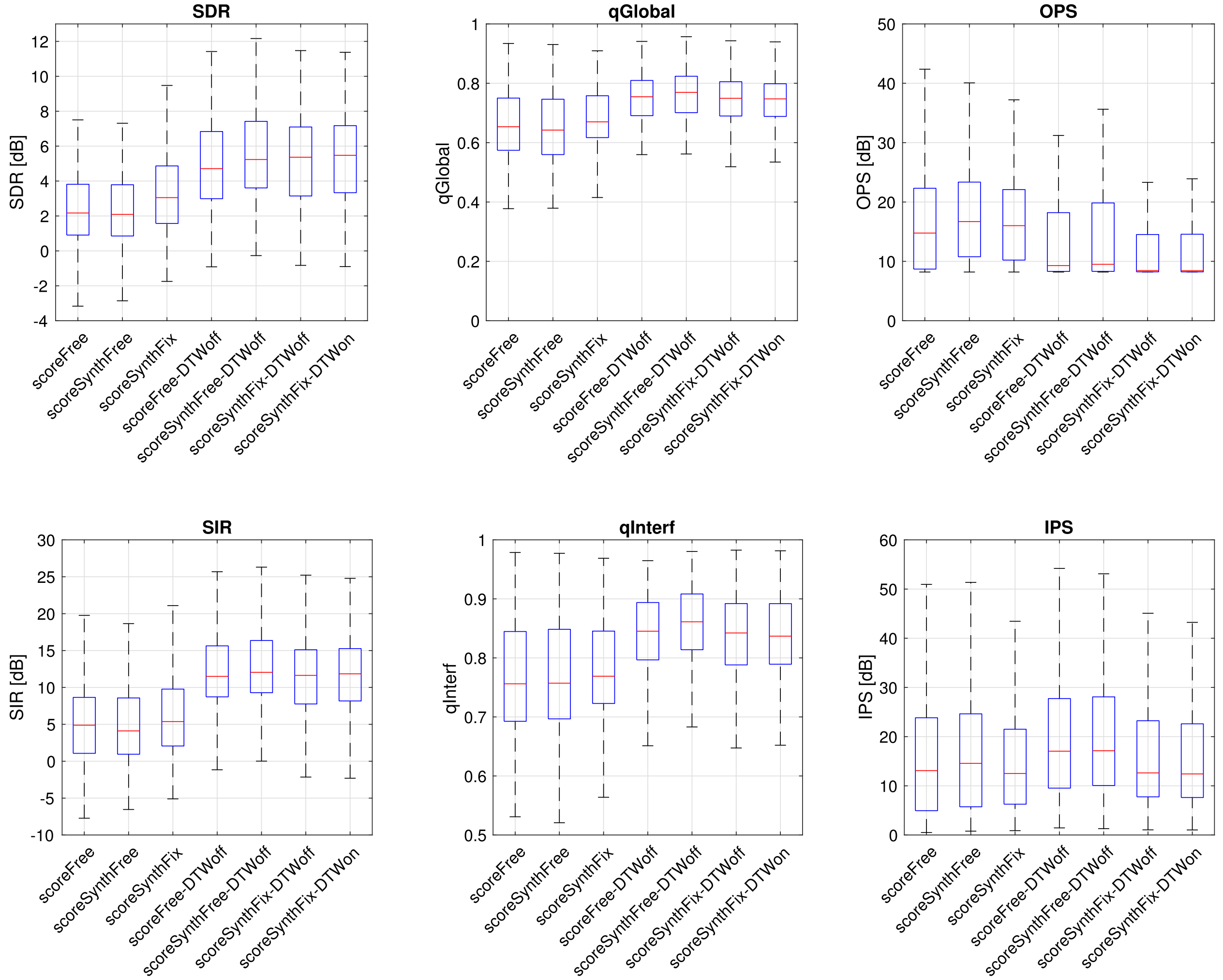


Fig. 6 Objective and perceptual results of the variants of the proposed model using the BBS_EVAL metrics and the PEASS metrics for the URMP dataset

5.5.2 Source separation results

In this section, we have studied the separation performance of the proposed method as a function of parameters **E** and **A**. The obtained results for each configuration described in Section 5.4.9 are presented in Fig. 6. The lower and upper limits of each box (in blue) represents the 25th and 75th percentiles of the sample. The red line in the middle of each box is the median. The lines extending above and below each box represent both the best and the worst performance, respectively, for each variant. For the sake of brevity and better understanding, only the main metrics used in the comparison are displayed, while the other metrics introduced in Section 5.3 are given in Appendix 2.

As can be observed, initializing the time-varying gains with information from the DTW-based alignment provide better separation results than random initialization of **A** in terms of SDR and SIR. However, imposing sparsity constraints in the gains provokes a slight underperformance in terms of SAR (see Appendix 2) which could be associated to loss of time-frequency coefficients (i.e., set to zero) in the reconstructed signal. A similar behavior can be observed by analyzing the results obtained with the perceptual similarity measures (PSM) provided by the PEMO-Q auditory model (qGlobal, qInterf, qArtif, and qTarget). However, it must be highlighted that attending to PEASS metrics (OPS, TPS, IPS, and APS), there is no correlation between these metrics and PSM ones. A possible reason could be that the nonlinear mapping provided by the PEASS neural network is not optimized to evaluate musical ensembles (see test material in [71]). We encourage the readers to listen to the audio demos[9].

[9] https://antoniojmm.github.io/OISS_Minus-One.github.io/

Regarding the performance as a function of parameter **E**, better results are obtained when the notes-to-units matrix **E** is initialized using the synthetized MIDI score. In fact, scoreSynthFree-DTWoff obtains the best performance in terms of objective and perceptual measures. In other words, under a proper initialization, adapting the parameters to the input signal outperforms the SS performance. scoreSynthFix-DTWoff and scoreSynthFix-DTWon provide similar separation results and rank second among the compared variants. Note that **E** is a global parameter (i.e., requires the full signal to be updated), and therefore, online DTW can only be implemented assuming **E** as a fixed parameter.

In the case of random gains initialization, scoreSynthFix provide better results than scoreSynthFree and scoreFree configurations in terms of SDR, SIR, and PSM. In fact, the model estimation is more biased towards local minima when the number of free and unconstrained parameter increases. Therefore, reliable adaptation of the parameters can only be made using a proper initialization.

Figure 7 illustrates a comparison of the best performing offline and online configurations of the proposed method (scoreSynthFree-DTWoff and scoreSynthFix-DTWon) and the state-of-the-art methods presented in Section 5.4.

As can be seen, the energy distribution (ED) baseline method obtains the worst results in terms of SDR and SIR which seems logical since no actual separation is performed. However, it obtains the best results in terms of SAR (see Fig. 13 in Appendix 2). As commented in [4], it is usual that constrained signal decomposition-based models underperform unconstrained models in terms of SAR. In fact, enforcing the system to model only the score information may cause an unreliable modeling of the recording conditions or the background noise, if present. On the contrary, using suitable constraints provides better

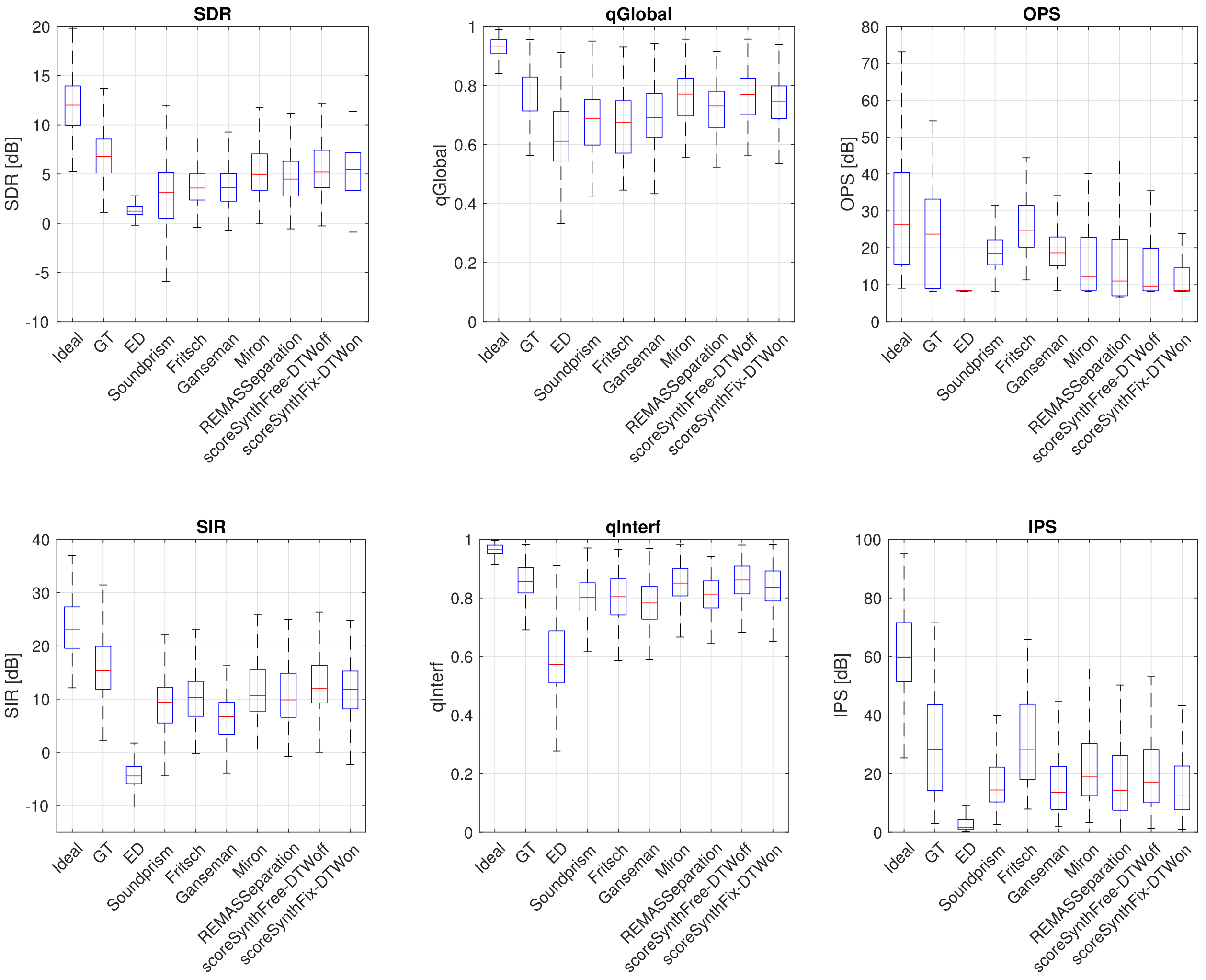


Fig. 7 Objective and perceptual results for the comparison of the SS methods using the BBS_EVAL metrics and the PEASS metrics for the URMP dataset

