Extended Semantic Initialization for NMF-based Audio Source Separation

Christian Rohlfing and Julian M. Becker
Institut für Nachrichtentechnik
RWTH Aachen University
D-52056 Aachen, Germany
Email: rohlfing@ient.rwth-aachen.de

***Abstract*—Nonnegative matrix factorization (NMF) is often used for source separation of audio signals. In most of these algorithms, the initialization step of the NMF, which has a strong impact on the separation performance, is based on random values or deterministic methods such as singular value decomposition (SVD). Another deterministic initialization approach, which is used e.g. for score-informed source separation algorithms, makes use of synthesized magnitude spectra of harmonic notes. It was shown that this semantic method leads to good separation results in blind source separation (BSS) as well; not only for harmonic but also for percussive mixtures with some harmonic components. In this paper, we present an extension to the semantic approach to enhance the separation quality for arbitrary audio mixtures. We evaluate this extension in a BSS scenario and compare it to other initialization schemes.**

I. Introduction

Nonnegative matrix factorization (NMF) based algorithms are frequently used for audio source separation. The NMF initialization is an important part of these algorithms as it influences the separation quality to a great extent. Furthermore, it can be used for adaptation of prior information (e.g. [1], [2], [3]). There exist different NMF initialization schemes based on random values or data-driven schemes such as the singular value decomposition (SVD) [4].

Other deterministic methods try to model the data's semantics: For example, [5] uses real piano notes for initializing the NMF for automatic music transcription. Synthetic initializations with sparse comb structures for harmonic notes and additional noise or wideband components to model percussive components are employed for e.g. score-informed [3], [6] or guided source separation [2] as well as for multiple pitch estimation [7].

To our knowledge, there exist only rather simple wideband or noise models for synthetic initialization of unpitched note spectra which are not as accurate as harmonic models for pitched note spectra. On the other hand, these simple models increase the separation quality of e.g. score-informed separation methods [3], [6] due to the fact, that these methods are able to estimate the onset time of each percussive event by evaluating the mixture's score.

In this paper, we try to combine semantic harmonic initialization schemes to model pitched components with data-driven initialization schemes to take all components into account, which cannot be modelled correctly by the harmonic scheme. We evaluate this novel approach as initialization for a basic NMF-based blind source separation (BSS) algorithm as proposed in [8]. Note, that the application of our approach is not limited to BSS.

The paper is structured as follows: In Section II, the BSS algorithm and in Section III a synthetic harmonic initialization scheme are summarized. Our novel extension to that scheme is introduced in Section IV and evaluated in Section V. Section VI concludes this paper and gives an outlook on future work.

II. Blind Audio Source Separation with Beta-NMF

The basic blind audio source separation algorithm is shown in Figure 1 and described in detail in [8]:

The time-domain mixture $\mathbf{x} = \sum_{m=1}^{M} \mathbf{s}_m$ consisting of M sources $\mathbf{s}_m$ is transformed to the time-frequency domain by the short-time Fourier transform (STFT). In the following dimension reduction step (DR), the spectral dimension of the mixture amplitude spectrogram is reduced to suppress vibrato effects and to speed up the subsequent processing steps. The dimension reduction is obtained by filtering the spectrogram with a mel-filterbank which consists of K triangular filters whose central frequencies are spaced linearly on the mel scale $f_{\mathrm{mel}} = 1127 \log\left(1 + f_{\mathrm{Hz}}/700\right)$.

The dimension-reduced magnitude spectrogram $\mathbf{X} \in \mathbb{R}_{+}^{K \times T}$ is factorized by the NMF into I components $\mathbf{X} \approx \tilde{\mathbf{X}} = \mathbf{B}\mathbf{G}^{\mathbf{T}}$ with the spectral basis matrix $\mathbf{B} \in \mathbb{R}_{+}^{K \times I}$, temporal gain matrix $\mathbf{G} \in \mathbb{R}_{+}^{T \times I}$ and number of time frames T. I is a user defined parameter and should be chosen much smaller than both K and T but larger than the number of sources M. The NMF variant used here denoted as β-NMF estimates $\mathbf{B}$ and $\mathbf{G}$ by evaluating multiplicative update rules which are derived by minimizing the β-Divergence [8]. Common choices in the literature for values of β for using β-NMF for audio source separation are $\beta = 0$ (Itakura-Saito distance) or $\beta = 1$ (Kullback-Leibler Divergence). The separation quality strongly depends on the initial matrices $\mathbf{B}_0$ and $\mathbf{G}_0$ which are computed in an initialization step in advance.

