Guitar String Separation Using Non-Negative Matrix Factorization and Factor Deconvolution

Dalia Senvaityte
Centre for Digital Music
Queen Mary University of London
London, UK
d.senvaityte@qmul.ac.uk

Johan Pauwels
Centre for Digital Music
Queen Mary University of London
London, UK
j.pauwels@qmul.ac.uk

Mark Sandler
Centre for Digital Music
Queen Mary University of London
London, UK
mark.sandler@qmul.ac.uk

ABSTRACT

Guitar string separation is a novel and complicated task. Guitar notes are not pure steady-state signals, hence, we hypothesize that neither Non-Negative Matrix Factorization (NMF) nor Non-Negative Matrix Factor Deconvolution (NMFD) are optimal for separating them. Therefore, we separate steady-state and transient parts using Harmonic-Percussive Separation (HPS) as a preprocessing step. Then, we use NMF for factorizing the harmonic part and NMFD for deconvolving the percussive part. We make use of a hexaphonic guitar dataset which allows for objective evaluation. In addition, we compare several types of time-frequency mask and introduce an intuitive way to combine a binary mask with a ratio mask. We show that the HPS mask type has an effect on source estimation. Our proposed method achieved results comparable to NMF without HPS. Finally, we show that the optimal mask at the final separation stage depends on the estimation algorithm.

CCS CONCEPTS

• **Computing methodologies** → **Source separation**; **Non-negative matrix factorization**; • **Applied computing** → **Sound and music computing**;

KEYWORDS

source separation, non-negative matrix factorization & deconvolution, time-frequency masks, harmonic-percussive separation, median filtering

AM'19, September 18–20, 2019, Nottingham, United Kingdom

ACM ISBN 978-1-4503-7297-8/19/09…$15.00
https://doi.org/10.1145/3356590.3356628

ACM Reference Format:
Dalia Senvaityte, Johan Pauwels, and Mark Sandler. 2019. Guitar String Separation Using Non-Negative Matrix Factorization and Factor Deconvolution. In *Audio Mostly (AM'19), September 18–20, 2019, Nottingham, United Kingdom.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3356590.3356628

1 INTRODUCTION

A guitar signal is usually captured with either a pickup or a microphone. These ways all the strings are recorded into a single track. Conversely, capturing each string individually can be achieved by fitting a hexaphonic pickup to the guitar. However, this is uncommon since it requires additional hardware.

One of the reasons why having each string in a separate track is beneficial is that non-linear processing causes intermodulation distortion when applied to signals containing more than one frequency component [8]. Although intermodulation distortion is desirable at times, it is generally unpleasant because it produces inharmonic partials. Non-linear processing applied to guitar notes that do not overlap in time does not produce obtrusive inharmonic partials.

Separating a guitar signal into a signal per string is a very difficult task compared to a more commonly tackled problem - separating different instruments. While guitar processing has received much attention [1, 13], little research is focused on guitar string separation. Previous work examined a special case where a non-standard tuning was used so that the strings were not harmonically related and only open strings were considered [15].

While Non-Negative Matrix Factorization (NMF) and its variations have been used for separation of both harmonic and percussive instruments [9, 10], it has been claimed that preserving time evolution of the spectra is especially important to non-stationary signals [11, 16]. Although the guitar is generally categorized as a quasi-harmonic instrument, plucked strings produce sharp transients in addition to the steady-state content (for convenience, we refer to these as percussive and harmonic parts, respectively). Based on the notion that it is important to preserve time evolution of the spectra, we hypothesize that Non-Negative Matrix Factor Deconvolution (NMFD) is more suited for the separation of

transient source elements like guitar onsets. However, since guitar notes vary greatly in length, it is not straightforward to apply NMFD to guitar string separation. For this reason, we first separate the guitar recordings into harmonic and percussive parts. Then, we factorize each part independently - NMF is applied to the harmonic part while NMFD is applied to the percussive part. Finally, the factorized harmonic and percussive parts are sorted into strings.

The rest of the paper is structured as follows. In Sec. 2 & 3 we discuss related work: NMF & NMFD, harmonic-percussive separation (HPS) and time-frequency (TF) masking. Our experiments are described in Sec. 4 while their results are discussed in Sec. 5. Finally, we draw the conclusions and suggest further work in Sec. 6.