results in terms of SDR, SIR, and perceptual measures which are more in line with the actual listening quality of the separation. Regarding the obtained results in terms of PEASS metrics, we can see that they are not in line with the PSM. In fact, as we commented before, the testing material used to train this metrics does not include classical ensembles music signals, and therefore, the obtained results for these metrics might not represent the actual listening quality of the separation (listen to the demos website[9]). As can be seen in Fig. 7, in general, the distribution of the compared methods using the energy-based (BSS_EVAL) and the perceptual (PSM) measures is similar. Therefore, for the sake of brevity, we will use the SDR and the SIR as the main metrics in order to compare the performance of the evaluated methods for the rest of the paper.

The best results are obtained with the baseline ideal separation method (SDR = 12 dB, SIR = 23.01 dB). This measure informs us about the best separation that can be achieved using the selected 1/4 semitone time-frequency representation. Besides, our oracle alignment (GT) variant achieves the best separation results (SDR = 6.80 dB, SIR = 15.35 dB) among the compared methods. This oracle approach provides information about the best separation results that can be obtained using our method when perfect annotation is available. In fact, the proposed variants, scoreSynthFree-DTWoff (SDR = 5.23 dB, SIR = 12.05 dB) and scoreSynthFix-DTWon (SDR = 5.47 dB, SIR = 11.84 dB), provide competitive separation results in terms of SDR and SIR in comparison with the oracle GT solution. This performance can be observed also in the perceptual metrics. Note that GT uses a hard-prior initialization of the gains, whereas the variants using score alignment use a soft-prior initialization (i.e., a tolerance window) to account for the possible missalignment at the onset and offset frames.

Regarding the state-of-the-art algorithms, the signal factorization-based methods Fritsch (SDR = 3.58 dB, SIR = 10.30 dB) and Ganseman (SDR = 3.64 dB, SIR = 6.68 dB) provide similar separation results. In the case of Miron (SDR = 4.96 dB, SIR = 10.69 dB), the obtained results are slightly below the proposed variants in terms of SDR and SIR, while they are slightly above our online variant in terms of qGlobal and qInterf and similar to our offline variant in terms of qGlobal and sligthly worse in terms of qInterf.

Concerning the online approaches, the Soundprism [22], REMASSeparation, and scoreSynthFix-DTWon methods are the only ones which can be implemented into a real-time system. As can be seen, our proposal outperforms both Soundprism (SDR = 3.15 dB, SIR = 9.43) and REMASSeparation (SDR = 4.49 dB, SIR = 9.84).

Note that all the compared offline score-informed methods in Fig. 7 use the same alignment DTW-based strategy presented in Section 3.3. Therefore, they have the same time-varying gains initialization, including the tolerance window, to allow a fair comparison between their separation performance. Attending to the compared online methods, Soundprism uses its own alignment-separation scheme, whereas the time-varying gains matrix in REMASSeparation is initialized to the online alignment method described in [31], including the tolerance window.

A reason of the superior results obtained by our method is due to the fact that in our method, each basis represents a unique combination of notes (score unit), whereas in the compared state-of-the-art, signal decomposition-based algorithms propose an iterative separation method where each component from the spectral basis represents a single note. Consequently, during the factorization within the tolerance window, these algorithms allow concurrent activation of notes that might not happen in the score nor in the performance, worsening in situations of successive short duration notes. In contrast, our method might only be active in the combinations of notes (score units) that are defined in the score.

5.5.3 Comparison with novel deep learning approaches

In this section, we compare the performance of our method with the DL approaches described in Section 5.4. Results with respect to the number of sources in the mixture are presented in Table 4 which are classified in three modules by continuous lines .

First, we analyze the results provided by Exp-Wave-U-Net method and our scoreSynthFix variant. Both methods know in advance the specific instruments which compound the mixture, but are not informed by the instrument activity. scoreSynthFix obtains better results than Exp-Wave-U-Net in terms of SDR and SIR for all the polyphony cases, while it is inferior in SAR. Regarding the methods informed by the alignment score, scoreSynthFree-DTWoff clearly outperforms the DL-based methods in terms of SDR, SIR, and SAR. CExp-Wave-U-Net obtains poor results despite knowing which instruments are presented in the mixture (i.e., conditioned labels). Note that the baseline Wave-U-Net method was designed for source separation of leading voice and accompaniment mixtures and obtained remarkable results in the SISEC campaign [47]. However, extending the original method for source separation of music ensembles without exploiting the instrument activity information does not provide reliable results as depicted in [16]. CNN method achieves good results for low polyphony, reaching 5 dB in SDR for two simultaneous sources. However, it underperforms our proposal around 3 dB in SDR and 10 dB in SIR for all the polyphony cases. Note that in [15], the evaluation was performed over a dataset compounded by four monophonic instruments (Bach chorales), whereas the evaluated dataset in this

Table 4 Comparison between methods based on deep learning techniques and our proposed variants in terms of SDR, SIR, and SAR with respect to the number of sources in the mixture

Method	nSources	SDR [dB]	SIR [dB]	SAR [dB]
Exp-Wave-U-Net	2	− 0.42	1.75	*10.98*
	3	− 3.85	− 2.74	*11.97*
	4	− 5.90	− 5.33	*12.87*
scoreSynthFix	2	*4.65*	*8.82*	9.22
	3	*4.23*	*8.30*	7.86
	4	*2.23*	*4.57*	5.54
CExp-Wave-U-Net	2	− 0.16	4.62	7.48
	3	− 0.68	2.88	5.91
	4	− 2.56	0.44	6.35
CNN	2	5.00	7.66	8.74
	3	3.03	5.47	6.03
	4	1.83	1.37	5.52
scoreSynthFree-DTWoff	2	*8.21*	*16.86*	*10.15*
	3	*6.46*	*14.16*	*8.03*
	4	*4.65*	*11.79*	*5.77*
scoreSynthFix-DTWon	2	*7.60*	*16.06*	*9.69*
	3	*6.20*	*13.39*	*7.84*
	4	*4.54*	*11.01*	*5.70*

*The best results for each module appears in italics.

paper is more complex and compounded by a combination of 14 different polyphonic instruments.

Finally, we present the results for our online variant. As can be observed, scoreSythFix-DTWon obtains similar results to the offline variants, losing less than 1 dB in terms of SDR, SIR, and SAR for all the cases.

5.6 Evaluation of large ensembles

In this section, the evaluation performed for a highly polyphonic large ensembles orchestra dataset is presented using the metrics defined in Section 5.3. Then, we show the results obtained for the acoustic minus one application.

5.6.1 Orchestra source separation results

Orchestra SS results are illustrated in Fig. 8. As in the URMP dataset, the best results are obtained with the baseline ideal separation method (SDR = 7 dB, SIR = 14.08 dB) and the GT variant achieves the best separation results (SDR = 2.05 dB, SIR = 7.33 dB) among the compared methods. scoreSynthFree-DTWoff (SDR = 1.81 dB, SIR = 6.43 dB) and scoreSynthFix-DTWon (SDR = 1.78 dB, SIR = 4.38 dB) provide similar separation results in comparison with the oracle GT solution and clearly outperform the state-of-the-art algorithms in terms of SDR and SIR. Fritsch (SDR = 0.53 dB, SIR = 1.90 dB), Ganseman (SDR = 0.91 dB, SIR = 0.78 dB), and Miron (SDR = 0.32 dB, SIR = − 1.17 dB) provide similar separation results. Attending to the perceptual metrics, the superiority of our proposed methods becomes more evident in terms of both qGlobal and qInterf.

Regarding the online approaches, REMASSeparation obtains 0.01 dB and 2.63 dB in terms of SDR and SIR, underperforming our proposal. On the other hand, Soundprism provides the worst results among the compared methods (SDR = − 9.11 dB, SIR = − 8.31 dB). A possible reason of this behavior is because Soundprism uses a multi-pitch estimation model without timbral information together with a HMM to estimate the score position for each frame of real audio. This model obtains reliable results for monotimbral (i.e., just one type of instrument) signals and low polyphony multitimbral mixtures. However, in the case of high polyphony multimbral dataset (see Table 2), the alignment is severely degraded, and consequently, Soundprism provides unreliable separation results. This result is in line with the MIREX 2010[10] task of real-time audio to score alignment where Soundprism obtains a 49.11% of total precision, whereas the alignment scheme used in our proposed ISS method provides a total precision of 95.53% over the same dataset, but in a different MIREX campaign (MIREX 2015[11]).