For synthesis, fine structure and phase information of the components are obtained by Wiener-like filtering with the original complex mixture spectrogram: First, the spectral dimension of $\mathbf{B}$ is transformed to the linear frequency domain by an inverse dimension reduction step. Afterwards, a so-called

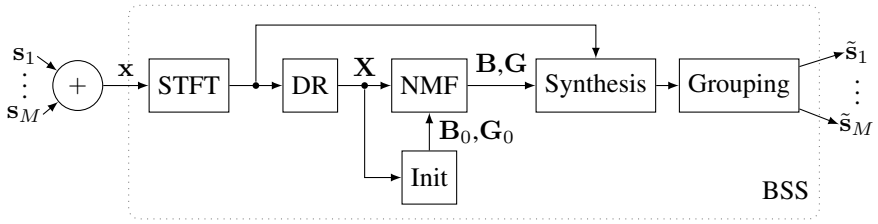


Fig. 1. Flowgraph of the BSS algorithm.

time-frequency mask is determined for each NMF component depending on the corresponding spectral basis and the temporal gain vectors. The actual Wiener-filtering is conducted by multiplying the complex mixture spectrogram with the mask corresponding to each component. The resulting spectrogram is finally transformed to time domain by an inverse STFT [8].

In the grouping step, the components are grouped together to form M estimated sources $\tilde{\mathbf{s}}_m$ since the number of NMF components I is usually larger than M. In this paper, we use a non-blind grouping procedure with knowledge of the original sources to rule out possible errors by a blind grouping step. This procedure yields the grouping assignment which maximizes the overall separation quality [8].

III. Semantic Harmonic Initialization

As discussed in Section I, several algorithms use a synthetic semantic initialization (SI) for the NMF basis matrix consisting of harmonic spectra. In this paper, we follow the procedure proposed in [8] to obtain the initial basis matrix $\mathbf{B}_0$ by synthesizing the spectra of $I_0 = 88$ piano notes: Each note with fundamental frequency $\mathbf{f}(i) = 27.5\,(2^{\frac{1}{12}})^{i-1}\,\mathrm{Hz}$ is calculated at first in time domain consisting of $N = 20$ partials and stored in a matrix $\mathbf{P} \in \mathbb{R}^{K_0 \times I_0}$

$$\mathbf{P}(t,i) = \sum_{n=1}^{N} \cos\left(2\pi\,(t-1)\frac{n\,\mathbf{f}(i)}{F_\mathrm{s}}\right) C_\mathrm{att}(n\,\mathbf{f}(i))\ , \qquad (1)$$

with $1 \le t \le K_0$. Factor C_att depending on the partial's frequency $n\,\mathbf{f}(i)$ ensures an attenuation of $3\,\mathrm{dB}$ per octave, F_s denotes the sampling frequency and K_0 the length of the STFT window.

Afterwards, the time-domain piano notes $\mathbf{P}$ are transformed to the mel domain by applying the windowed Fourier transform and the mel filterbank consecutively to each column of $\mathbf{P}$. This procedure mimics the transformation of the mixture in time domain $\mathbf{x}$ to the mel-filtered spectrogram $\mathbf{X}$ (see Section II). As a last step, each mel-domain piano note spectrum is normalized to unit energy and stored in $\mathbf{B}_0 \in \mathbb{R}^{K \times I_0}$.

The gain matrix $\mathbf{G}_0 \in \mathbb{R}^{T \times I_0}$ is then initialized as

$$\mathbf{G}_0 = \mathbf{X}^\mathbf{T}\mathbf{B}_0 \quad \text{or} \quad \mathbf{G}_0(t,i) = \sum_k \mathbf{X}(k,t)\mathbf{B}_0(k,i) \qquad (2)$$

which can be interpreted as a correlation at lag zero between the ith piano note spectrum (ith column of $\mathbf{B}_0$) and each mixture frame at time bin t (tth column of $\mathbf{X}$). Therefore, $\mathbf{G}_0$ gives information about the similarity (defined by high correlation) of the ith spectrum and the frame at time bin t. As Eq. 2 is computed in the mel domain, possible errors due to tuning differences between the synthesized notes and the mixture frames or vibrato effects are reduced.