2 NMF & NMFD

NMF is defined as follows [12, 16]:

$$V \approx \widetilde{V} = W \cdot H \tag{1}$$

where $V \in \mathbb{R} \geq 0^{M \times N}$ is the original mixture, $\widetilde{V} \in \mathbb{R} \geq 0^{M \times N}$ is the estimated mixture, $W \in \mathbb{R} \geq 0^{M \times R}$ contains bases and $H \in \mathbb{R} \geq 0^{R \times N}$ contains activations (M, N & R are the number of frequency bins, the number of time frames and decomposition rank, respectively), $\cdot$ is a matrix product. It can represent each source with several basis spectra but it does not make use of spectro-temporal information.

Smaragdis proposed an NMF extension which makes use of time-wise patterns - NMFD [16]. It is defined as follows:

$$V \approx \widetilde{V} = \sum_{t=0}^{T-1} W_t \cdot \overset{t\rightarrow}{H}, \tag{2}$$

where $W_t \in \mathbb{R} \geq 0^{M \times R}$ are bases, $\overset{t\rightarrow}{\cdot}$ is a shift operation where matrix columns are shifted right with resulting empty columns on the left being filled with zeros, and T is convolution depth. Kullback-Leibler (KL) divergence is often used to find the factors.

Kwon et al. showed that NMF results are highly dependent on initialisation of W due to the problem being non-convex [10] when W and H are updated concurrently [2]. Hence, supervised/informed NMF based methods are common. Often the bases are not updated.

3 HPS & TF MASKS

FitzGerald applied Median Filtering to magnitude spectrograms of audio signals to achieve HPS [6]. Based on the notion that harmonic and percussive content forms orthogonal structures, filtering a spectrogram along the time axis with a median kernel results in one with enhanced harmonic components and vice versa. The resulting spectrograms can be used to create TF masks for extracting harmonic and percussive components from the original spectrogram.

TF masking is required not only for HPS but also for recovering sources from a mixture since sources estimated with NMF or NMFD often contain artefacts. In the following section, we discuss three types of mask - binary, ratio and sigmoid - and introduce an intuitive way to combine a binary mask with a ratio mask.

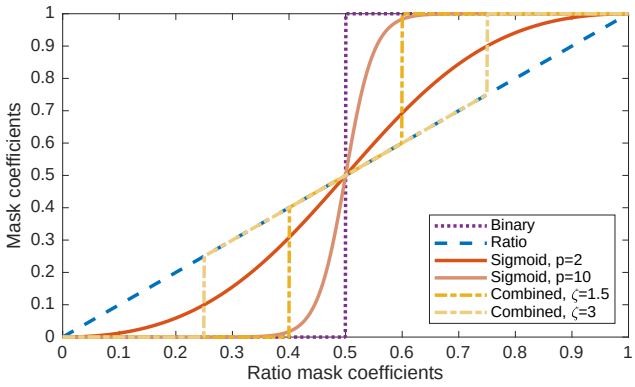

Figure 1: Mapping from ratio mask coefficients to binary, combined, sigmoid and ratio mask coefficients assuming a mixture of two sources.

Binary Masks

A Binary Mask (BM), is defined as follows:

$$M_{ij}^{B} = \begin{cases} 1 & \text{if} \quad X_{ij} > Y_{ij} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where X and Y are magnitude spectrograms of the target and the interfering signals, respectively, while i and j correspond to time and frequency indices. According to Yılmaz and Rickard "perfect demixing via binary TF masks is possible provided the TF representations of the sources do not overlap" [22]. Nevertheless, BMs are used for sources that overlap more often than not. In the situations where they do, BMs are likely to produce artefacts.

Sigmoid and Ratio Masks

Another type of TF mask is a Sigmoid Mask (SM) [7]. As opposed to a BM, an SM is a continuous mask since its coefficients belong to $[0, 1]$. For more than one interfering source, it is defined as follows:

$$M_{ij}^{S} = \frac{X_{ij}^{p}}{X_{ij}^{p} + \sum_{Y} Y_{ij}^{p}} \tag{4}$$

where $p \in [1; \infty)$, $p = 1$ is a special case - a Ratio Mask (RM). It was defined for energy spectrograms by Srinivasan et al. [17, eq. 1]. An SM can be considered as a compromise between an RM and a BM. The lower the p the closer the mask is to an RM. Conversely, the higher the p the closer it is to a BM, see Fig 1.