Finally, we have studied the performance of the compared approaches as a function of the instrument. Since the separation quality varies among the different instruments which compound the mixture, we have considered using the delta-metrics ΔSDR and ΔSIR as in [75, 76]. These metrics inform about the difference between the performance of the ideal Wiener separation (i.e., using true source spectrograms) and the separation score. The obtained results are depicted in Fig. 9. Note that in Fig. 9, the best performance corresponds to the value 0 dB (i.e., the ideal Wiener separation). As can be observed, on average, our proposed variants are over the state-of-art level in terms of ΔSDR and obtain significantly superior results in terms of ΔSIR, reaching in both cases values very close to the GT variant. Similar averaged results are obtained for Fritsch (ΔSDR = − 7.54 dB, ΔSIR = − 18.27 dB), Ganseman (ΔSDR = − 7.18 dB, ΔSIR = − 19.69 dB), and Miron (ΔSDR = − 6.74 dB, ΔSIR = − 17.51 dB).

Comparing our variants, better results are obtained in the offline case. In fact, adapting the relative amplitude between instruments in the notes-to-units matrix **E** improves the separation results. Note that **E** is a global parameter, and thus, it requires the whole audio signal to be adjusted offline.

[10] https://www.music-ir.org/mirex/wiki/2010:Real-time_Audio_to_Score_Alignment_(a.k.a._Score_Following)_Results

[11] https://www.music-ir.org/mirex/wiki/2015:Real-time_Audio_to_Score_Alignment_(a.k.a._Score_Following)_Results

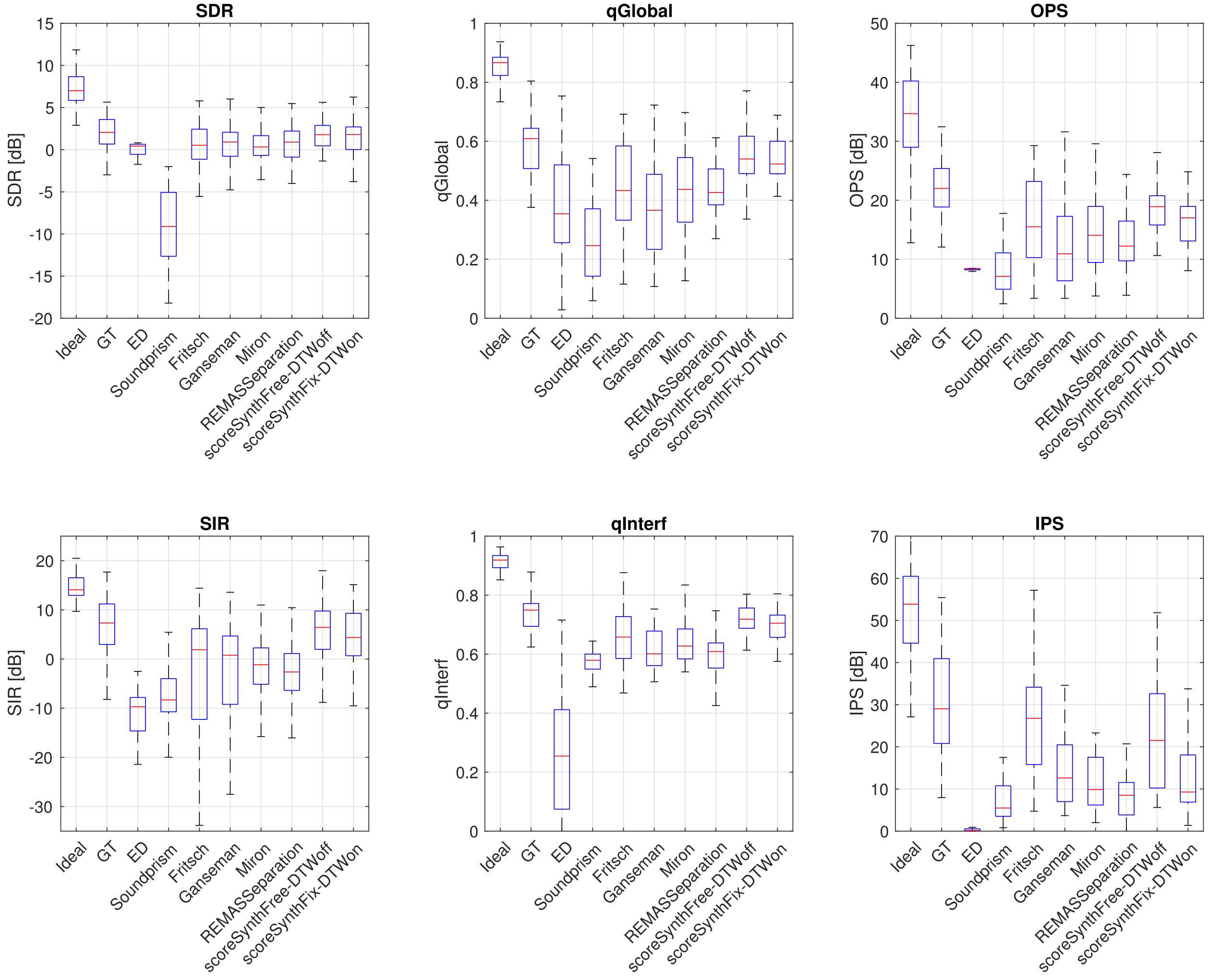


Fig. 8 Objective and perceptual results for the comparison of the SS methods using the BBS_EVAL metrics and the PEASS metrics for the orchestra dataset

5.6.2 Results for minus one application

In this section, we present the results obtained for the acoustic minus one application. The goal is to provide an audio signal where one of the sources is removed from the mixture. In this sense, minimizing the interference produced by the removed instrument, which implies maximizing SIR and qInterf of the target signal, is the main task to obtain the best minus one performance.

In order to study the performance of the minus one application, we have evaluated all the all-except-one instrument combinations using the orchestra dataset presented in Section 5.1. The overall results are presented in Fig. 10. Here, we compared the best performing separation methods for offline and online approaches applying the psychoacoustic model presented in Section 4 (scoreSynthFree-DTWoff-psycho and scoreSynthFix-DTWon-psycho) with the baseline methods (ideal separation and energy distribution) to set the "unrealistic" extreme results. In order to evaluate the reliability of the psychoacoustical masking procedure, we have included the scoreSynthFree-DTWoff and scoreSynthFix-DTWon variants of our proposed method, which are computed, in this minus one context, by combining all-except-one instruments obtained in the separation process.

As can be observed, the baseline ideal method (SDR = 16.5 dB, SIR = 22.1 dB) obtains the best results. This measure indicates the best minus one composition that can be achieved using the frequency resolution presented in Section 5.2. Regarding the effect of the psychoacoustic mask on the proposed variants scoreSynthFree-DTWoff and scoreSynthFix-DTWon, we can observe that the masking model underperforms in terms of SDR ($\sim$ 2 dB and $\sim$ 3 dB in the offline and online

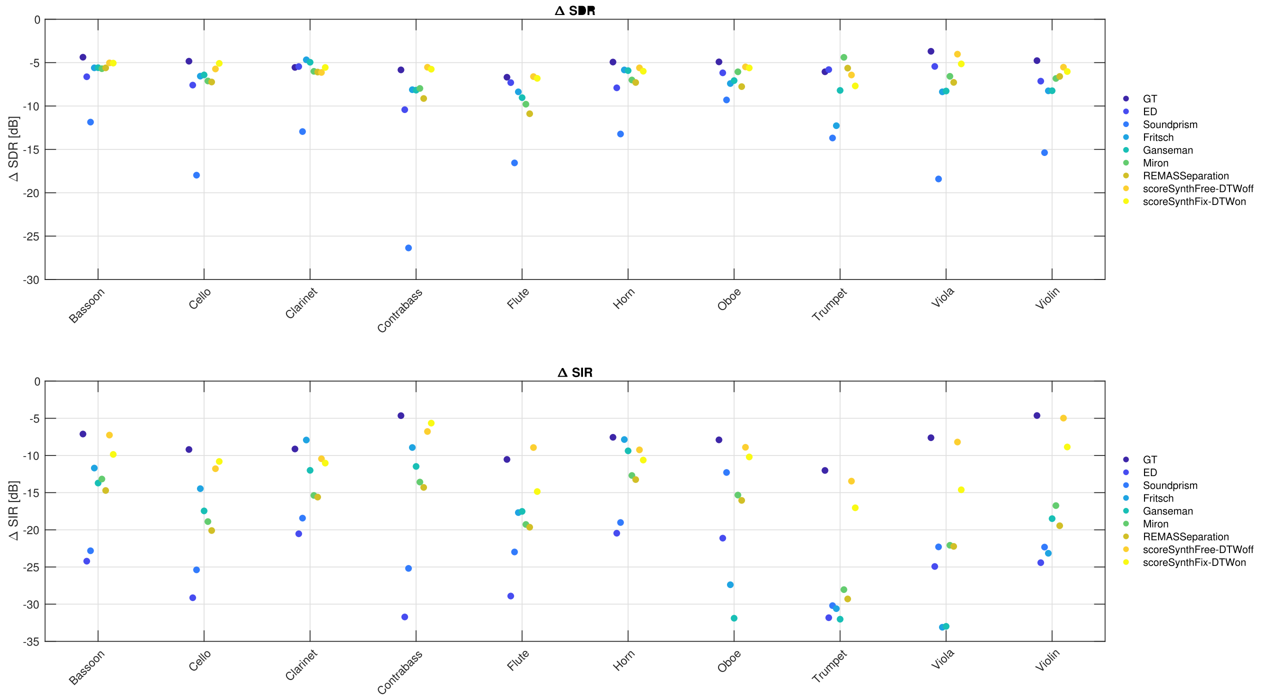


Fig. 9 Overall ΔSDR and ΔSIR results per instrument over the orchestra dataset for the methods presented in Section 5.4