To decrease the number of components from $I_0 = 88$ to the user-specified I, the component with the lowest energy $\mathbf{E}(i) = \left[\sum_k \mathbf{B}(k,i)^2\right]\left[\sum_t \mathbf{G}(t,i)^2\right]$ is discarded after each of the first $I_0 - I$ NMF iteration steps until the condition $I_0 = I$ is reached [8]. Afterwards, the NMF proceeds unmodified. This modification is a disadvantage as the initialization requires the NMF algorithm to be changed. However, this is only a subtle change compared to the differences between the NMF and more complex algorithms to reduce the number of components automatically such as [9].

IV. Data-driven Extension

The semantic initialization (SI) described in the previous section assumes a harmonic structure of the mixture's spectrogram $\mathbf{X}$. Instead of appending fixed wide-band spectra to deal with non-harmonic components, we want to gain information out of the mixture itself. Hence, we combine the SI with a data-driven initialization. Figure 2 shows the proposed procedure:

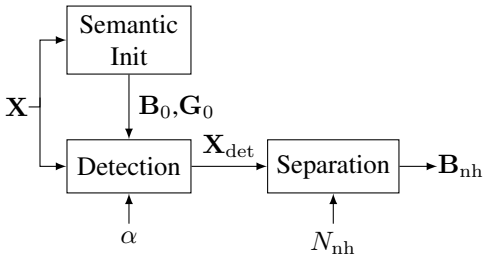


Fig. 2. Flowgraph of the proposed method.

Based on the initial gain matrix $\mathbf{G}_0$ calculated by SI, all the frames of the spectrogram $\mathbf{X}$ which are not properly represented by the harmonic spectra in $\mathbf{B}_0$ are detected and stored in a matrix $\mathbf{X}_\mathrm{det}$[1]. After detecting these frames, a separation method is used to extract N_nh new basis vectors $\mathbf{B}_\mathrm{nh}$ which are simply appended to $\mathbf{B}_0$ calculated by SI to form a new initial basis matrix. N_nh is a user-defined parameter. The corresponding gain vectors are obtained by evaluating Eq. (2) and the number of initial NMF components is updated to $I_0 = 88 + N_\mathrm{nh}$.

[1] In the following, we denote all events, which are not properly represented by $\mathbf{B}_0$ as "non-harmonic".

A. Detection of Non-harmonic Frames

Figure 3 shows the mixture spectrogram $\mathbf{X}$, the spectral basis matrix $\mathbf{B}_0$ and the (transposed) initial gain matrix $\mathbf{G}_0^{\mathbf{T}}$ obtained by the semantic initialization for an exemplary mixture of a guitar and a drum recording. As shown in Fig. 3c, the tth column of $\mathbf{G}_0^{\mathbf{T}}$ gives a good indication if frame t of the spectrogram $\mathbf{X}$ can be represented by the ith column of $\mathbf{B}_0$ as Eq. (2) evaluates the correlation at lag zero between these two vectors. For harmonic notes in frame t of $\mathbf{X}$, most of the energy in the tth column of $\mathbf{G}_0^{\mathbf{T}}$ is concentrated at bin i corresponding to the ith piano note spectrum (ith column of $\mathbf{B}_0$) with similar fundamental frequency and at the harmonics of the note. Regarding non-harmonic events, a wide range of spectra of $\mathbf{B}_0$ are necessary to represent frame t. In the following, we discuss several methods to detect non-harmonic frames which are combined to a matrix $\mathbf{X}_{\mathrm{det}}$.

1) Spectral Flatness: A simple measure for harmonicity is the spectral flatness as non-harmonic spectra tend to be more flat than harmonic ones. In [10], a robust entropy based variant $d_{\mathrm{sf}}(t)$ was proposed:

$$\log_2\left(d_{\mathrm{sf}}(t)+1\right) = -\frac{1}{\log_2(I_0)}\sum_i \hat{\mathbf{G}}_0(t,i)\log_2\left(\hat{\mathbf{G}}_0(t,i)\right) \quad (3)$$

with $\hat{\mathbf{G}}_0(t,i) = \mathbf{G}_0(t,i)/\left[\sum_j \mathbf{G}_0(t,j)\right]$. The values of $d_{\mathrm{sf}}(t)$ are restricted to $0 \leq d_{\mathrm{sf}}(t) \leq 1$. Note that we evaluate the flatness of each column of the temporal gain matrix $\mathbf{G}_0^{\mathbf{T}}$ instead of the corresponding spectral basis vectors or columns of $\mathbf{X}$ as we try to detect the frames, for which a wide range of piano notes was selected. As this measure is called spectral flatness in the literature, we adopt this name in the following.

Non-harmonic frames are detected with a threshold-decision $d_{\mathrm{sf}}(t) > \alpha_{\mathrm{sf}}$. In Figure 3d, $d_{\mathrm{sf}}(t)$ is exemplary depicted.

2) Correlation: We propose another detection function for non-harmonic frames which evaluates the autocorrelation of each column t of $\mathbf{G}_0^{\mathbf{T}}$

$$\mathbf{C}(j,t) = \frac{1}{\sum_{i=1}^{I_0}\mathbf{G}_0(t,i)^2}\sum_{i=1}^{I_0}\mathbf{G}_0(t,i)\mathbf{G}_0(t,i-j+1) \quad (4)$$

with $1 \leq j \leq I_0$ and $\mathbf{C}(j,t)$ being normalized such that $\mathbf{C}(j=1,t)=1$. As shown in Figure 3f, the autocorrelation $\mathbf{C}$ discards the pitch-information itself but provides a good measure to discriminate harmonic from non-harmonic notes: For most harmonic components, the corresponding column t of $\mathbf{C}$ features a strong peak at lag $j_{\mathrm{oct}} = 13$ which corresponds to one octave. For non-harmonic notes, the autocorrelation function does not show this but a rather smeared structure. The detection function itself evaluates the height of the peak at lag j_{oct}

$$d_{\mathrm{c}}(t) = \frac{1}{2}\left[\frac{\mathbf{C}(j_{\mathrm{oct}},t)}{\mathbf{C}(j_{\mathrm{oct}}-j_0,t)} + \frac{\mathbf{C}(j_{\mathrm{oct}},t)}{\mathbf{C}(j_{\mathrm{oct}}+j_0,t)}\right] \quad (5)$$

with $j_0 = 2$ to take small smearing effects in $\mathbf{C}(j,t)$ into account.

Similar to the spectral flatness detection function Eq. (3), we select non-negative frames with thresholding Eq. (5) $d_{\mathrm{c}}(t) < \alpha_{\mathrm{c}}$. In Figure 3e, $d_{\mathrm{c}}(t)$ and α_{c} are shown for the exemplary guitar-drum mixture.

B. Separation Step

The detected non-harmonic frames $\mathbf{X}_{\mathrm{det}}$ could still contain some harmonic components if a non-harmonic and a harmonic source are active at the same time frame t or in case of a faulty detection. Besides, multiple non-harmonic sources (such as different percussive instruments or a percussive instrument and noise, for example) could be active as well.

Therefore, we propose to separate $\mathbf{X}_{\mathrm{det}}$ into N_{nh} additional basis functions $\mathbf{B}_{\mathrm{nh}}$ which are then simply appended to $\mathbf{B}_0$. As a separation method, we chose the nonnegative variant of the SVD (NNDSVD) which was introduced in [4]. This algorithm evaluates the SVD, selects the N_{nh} highest singular values and cancels out iteratively negative components in the output matrices of the SVD. Another possible method would be a random-initialized NMF. In this paper, however, we focus on deterministic initialization methods.

V. Evaluation

A. Setup

For evaluation of the extended semantic initialization (ESI), we performed source separation as summarized in Section II on a database of 60 audio sources (harmonic, percussive, vocals, speech and noise signals) sampled at $F_{\mathrm{s}} = 44\,100\,\mathrm{Hz}$ as described in [8]. We evaluated mixtures consisting of $M = 2$ sources, resulting in a total of 1770 mixtures.