variants, respectively). However, using the psychoacoustic model outperforms in terms of SIR ($\sim$ 3 dB and $\sim$ 3.5 dB in the offline and online variants, respectively). Note that Eq. (29) involves setting to zero those time-frequency bins (f, t) where the interference of the source to be removed might be noticeable. As a consequence, some artifacts are generated in the minus one signal, resulting in worse SAR and SDR values. Nevertheless, the listening performance of the minus one signals obtained with the phychoacoustic model is superior; some examples can be found at demo website[9] . This page contains the audio sources from the test database and the corresponding files obtained by using the proposed method. We have also included an example where the melody line has been removed from the Mozart's composition. In the case of Mozart's piece, the main melody is guided by the bassoon and clarinet. Both instruments have been removed in the mixture applying the psychoacoustical mask described in Eq. (29). This example shows the improvement achieved when the psychoacoustical mask is applied instead of just mixing the corresponding instruments after the separation process.

6 Conclusions

In this paper, we present a signal decomposition-based method for score-informed SS that is suitable for both offline and online applications. Our framework is composed of three stages. First, the score information is encoded as a sequence of individual occurrences of unique combinations of notes (score units). The basis functions for each unit are obtained from a trained in advance dictionary of spectral patterns for each note and instrument in the score. Secondly, a cost function is obtained from the projection of each unit over the whole spectrogram of the input mixture signal and a DTW-based scheme to estimate the activations of each score unit. Finally, a local low-rank NMF approach is used to estimated the time-varying gains, and the source signal is reconstructed using a soft-filter-based strategy.

The proposed method has been evaluated for SS of small and large ensembles single-channel signals obtaining reliable results in terms of SDR, SIR, and perceptual measures in comparison with other signal decomposition and score-informed deep learning approaches. To our best knowledge, the proposed method is the only informed SS method that can be implemented in real-time and obtains reliable results for highly polyphonic large ensembles.

In addition, we have evaluated the proposed method for the task of minus one. To this end, the soft-filter strategy has been modified using a psychoacoustic sinusoidal model. It consists of applying an auditory spectral masking to reduce the interference introduced by the source removed from the audio mixture. The proposed minus one model provide robust results and outperforms in

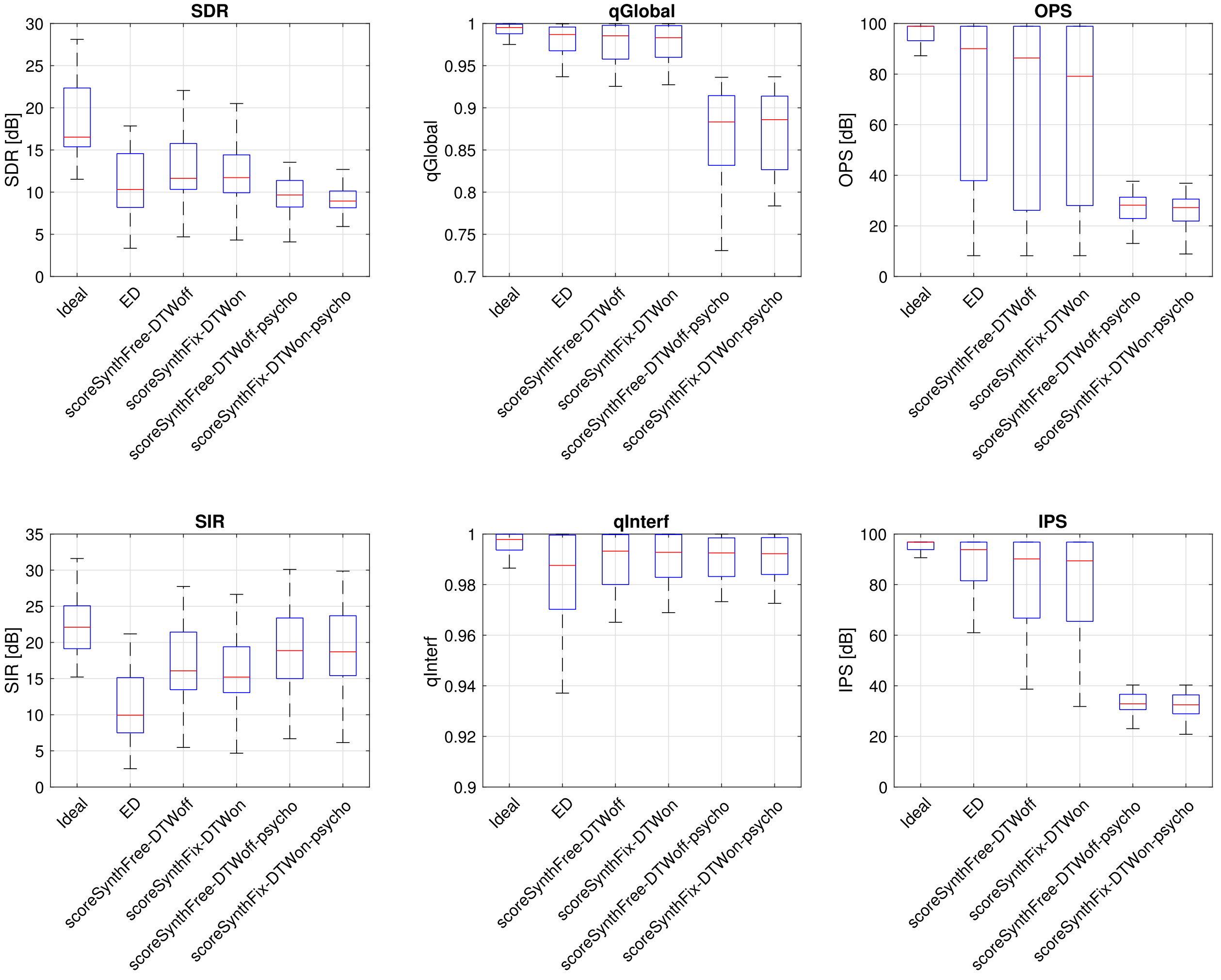


Fig. 10 Objective and perceptual results for minus one application using the BBS_EVAL metrics and the PEASS metrics for the orchestra dataset

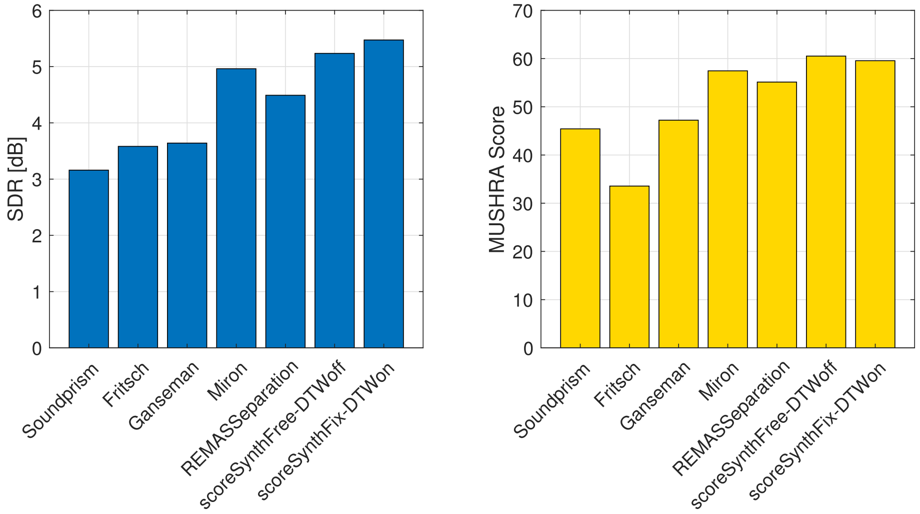


Fig. 11 Comparison between the MUSHRA results and SDR for all the all compared methods in Section 5.4

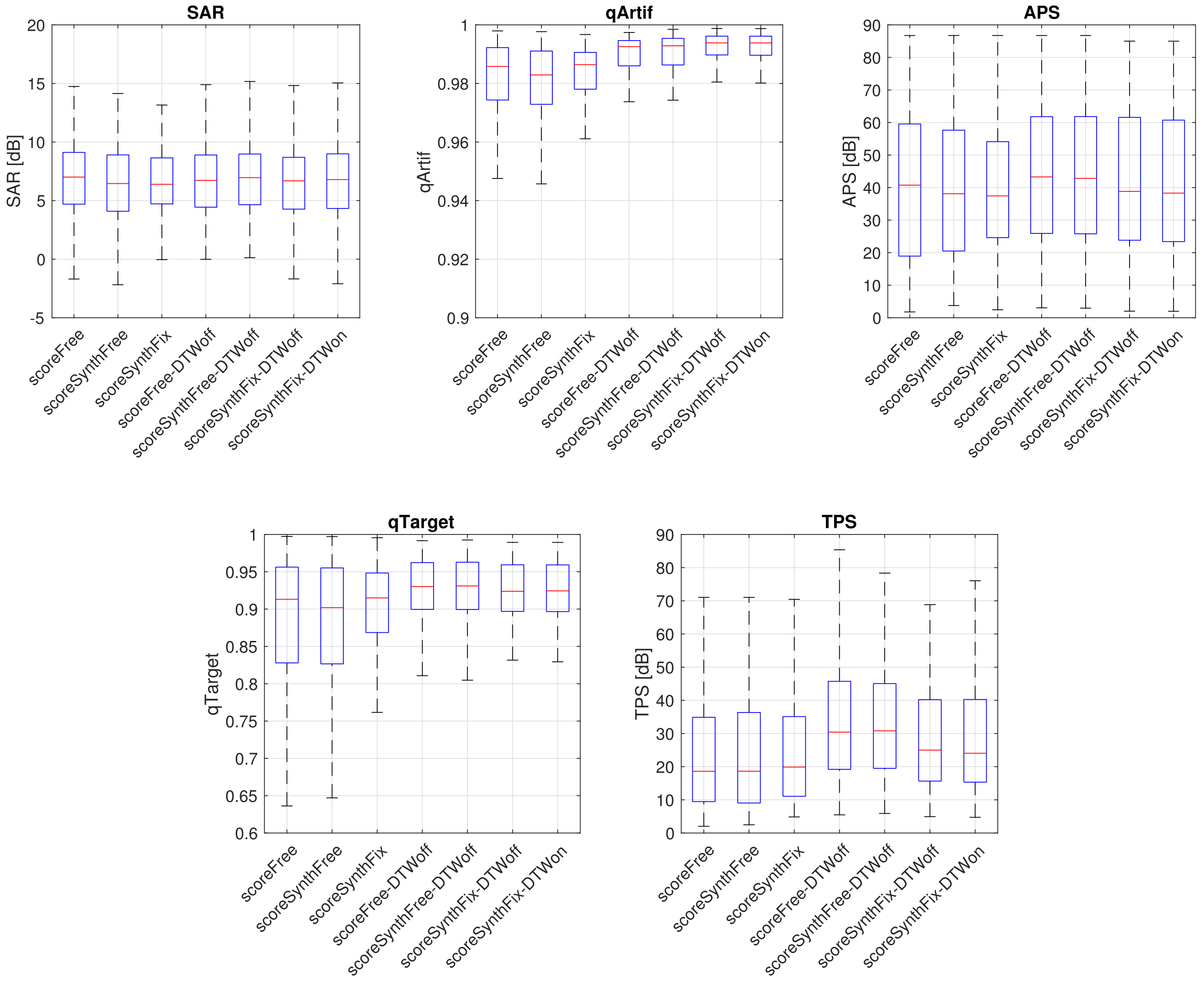


Fig. 12 Additional objective and perceptual measures for Fig. 6

terms of SIR the best performing variants of our proposed method using the soft-filter strategy.

Finally, for future work, we would extend the current framework to a multichannel approach and the use of phase information to mitigate the overlapping partial problem of accompaniment instruments.

Appendix 1: Listening tests

In this section, listening tests were conducted to subjectively assess the audio quality of our SS methods. For this purpose, the MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA) has been employed, which is an ITU-R Recommendation BS.1534-1 [77] implemented in [78].

The MUSHRA listening test [77] is a commonly used method for the subjective evaluation of audio quality. It does not require a huge number of participants to obtain a statistically significant result [79, 80] reference. For this reason, we have used the MUSHRA listening test to evaluate the subjective quality of the SS results of the URMP database. In the MUSHRA test, the participants are provided with the signals under test as well as one reference. The listeners have to grade the different signals on a quality scale between 0 and 100. The participants were allowed to listen to each test signal several times and always had access to the clean reference.

Forty-two listeners, whose ages are from 20 to 40 years old, participated in the MUSHRA test. Five classical music pieces per each participant were randomly chosen from the 44 arrangements of the URMP database, and the separation of the proposed methods (scoreSynthFree-DTWoff and scoreSynthFix-DTWon) and the state-of-the-art methods presented in Section 5.4 were compared for each classical music piece.