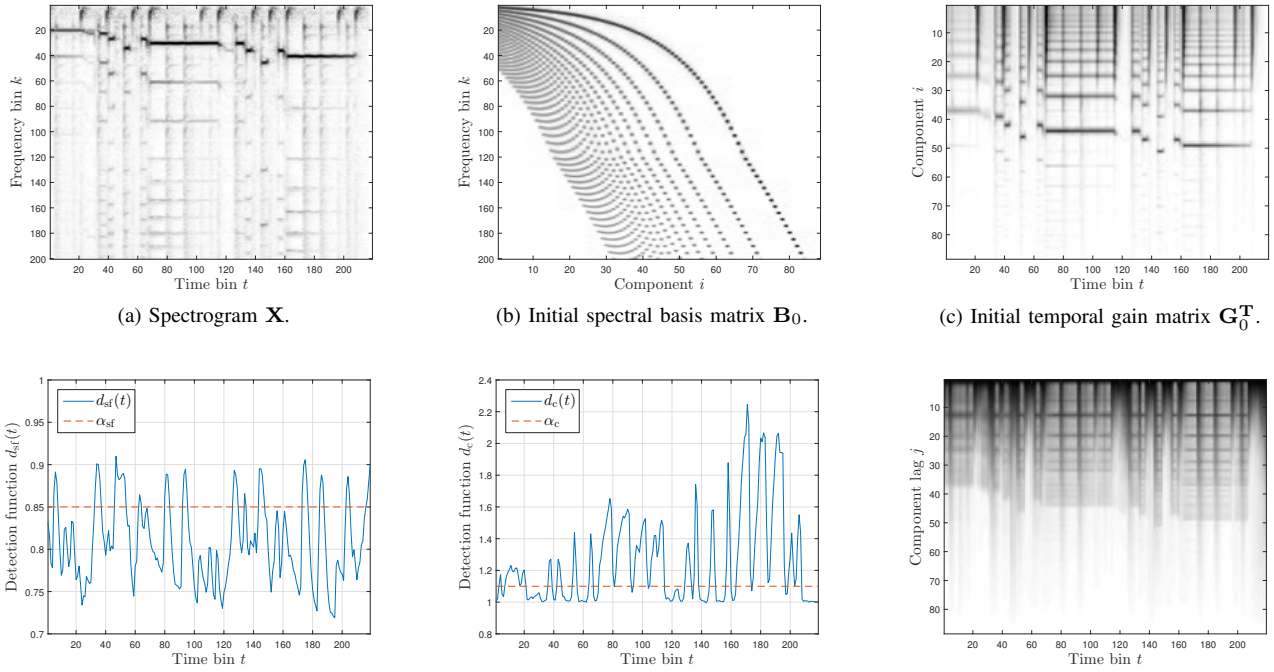
Regarding the STFT, we chose a window size corresponding to $93\,\mathrm{ms}$ and a hop size of $23\,\mathrm{ms}$. The mel-filtering was done with $K = 400$ filters. We set the number of NMF components to $I = 20$ and performed 300 NMF iterations. We chose $\beta = 0$ as it is more robust regarding loudness differences between the sources [8].

The parameters for thresholding the detection functions were chosen as follows: For the spectral flatness function $d_{\mathrm{sf}}(t)$, we chose eleven different values in the interval $\alpha_{\mathrm{sf}} \in [0.8, 0.9]$. For the correlation-based detection $d_{\mathrm{c}}(t)$, we chose eleven values in the interval $\alpha_c \in [1, 1.1]$. We also evaluated the case, where all frames are detected by choosing $\alpha_{\mathrm{sf}} = 0$. For the separation step, the number of additional non-harmonic components $\mathbf{B}_{\mathrm{nh}}$ calculated by the NNDSVD was chosen to $N_{\mathrm{nh}} \in \{2, 3, 5\}$.

As quality measures we chose the signal-to-distortion ratio (SDR), the signal-to-interference ratio (SIR) and the signal-to-artifact ratio (SAR) which we calculated by using the $\mathrm{BSS}_{\mathrm{EVAL}}$ toolbox [11] and averaged over all mixtures.