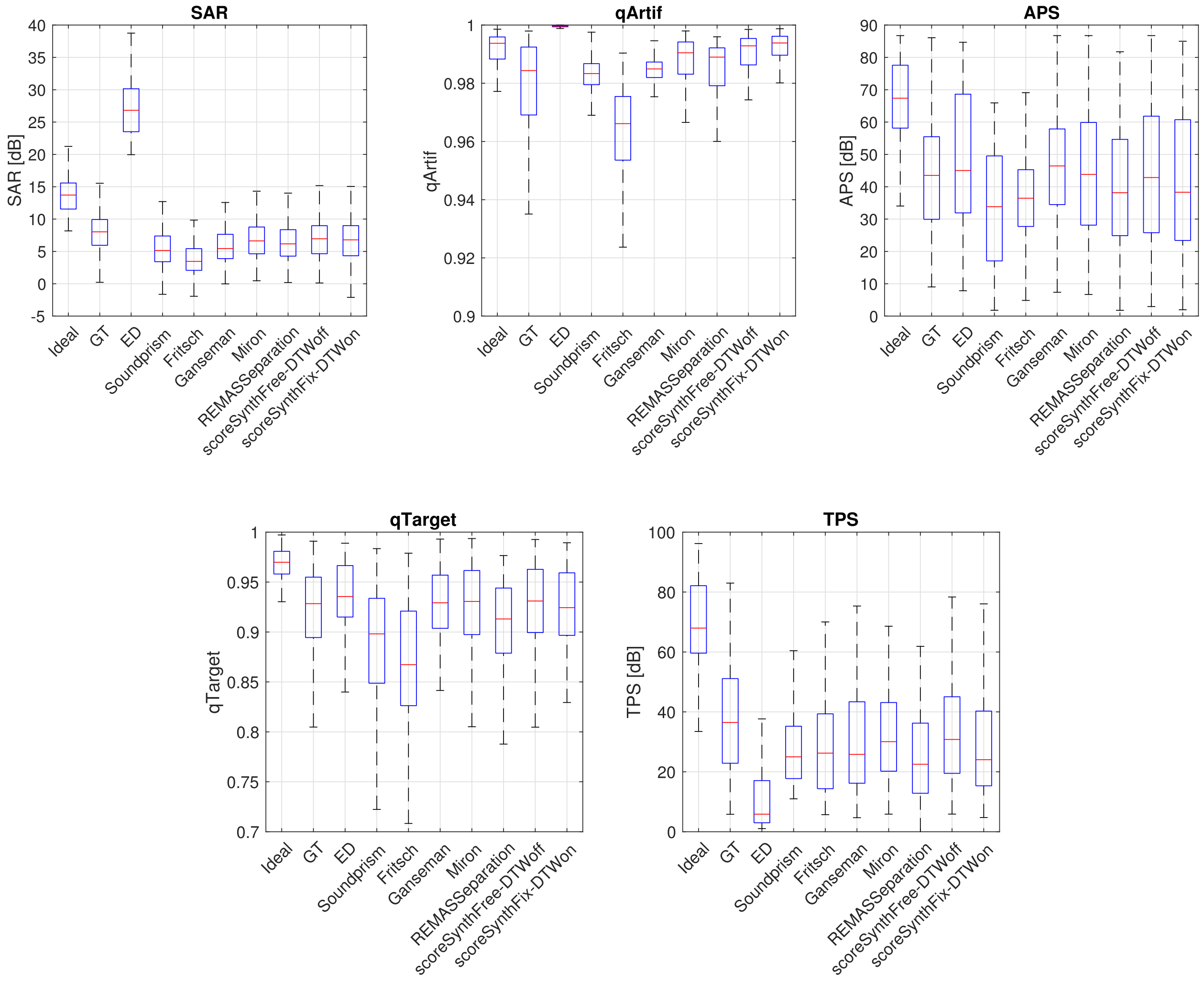


Fig. 13 Additional objective and perceptual measures for Fig. 7

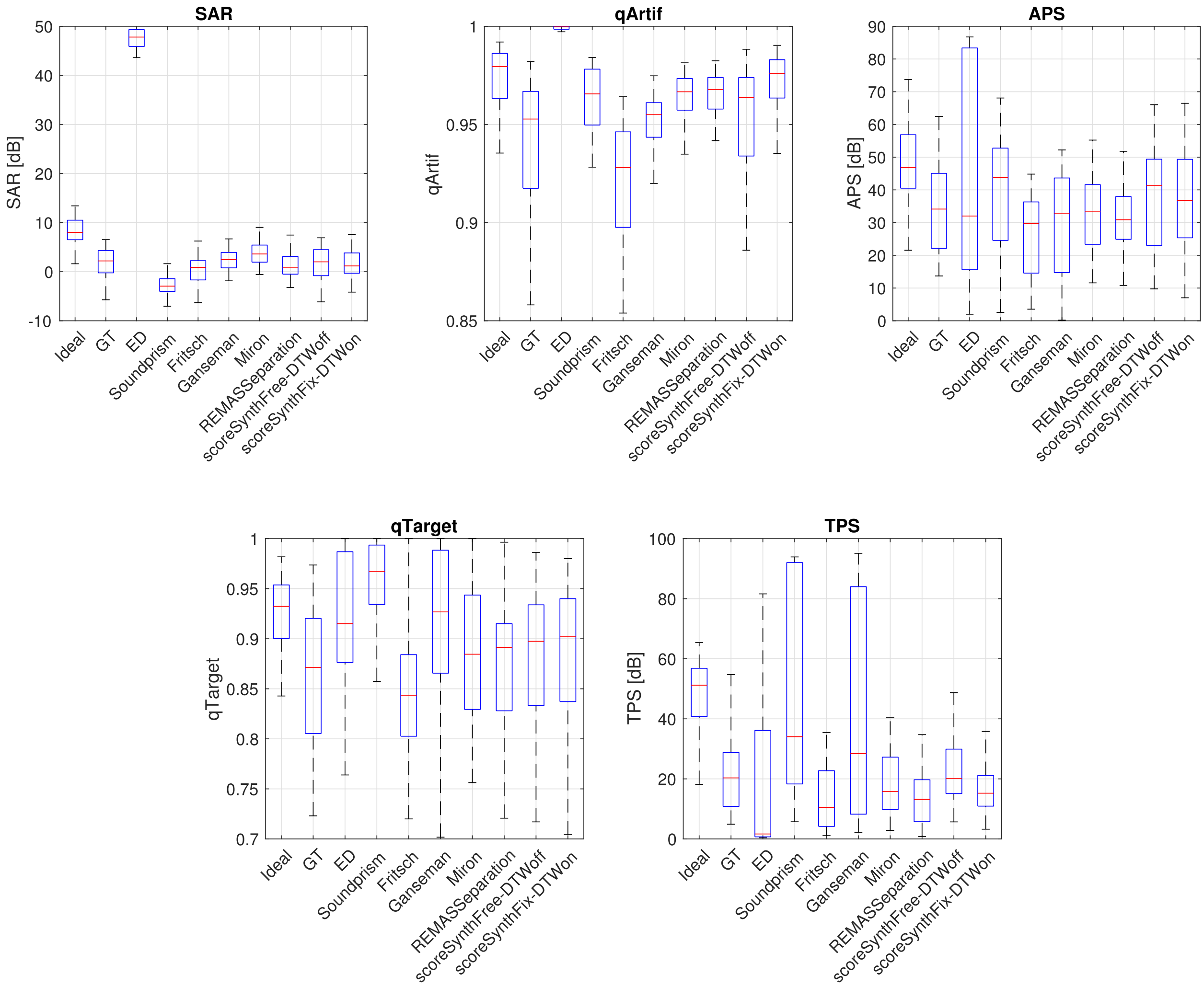


Fig. 14 Additional objective and perceptual measures for Fig. 8

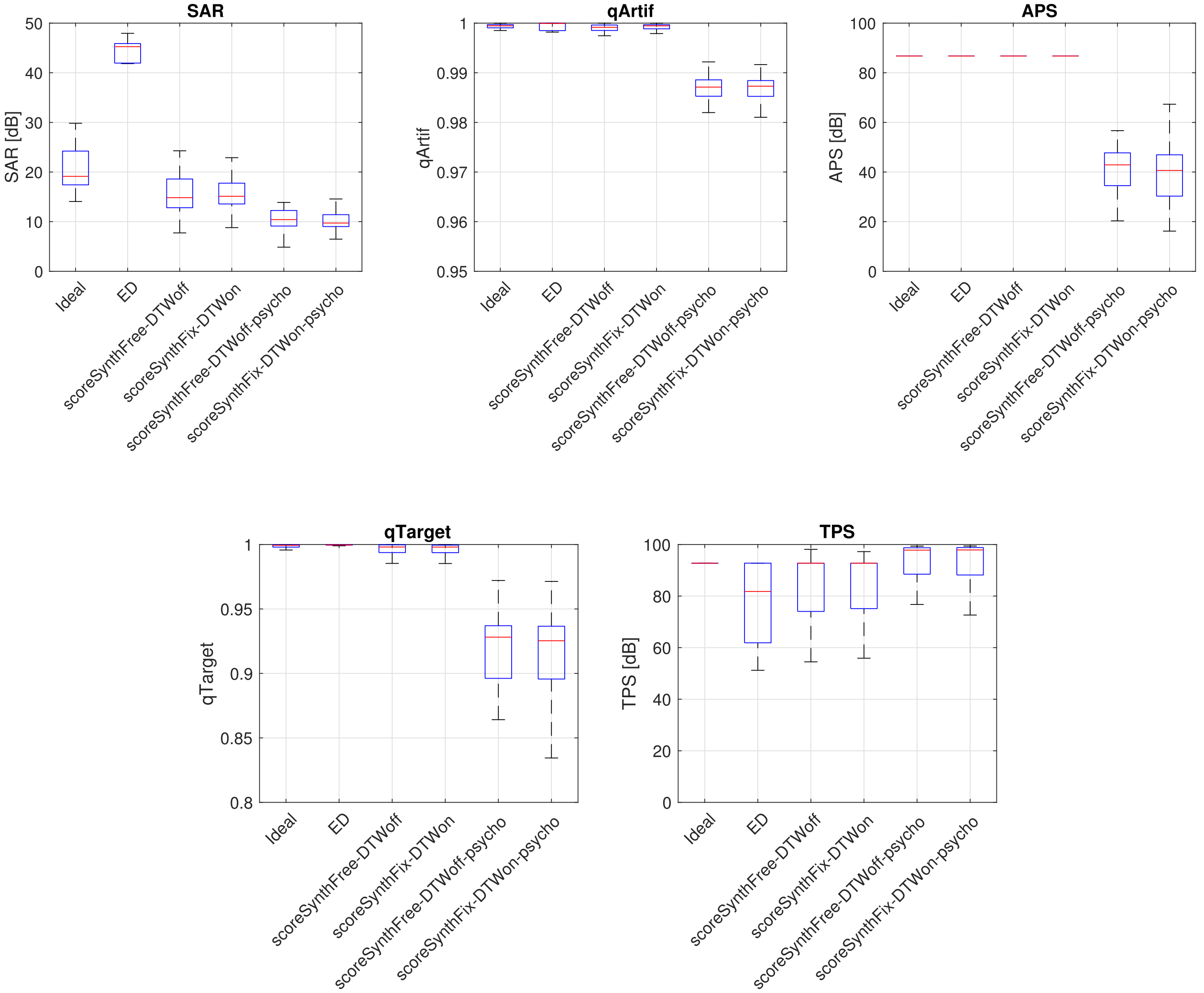


Fig. 15 Additional objective and perceptual measures for Fig. 10

After all the listeners had graded the test signals, a statistical analysis of the results was conducted. Figure 11 shows the average results of SDR and the MUSHRA listening test for all compared methods. As can be observed, the subjective results are in line with the objective ones. Note that the proposed methods yield higher average MUSHRA scores than the other reference methods.

Appendix 2: Complementary material

This section presents additional measures related to Figs. 6, 7, 8, and 10.

Abbreviations

APS: Artifacts-related perceptual score; AR: Aligned rate; CNN: Convolutional neural network; DL: Deep learning; DNN: Deep neural networks; DTW: Dynamic time warping; ED: Energy distribution; EUC: Euclidean distance; GT: Ground-truth; HMM: Hidden Markov model; ICA: Independent component analysis; IPS: Interference-related perceptual score; IS: Itakura-Saito divergence; ISR: Image spatial distortion ratio; KL: Kullback-Leibler divergence; MIREX: Music information retrieval evaluation eXchange; MUSHRA: Multi stimulus test with hidden reference and anchor; NMF: Non-negative matrix factorization; OPS: Overall perceptual score; PCA: Principal component analysis; PLCA: Probabilistic latent component analysis; PSM: Perceptual similarity measures; RWC: Real world computing musical instrument sound database; SAR: Source to artifacts ratio; SDR: Source to distortion ratio; SIR: Source to interference ratio; SS: Source separation; STFT: Short-time fourier transform; TPS: Target-related perceptual score; URMP: University of Rochester Multimodal Music Performance

Acknowledgements

This work was supported by the "Ministerio de Economía y Competitividad" of Spain/FEDER under project TEC2018.

Authors' contributions

All of the authors contributed to the design and implementation of the research, to the analysis of the results, and to the writing of the manuscript. All authors read and approved the final manuscript.

Funding

This work was funded by the "Ministerio de Economía y Competitividad" of Spain/FEDER under projects TEC2018.

Availability of data and materials

The datasets used and analyzed during the current study are available in our repository: https://antoniojmm.github.io/OISS_Minus-One.github.io/. The code of this experimental study is available in our repository: https://github.com/AntonioJMM/OISS_Minus-One.github.io.

Competing interests

The authors declare that they have no competing interests.

Received: 27 March 2019 Accepted: 18 November 2019
Published online: 23 December 2019

References

1. F. J. Canadas-Quesada, D. Fitzgerald, P. Vera-Candeas, N. Ruiz-Reyes, in *Proceedings of the 20th International Conference on Digital Audio Effects (DAFx-17)*. Harmonic-percussive sound separation using rhythmic information from non-negative matrix factorization in single-channel music recordings, (Edinburgh, 2017), pp. 276–282
2. J.-L. Durrieu, G. Richard, B. David, C. Fevotte, Source/filter model for unsupervised main melody extraction from polyphonic audio signals. IEEE Trans. Audio Speech Lang. Process. **18**(3), 564–575 (2010). https://doi.org/10.1109/TASL.2010.2041114
3. J. Nikunen, T. Virtanen, Direction of arrival based spatial covariance model for blind sound source separation. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(3), 727–739 (2014). https://doi.org/10.1109/TASLP.2014.2303576
4. J. J. Carabias-Orti, J. Nikunen, T. Virtanen, P. Vera-Candeas, Multichannel blind sound source separation using spatial covariance model with level and time differences and nonnegative matrix factorization. IEEE/ACM Trans. Audio Speech Lang Process. **26**(9), 1512–1527 (2018). https://doi.org/10.1109/TASLP.2018.2830105
5. L. Wang, H. Ding, F. Yin, Combining superdirective beamforming and frequency-domain blind source separation for highly reverberant signals. EURASIP J. Audio Speech Music. Process. **2010**(1), 1–13 (2010). https://doi.org/10.1155/2010/797962
6. F. J. Rodriguez-Serrano, Z. Duan, P. Vera-Candeas, B. Pardo, J. J. Carabias-Orti, Online score-informed source separation with adaptive instrument models. J. New Music. Res. **44**(2), 83–96 (2015). https://doi.org/10.1080/09298215.2014.989174
7. Y. Mitsufuji, A. Roebel, in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Sound source separation based on non-negative tensor factorization incorporating spatial cue as prior knowledge (IEEE, Vancouver, 2013), pp. 71–75. https://doi.org/10.1109/ICASSP.2013.6637611
8. J. Woodruff, B. Pardo, R. Dannenberg, in *Proceedings of the International Conference on Music Information Retrieval (ISMIR 2006)*. Remixing Stereo Music with Score-Informed Source Separation, (Victoria, 2006), pp. 314–319. https://doi.org/10.5281/zenodo.1414898
9. J. Ganseman, P. Scheunders, G. J. Mysore, J. S. Abel, in *Proceedings of the 2010 International Computer Music Conference, ICMC 2010*. Source separation by score synthesis, (New York, 2010), pp. 1–4. http://hdl.handle.net/2027/spo.bbp2372.2010.108
10. R. Hennequin, B. David, R. Badeau, *Score informed audio source separation using a parametric model of non-negative spectrogram*, (2011), pp. 45–48. https://doi.org/10.1109/ICASSP.2011.5946324
11. S. Ewert, M. Muller, *Estimating note intensities in music recordings*, (2011), pp. 385–388. https://doi.org/10.1109/ICASSP.2011.5946421
12. S. Ewert, M. Muller, *Using score-informed constraints for NMF-based source separation*, (2012), pp. 129–132. https://doi.org/10.1109/ICASSP.2012.6287834
13. M. Miron, J. J. Carabias-Orti, J. J. Bosch, E. Gómez, J. Janer, Score-informed source separation for multi-channel orchestral recordings. J. Electr. Comput. Eng. **2016**, 1–27 (2016). https://doi.org/10.1155/2016/8363507
14. S. Ewert, M. B. Sandler, *Structured dropout for weak label and multi-instance learning and its application to score-informed source separation*, (2017), pp. 2277–2281. https://doi.org/10.1109/ICASSP.2017.7952562
15. M. Miron, J. Janer, E. Gómez, in *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR 2017)*. Monaural score-informed source separation for classical music using convolutional neural networks, (Suzhou, 2017), pp. 55–62. https://doi.org/10.5281/zenodo.1416498
16. O. Slizovskaia, L. Kim, G. Haro, E. Gomez, *End-to-End sound source separation conditioned on instrument labels*, (2018), pp. 306–310. https://doi.org/10.1109/ICASSP.2019.8683800. 1811.01850
17. R. B. Dannenberg, C. Raphael, Music score alignment and computer accompaniment. Commun. ACM. **49**(8), 38 (2006). https://doi.org/10.1145/1145287.1145311
18. A. Cont, A coupled duration-focused architecture for real-time music-to-score alignment. IEEE Trans. Pattern Anal. Mach. Intell. **32**, 974–987 (2010)
19. N. Hu, R. B. Dannenberg, G. Tzanetakis, *Polyphonic audio matching and alignment for music retrieval*, (2003), pp. 185–188. https://doi.org/10.1109/ASPAA.2003.1285862
20. O. Izmirli, R. B. Dannenberg, in *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010)*. Understanding features and distance functions for music sequence alignment, (Utrecht, 2010), pp. 411–416. https://doi.org/10.5281/zenodo.1418353
21. M. S. Puckette, in *Proceedings of the 1995 International Computer Music Conference, ICMC 1995*. Score following using the sung voice, (Banff, 1995), pp. 175–178. http://hdl.handle.net/2027/spo.bbp2372.1995.053
22. Z. Duan, B. Pardo, Soundprism: an online system for score-informed source separation of music audio. IEEE J. Sel. Top. Sign. Process. **5**(6), 1205–1215 (2011). https://doi.org/10.1109/JSTSP.2011.2159701
23. A. Cont, in *IEEE International Conference in Acoustics and Speech Signal Processing (ICASSP)*. Realtime audio to score alignment for polyphonic music instruments using sparse non-negative constraints and hierarchical HMMs (IEEE, Toulouse, 2006). https://doi.org/10.1109/ICASSP.2006.1661258