B. Results

We evaluated the so-called oracle estimator [12] which yields an upper bound for separation algorithms which use the Wiener-like filtering for synthesis (refer to Section II) on the database described in Section V-A as well as the harmonic semantic initialization scheme (SI) described in Section III



(a) Spectrogram $\mathbf{X}$.

(b) Initial spectral basis matrix $\mathbf{B}_0$.

(c) Initial temporal gain matrix $\mathbf{G}_0^{\mathbf{T}}$.

(d) Detection function $d_{\mathrm{sf}}(t)$.

(e) Detection function $d_{\mathrm{c}}(t)$.

(f) Correlation matrix $\mathbf{C}$.

Fig. 3. Mel-filtered mixture spectrogram $\mathbf{X}$, initial spectral basis matrix $\mathbf{B}_0$ and temporal gain matrix $\mathbf{G}_0^{\mathbf{T}}$ obtained by the basic semantic initialization, spectral flatness and correlation-based detection functions $d_{\mathrm{sf}}(t)$ and $d_{\mathrm{c}}(t)$ as well as the correlation matrix $\mathbf{C}$ for an exemplary guitar-drum mixture.

and the non-negative SVD variant NNDSVD summarized in Section IV-B as it is used for separation of the detected non-harmonic frames[2].

To compare our method to simplistic models for percussive note spectra, we appended these spectra to $\mathbf{B}_0$ obtained by the semantic initialization scheme (SI): We added either a constant spectrum (SI-1, similar to [6]) or N_{nh} mel-filters (SI-WB) (similar to [2], [7]). We also evaluated the combination of SI with additional random (absolute Gaussian distributed) components. This procedure was outperformed by SI-WB in our simulations and is therefore not considered here.

TABLE I
SDR, SIR AND SAR RESULTS IN dB FOR REFERENCE METHODS: ORACLE ESTIMATOR (ORACLE), SEMANTIC INITIALIZATION (SI), NON-NEGATIVE SVD (NNDSVD) AND SEMANTIC INITIALIZATION WITH ONE EXTRA COMPONENT CONSISTING OF A CONSTANT SPECTRUM (SI-1).

	Oracle	SI	NNDSVD	SI-1
SDR [dB]	20.35	11.75	10.52	11.88
SIR [dB]	28.01	18.63	17.26	18.74
SAR [dB]	22.16	14.17	12.87	14.38

Table I shows the averaged results over all 1770 mixtures for the following reference methods: The oracle estimator, the semantic initialization without (SI) and with an extra constant spectrum (SI-1) as well as initialization with the NNDSVD. Appending an additional constant spectrum to the harmonic notes considered in SI increases the separation performance significantly. The NNDSVD is clearly outperformed by the other methods.

Regarding ESI, non-harmonic frames were detected with either the spectral flatness (ESI-SF) or the correlation (ESI-C) based detection function described in Section IV-A with thresholding values given in Section V-A. The detected non-harmonic frames were separated with the NNDSVD into N_{nh} additional basis vectors as discussed in Section IV-B. We performed simulations for all $N_{\mathrm{nh}} \in \{2, 3, 5\}$ and for all threshold values and selected the thresholds for each detection function which resulted in the best SDR value averaged over all values of N_{nh}: This procedure yields $\alpha_{\mathrm{sf}} = 0.86$ for ESI-SF and $\alpha_{\mathrm{c}} = 1.05$ for ESI-C. We also evaluated the case where all time frames of the mixture spectrogram are detected (ESI-all) by setting $\alpha_{\mathrm{sf}} = 0$.

Table II shows results for initialization with additional mel-filters (SI-WB) as well as results for the proposed extended semantic initialization (ESI) for different numbers of additional basis vectors N_{nh} averaged over all 1770 mixtures:

- It becomes clear, that all extensions of the semantic initialization outperform the basic variant SI (see Table I). Our proposed data-driven extension improves the separation quality of SI significantly: ESI-SF yields an

[2]In [13], we evaluated SVD-based initialization schemes and compared them to SI on the same database as described in Section V-A. SI outperformed all other data-driven methods as well as a random initialization with absolute values of Gaussian random values for $\mathbf{B}_0$ and $\mathbf{G}_0$. Therefore, we consider here only SI as a reference and refer to [13] for the evaluation of the SVD-based methods.