24. C. Joder, S. Essid, G. Richard, Learning optimal features for polyphonic audio-to-score alignment. IEEE Trans. Audio Speech Lang. Process. **21**(10), 2118–2128 (2013). https://doi.org/10.1109/TASL.2013.2266794
25. P. Cuvillier, A. Cont, in *MLSP 2014 - IEEE International Workshop on Machine Learning for Signal Processing 2014*. Coherent time modeling of semi-Markov models with application to real-time audio-to-score alignment (IEEE, Reims, 2014). https://doi.org/10.1109/MLSP.2014.6958908
26. J. Paulus, A. Klapuri, Drum sound detection in polyphonic music with hidden Markov models. EURASIP J. Audio Speech Music. Process. **2009**(1), 1–9 (2009). https://doi.org/10.1155/2009/497292
27. N. Orio, D. Schwarz, in *Proceedings of the 2001 International Computer Music Conference, ICMC 2001*. Alignment of monophonic and polyphonic music to a score, (Havana, 2001), pp. 155–158. http://hdl.handle.net/2027/spo.bbp2372.2001.104
28. S. Dixon, in *Proc. of the 8th Int. Conference on Digital Audio Effects (DAFx'05)*. Live tracking of musical performances using on-line time warping, (Madrid, 2005), pp. 1727–1728
29. F. J. Rodriguez-Serrano, J. J. Carabias-Orti, P. Vera-Candeas, D. Martinez-Munoz, Tempo driven audio-to-score alignment using spectral decomposition and online dynamic time warping. ACM Trans. Intell. Syst. Technol. **8**(2), 1–20 (2016). https://doi.org/10.1145/2926717
30. A. J. Muñoz-Montoro, P. Vera-Candeas, R. Cortina, E. F. Combarro, P. Alonso-Jordá, Online score-informed source separation in polyphonic mixtures using instrument spectral patterns. Comput. Math. Methods. **1**(4), 1040 (2019). https://doi.org/10.1002/cmm4.1040
31. J. J. Carabias-Orti, F. J. Rodriguez-Serrano, P. Vera-Candeas, N. Ruiz-Reyes, F. J. Canadas-Quesada, in *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR 2015)*. An audio to score alignment framework using spectral factorization and dynamic time warping, (Málaga, 2015), pp. 742–748. https://doi.org/10.5281/zenodo.1418371
32. P. Alonso, R. Cortina, F. J. Rodríguez-Serrano, P. Vera-Candeas, M. Alonso-González, J. Ranilla, Parallel online time warping for real-time audio-to-score alignment in multi-core systems. J. Supercomput. **73**(1), 126–138 (2017). https://doi.org/10.1007/s11227-016-1647-5
33. D. D. Lee, M. Hill, H. S. Seung, Algorithms for non-negative matrix factorization. Adv. Neural Inf. Process. Syst, 556–562 (2001). https://doi.org/10.1007/11785231_58
34. S. Raczyński, N. Ono, in *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*. Multipitch analysis with harmonic nonnegative matrix approximation, (Vienna, 2007), pp. 281–386. https://doi.org/10.5281/zenodo.1417809
35. C. Févotte, N. Bertin, J.-L. Durrieu, Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis. Neural Comput. **21**(3), 793–830 (2009). https://doi.org/10.1162/neco.2008.04-08-771
36. B. Zhu, W. Li, R. Li, X. Xue, Multi-stage non-negative matrix factorization for monaural singing voice separation. IEEE Trans. Audio Speech Lang. Process. **21**(10), 2096–2107 (2013). https://doi.org/10.1109/TASL.2013.2266773
37. F. J. Canadas-Quesada, P. Vera-Candeas, N. Ruiz-Reyes, J. Carabias-Orti, P. Cabanas-Molero, Percussive/harmonic sound separation by non-negative matrix factorization with smoothness/sparseness constraints. EURASIP J. Audio Speech Music Process. **2014**(1), 1–17 (2014). https://doi.org/10.1186/s13636-014-0026-5
38. J. Park, J. Shin, K. Lee, Exploiting continuity/discontinuity of basis vectors in spectrogram decomposition for harmonic-percussive sound separation. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(5), 1061–1074 (2017). https://doi.org/10.1109/TASLP.2017.2681742
39. A. Hyvärinen, E. Oja, Independent component analysis: algorithms and applications. Neural Netw. **13**(4), 411–430 (2000). https://doi.org/10.1016/S0893-6080(00)00026-5
40. A. I. T. Jolliffe, *Principal Component Analysis*. (Springer, New York, 2002)
41. C. Févotte, N. Bertin, J.-L. Durrieu, Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis. Neural Comput. **21**(3), 793–830 (2009)
42. C. Févotte, J. Idier, Algorithms for nonnegative matrix factorization with the β-divergence. Neural Comput. **23**(9), 2421–2456 (2011). https://doi.org/10.1162/NECO_a_00168
43. T. Virtanen, Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. IEEE Trans. Audio Speech Lang. Process. **15**(3), 1066–1074 (2007). https://doi.org/10.1109/TASL.2006.885253
44. P. O. Hoyer, Non-negative matrix factorization with sparseness constraints. J. Mach. Learn. Res. **5**, 1457–1469 (2004)
45. E. Vincent, N. Bertin, R. Badeau, Adaptive harmonic spectral decomposition for multiple pitch estimation. IEEE Trans. Audio Speech Lang. Process. **18**(3), 528–537 (2010). https://doi.org/10.1109/TASL.2009.2034186
46. N. Bertin, R. Badeau, E. Vincent, Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization Applied to Polyphonic Music Transcription. IEEE Trans. Audio Speech Lang. Process. **18**(3), 538–549 (2010). https://doi.org/10.1109/TASL.2010.2041381
47. The 2018 Signal separation evaluation campaign. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture, Notes in Bioinformatics). **10891 LNCS**, 293–305 (2018). https://doi.org/10.1007/978-3-319-93764-9_28
48. P. Chandna, M. Miron, J. Janer, E. Gómez, in *Latent Variable Analysis and Signal Separation. LVA/ICA 2017. Lecture Notes in Computer Science, vol 10169*, ed. by Tichavský P., Babaie-Zadeh M., Michel O., and Thirion-Moreau N. Monoaural Audio Source Separation Using Deep Convolutional Neural Networks (Springer, Cham, 2017), pp. 258–266. https://doi.org/10.1007/978-3-319-53547-0_25
49. A. Pandey, D. Wang, in *Interspeech 2018*. A new framework for supervised speech enhancement in the time domain (ISCA, ISCA, 2018), pp. 1136–1140. https://doi.org/10.21437/Interspeech.2018-1223
50. E. M. Grais, M. U. Sen, H. Erdogan, in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Deep neural networks for single channel source separation (IEEE, 2014), pp. 3734–3738. https://doi.org/10.1109/ICASSP.2014.6854299. 1311.2746
51. P.-S. Huang, M. Kim, M. Hasegawa-Johnson, P. Smaragdis, in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Deep learning for monaural speech separation (IEEE, 2014), pp. 1562–1566. https://doi.org/10.1109/ICASSP.2014.6853860
52. S. Uhlich, F. Giron, Y. Mitsufuji, in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Deep neural network based instrument extraction from music (IEEE, 2015), pp. 2135–2139. https://doi.org/10.1109/ICASSP.2015.7178348
53. P. Smaragdis, S. Venkataramani, in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. A neural network alternative to non-negative audio models (IEEE, 2017), pp. 86–90. https://doi.org/10.1109/ICASSP.2017.7952123. 1609.03296
54. Y. Luo, Z. Chen, J. R. Hershey, J. Le Roux, N. Mesgarani, in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Deep clustering and conventional networks for music separation: stronger together (IEEE, 2017), pp. 61–65. https://doi.org/10.1109/ICASSP.2017.7952118. 1611.06265. http://ieeexplore.ieee.org/document/7952118/
55. F. J. Canadas-Quesada, P. Vera-Candeas, D. Martinez-Munoz, N. Ruiz-Reyes, J. J. Carabias-Orti, P. Cabanas-Molero, Constrained non-negative matrix factorization for score-informed piano music restoration. Dig. Signal Process. Rev. J. (2016). https://doi.org/10.1016/j.dsp.2016.01.004
56. J. Fritsch, M. D. Plumbley, *Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis*, (2013), pp. 888–891. https://doi.org/10.1109/ICASSP.2013.6637776
57. J. J. Carabias-Orti, F. J. Rodriguez-Serrano, P. Vera-Candeas, F. J. Cañadas-Quesada, N. Ruiz-Reyes, Constrained non-negative sparse coding using learnt instrument templates for realtime music transcription. Eng. Appl. Artif. Intell. **26**(7), 1671–1680 (2013). https://doi.org/10.1016/j.engappai.2013.03.010
58. Dynamic time warp (dtw), In Matlab, web resource. http://www.ee.columbia.edu/pwe/resources/matlab/dtw/
59. M. Goto, H. Hashiguchi, T. Nishimura, R. Oka, in *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)*. RWC music database: popular, classical and jazz music databases, (Paris, 2002), pp. 287–288. https://doi.org/10.5281/zenodo.1416474
60. M. Goto, in *Proceedings of the 18th International Congress on Acoustics (ICA 2004) 4-9 April 2004*. Development of the RWC music database, (Kyoto, 2004), pp. 553–556
61. J. J. Carabias-Orti, T. Virtanen, P. Vera-Candeas, N. Ruiz-Reyes, F. J. Canadas-Quesada, Musical instrument sound multi-excitation model for non-negative spectrogram factorization. IEEE J. Sel. Top. Signal Process. **5**(6), 1144–1158 (2011). https://doi.org/10.1109/JSTSP.2011.2159700
62. G. Zhou, A. Cichocki, S. Xie, Fast nonnegative matrix/tensor factorization based on low-rank approximation. IEEE Trans. Signal Process. **60**(6), 2928–2940 (2012). https://doi.org/10.1109/TSP.2012.2190410

63. S. van de Par, A. Kohlrausch, G. Charestan, R. Heusdens, *A new psychoacoustical masking model for audio coding applications* (IEEE, 2002), pp. 1805–1808. https://doi.org/10.1109/ICASSP.2002.5744974
64. H. Fastl, E. Zwicker, *Psychoacoustics*. (Springer, 2007). https://doi.org/10.1007/978-3-540-68888-4. arXiv:1011.1669v3
65. B. Li, X. Liu, K. Dinesh, Z. Duan, G. Sharma, Creating a multitrack classical music performance dataset for multimodal music analysis: challenges, insights, and applications. IEEE Trans. Multimed. **21**(2), 522–535 (2019). https://doi.org/10.1109/TMM.2018.2856090
66. J. Pätynen, V. Pulkki, T. Lokki, Anechoic recording system for symphony orchestra. Acta Acustica United Acustica. **94**(6), 856–865 (2008). https://doi.org/10.3813/AAA.918104
67. J. Parras-Moral, F. Canadas-Quesada, P. Vera-Candeas, *Audio restoration of solo guitar excerpts using a excitation-filter instrument model* (In Stockholm Music Acoustics Conference jointly with Sound And Music Computing Conference, 2013), pp. 654–659
68. J. J. Carabias-Orti, M. Cobos, P. Vera-Candeas, F. J. Rodríguez-Serrano, Nonnegative signal factorization with learnt instrument models for sound source separation in close-microphone recordings. EURASIP J. Adv. Signal Process. **2013**(1), 184 (2013). https://doi.org/10.1186/1687-6180-2013-184
69. F. J. Rodríguez-Serrano, J. J. Carabias-Orti, P. Vera-Candeas, F. J. Canadas-Quesada, N. Ruiz-Reyes, Monophonic constrained non-negative sparse coding using instrument models for audio separation and transcription of monophonic source-based polyphonic mixtures. Multimed. Tools Appl. **72**(1), 925–949 (2014). https://doi.org/10.1007/s11042-013-1398-8
70. E. Vincent, R. Gribonval, C. Fevotte, Performance measurement in blind audio source separation. IEEE Trans. Audio Speech Lang. Process. **14**(4), 1462–1469 (2006). https://doi.org/10.1109/TSA.2005.858005
71. V. Emiya, E. Vincent, N. Harlander, V. Hohmann, Subjective and objective quality assessment of audio source separation. IEEE Trans. Audio Speech Lang. Process. **19**(7), 2046–2057 (2011). https://doi.org/10.1109/TASL.2011.2109381
72. R. Huber, B. Kollmeier, PEMO-Q-A new method for objective audio quality assessment using a model of auditory perception. IEEE Trans. Audio Speech Lang. Process. **14**(6), 1902–1911 (2006). https://doi.org/10.1109/TASL.2006.883259
73. J. Fritsch, J. Ganseman, M. D. Plumbley, in *International Conference on Machine Learning - June 26-July 1, 2012*. A comparison of two different methods for score-informed source separation, (Edinburgh, 2012), pp. 1–2
74. R. Turetsky, D. Ellis, *Ground-truth transcriptions of real music from force-aligned MIDI syntheses*, (2003), pp. 135–141. https://doi.org/10.7916/D8S472CZ
75. A. Liutkus, D. Fitzgerald, Z. Rafii, B. Pardo, L. Daudet, Kernel additive models for source separation. IEEE Trans. Signal Process. **62**(16), 4298–4310 (2014). https://doi.org/10.1109/TSP.2014.2332434
76. T. Jan, W. Wang, in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. Joint blind dereverberation and separation of speech mixtures (IEEE, Bucharest, 2012), pp. 2343–2347
77. ITU-R BS.1534-3, Method for the subjective assessment of intermediate quality level of audio systems. International Telecommunication Union (2015)
78. E. Vincent, MUSHRAM: A MATLAB interface for MUSHRA listening tests (2005). http://c4dm.eecs.qmul.ac.uk/downloads/
79. F. Deng, C. C. Bao, Speech enhancement based on Bayesian decision and spectral amplitude estimation. Eurasip J. Audio Speech Music Process. **2015**(1) (2015). https://doi.org/10.1186/s13636-015-0073-6
80. Z. Ben-Hur, D. L. Alon, B. Rafaely, R. Mehra, Loudness stability of binaural sound with spherical harmonic representation of sparse head-related transfer functions. Eurasip J. Audio Speech Music Process. **2019**(1) (2019). https://doi.org/10.1186/s13636-019-0148-x

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.