TABLE II
SDR, SIR AND SAR RESULTS IN dB FOR SEMANTIC INITIALIZATION (SI) WITH N_{nh} ADDITIONAL WIDEBAND SPECTRA (SI-WB) AND THE PROPOSED EXTENDED SI (ESI) WITH SPECTRAL FLATNESS (ESI-SF AND $\alpha_{\mathrm{sf}} = 0.86$) AND CORRELATION (ESI-C AND $\alpha_{\mathrm{c}} = 1.05$) BASED DETECTION FUNCTIONS. ESI-ALL DENOTES THE CASE WHERE ALL FRAMES WERE DETECTED ($\alpha_{\mathrm{sf}} = 0$). REGARDING ESI, THE DETECTED FRAMES WERE SEPARATED WITH THE NNDSVD WITH DIFFERENT NUMBERS OF ADDITIONAL NON-HARMONIC COMPONENTS N_{nh}.

	SDR [dB]				SIR [dB]				SAR [dB]			
N_{nh}	SI-WB	ESI-SF	ESI-C	ESI-all	SI-WB	ESI-SF	ESI-C	ESI-all	SI-WB	ESI-SF	ESI-C	ESI-all
2	11.88	**11.91**	11.91	11.84	18.77	**18.83**	18.77	18.76	**14.37**	14.33	14.35	14.31
3	11.94	**12.00**	11.91	11.92	18.86	**18.89**	18.77	18.78	14.40	**14.44**	14.38	14.40
5	11.93	11.97	**12.02**	11.87	18.74	18.83	**18.92**	18.70	14.36	14.46	**14.50**	14.38

SDR increase of $+0.25\,\mathrm{dB}$ for $N_{\mathrm{nh}} = 3$, whereas ESI-C outperforms SI by $+0.27\,\mathrm{dB}$ for $N_{\mathrm{nh}} = 5$.

- Both variants of ESI outperform SI-WB for all N_{nh} as well, although the increase of the quality measures is only significant for ESI-C with $N_{\mathrm{nh}} = 5$: The SDR increases by $+0.09\,\mathrm{dB}$ and the SAR by $+0.14\,\mathrm{dB}$.
- Comparing ESI-SF and ESI-C, it can be noted that ESI-SF performs better for $N_{\mathrm{nh}} \in \{2, 3\}$ whereas ESI-C outperforms ESI-SF only for $N_{\mathrm{nh}} = 5$.
- Regarding ESI-all, all frames were detected which means that the complete spectrogram $\mathbf{X}$ was separated into N_{nh} components and appended to $\mathbf{B}_0$. This procedure results in worse separation results compared to detecting non-harmonic frames which clearly motivates the detection step.

C. Evaluation of Optimal Parameters

In the previous section, we optimized the parameters of our proposed initialization method ESI by simulating over different parameters regarding the threshold values α_{sf}, α_{c} and the number of separated components N_{nh}.

In this section, we used a different database for evaluation of the robustness of choice of parameters: The second database[3] consists of 26 recordings of harmonic and percussive instruments as well as vocals, that were extracted from the QUASI database [15]. Combining $M = 2$ sources results in a total of 325 mixtures. Oracle estimation yields an SDR value of $30.10\,\mathrm{dB}$, an SIR of $20.94\,\mathrm{dB}$ and an SAR of $22.10\,\mathrm{dB}$.

We evaluated SI, SI-WB and ESI with the parameters optimized on the first testset (refer to Table II): For ESI-SF, we chose the two parameters to be $(\alpha_{\mathrm{sf}}, N_{\mathrm{nh}}) = (0.86, 3)$, for ESI-C $(\alpha_{\mathrm{c}}, N_{\mathrm{nh}}) = (1.05, 5)$ and for SI-WB $N_{\mathrm{nh}} = 5$.

TABLE III
SDR, SIR AND SAR RESULTS IN dB FOR SI, SI-WB, ESI-SF AND ESI-C WITH OPTIMAL PARAMETERS.

	SI	SI-WB	ESI-SF	ESI-C
SDR [dB]	13.30	13.37	**13.47**	13.32
SIR [dB]	20.37	20.44	**20.65**	20.38
SAR [dB]	15.50	**15.59**	15.48	15.45

[3]We used the same database previously in [14].

Table III shows the corresponding separation results averaged over all mixtures:

- SI-WB gives better separation results than SI as it was observed already in Section V-B for the first database.
- ESI-SF clearly outperforms both SI-WB regarding SDR ($+0.13\,\mathrm{dB}$) and SIR ($+0.25\,\mathrm{dB}$) while SAR slightly decreases ($-0.09\,\mathrm{dB}$).
- ESI-C results in similar quality as SI for all measures.

This motivates the choice of the spectral flatness detection function as the optimal parameter choice seems to be more robust compared to ESI-C and ESI-C outperformes ESI-SF only by a small percentage on the first testset.

VI. SUMMARY

The goal of this preliminary work was to introduce a novel approach to initialize the basis matrix of the nonnegative matrix factorization by combining a semantic initialization scheme (SI) based on synthesized harmonic spectra and a data-driven initialization. Our method generates additional basis vectors by taking non-harmonic frames of the mixture into account. We proposed a novel detection method for non-harmonic frames and compared it to the spectral flatness measure.

We evaluated this scheme as initialization for a basic NMF-based blind source separation algorithm and compared it to other initialization approaches, e.g. SI and a combination of SI with wideband spectra. Our algorithm outperforms both approaches and is open for future work:

- More detailed evaluation of the non-harmonicity detection functions or possible combinations on data with ground truth.
- Analysis of the appended non-harmonic frames and the semantic piano note basis spectra to cancel out remaining harmonic components of the appended frames.
- Incorporation of a novel SVD-based initialization proposed in [13], which calculates the SVD on the complex output of the STFT directly and shows very good separation results; requires dealing with the phase information of the detected STFT frames.
- Computation of more general models such as a Gaussian mixture models at run-time instead of appending the separated detected frames directly.

REFERENCES

[1] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation: How models and side information can improve the separation of sound," *Signal Processing Magazine, IEEE*, vol. 31, no. 3, pp. 107–115, May 2014.

[2] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 4, pp. 1118–1133, May 2012.

[3] J. Fritsch and M. Plumbley, "Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013, pp. 888–891.

[4] C. Boutsidis and E. Gallopoulos, "SVD based initialization: A head start for nonnegative matrix factorization," *Pattern Recognition*, vol. 41, no. 4, pp. 1350–1362, April 2008.

[5] N. Bertin, R. Badeau, and G. Richard, "Blind signal decompositions for automatic transcription of polyphonic music: NMF and K-SVD on the benchmark," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 1, April 2007, pp. 65–68.

[6] S. Ewert and M. Müller, "Using score-informed constraints for nmf-based source separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, Mar 2012, pp. 129–132.

[7] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 3, pp. 528–537, March 2010.

[8] M. Spiertz, *Underdetermined Blind Source Separation for Audio Signals*, ser. Aachen Series on Multimedia and Communications Engineering. Aachen: Shaker Verlag, July 2012, vol. 10.

[9] V. Tan and C. Fevotte, "Automatic relevance determination in nonnegative matrix factorization with the beta-divergence," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 7, pp. 1592–1605, July 2013.

[10] N. Madhu, "Note on measures for spectral flatness," *Electronics Letters*, vol. 45, no. 23, pp. 1195–1196, November 2009.

[11] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, pp. 1462–1469, Jul 2006.

[12] E. Vincent, R. Gribonval, and M. D. Plumbley, "Oracle estimators for the benchmarking of source separation algorithms," *Signal Processing*, vol. 87, no. 8, pp. 1933–1950, Dec 2007.

[13] J. M. Becker, M. Menzel, and C. Rohlfing, "Complex SVD initialization for NMF source separation on audio spectrograms," in *DAGA 2015*, Nürnberg, Germany, 2015.

[14] J. Becker and C. Rohlfing, "Custom sized non-negative matrix factor deconvolution for sound source separation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP '14*. Florence, Italy: IEEE, Piscataway, May 2014.

[15] "QUASI database - a musical audio signal database for source separation," http://www.tsi.telecom-paristech.fr/aao/en/2012/03/12/quasi/.