Ratio and Binary Masks Combined

On average, continuous masks outperform BMs in terms of Signal to Distortion Ratio (SDR) [20]. This is supported by perceptual evaluation results in [18] which showed that, out of three types of mask - binary, ratio and sigmoid - applied to energy spectrograms, BMs were the least appealing. Nevertheless, a BM is useful because it offers more interference suppression. However, it is too harsh when the target and interfering sources contribute to the magnitude of a particular spectrogram bin to a similar extent.

While an SM interpolates between an RM and a BM, it does not allow for cases where most coefficients are very close to a BM while the rest still make up a significant part and are very close to an RM and vice versa. Hence, it may be beneficial to create a mask with the coefficients equal to RM coefficients when $\frac{X_{ij}}{Y_{ij}} \approx 1$ and BM coefficients when $\frac{X_{ij}}{Y_{ij}} \gg 1$ or $\frac{X_{ij}}{Y_{ij}} \ll 1$. This can be achieved by combining an RM and a BM. We introduce a convenient way of doing that - a Combined Mask (CM) - defined as follows:

$$M_{ij}^{C} = \begin{cases} 1 & \text{if } \frac{X_{ij}}{Y_{ij}} > \zeta \\ \frac{X_{ij}}{X_{ij}+Y_{ij}} & \text{if } \frac{1}{\zeta} \leq \frac{X_{ij}}{Y_{ij}} \leq \zeta \\ 0 & \text{if } \frac{X_{ij}}{Y_{ij}} < \frac{1}{\zeta} \end{cases} \quad (5)$$

where the parameter ζ allows for intuitively balancing between an RM and a BM. $\zeta \in [1; \infty)$, $\zeta = 1$ corresponds to a BM (except for the bins where $\frac{X_{ij}}{Y_{ij}} = 1$) while $\zeta \to \infty$ corresponds to an RM.

4 EXPERIMENTS

Dataset

The GuitarSet dataset [21] includes 180 hexaphonic accompaniment tracks, i.e. chords, played on an acoustic guitar sampled at 44.1kHz.

Each pickup in a hexaphonic pickup captures not only its target string but also the neighbouring strings albeit with lower amplitude. These tracks are not ideal for use as ground truth because the goal is to estimate the signals produced by each string individually. Conveniently, the dataset includes tracks de-bleeded using Kernel Additive Modeling for Interference Reduction (KAMIR) [14]. These tracks were used for creating ground truth for objective evaluation.

This dataset also includes note level annotations. It is highly imbalanced in terms of note occurrence, i.e. some notes appear much more frequently than others with the number of occurrences ranging from 1 to 2485 per note. In order to decrease this imbalance, we considered only the notes that appear at least a given number of times (= 56) in the training set. In addition, 0.2% of the accompaniment annotations are clearly incorrect since they are lower than the corresponding open strings. We excluded these notes. As a result, a list of 32 MIDI notes (40-73, excluding 42 & 43) was compiled.

The accompaniment tracks were split into a training set (148 tracks) and a testing set (32 tracks) so that all notes in the testing set were sufficiently represented in the training set i.e. so that each note appeared in the training set at least the given number of times. The resulting testing set contained 6876 note instances. See the accompanying GitHub repository for details[1].

Experiment Flow

Figure 2 shows an overview of the experiment flow. Firstly, linear mixtures were created by summing the channels of the hexaphonic recordings from the testing set (32 mixtures $\approx$ 14 minutes of audio).

Then, both the resulting mixtures and the note instances from the training set were separated into harmonic and percussive parts using a Median Filtering implementation from the TSM toolbox by Driedger and Müller [4]. The HPS parameters were a 1024 samples long Hann window with 75% overlap for Short-Time Fourier Transform (STFT) along with filter lengths of 0.2 seconds for the harmonic part and 500Hz for the percussive part. The masks described in Sec. 3 were used to extract the harmonic and the percussive parts.

Next, NMFD was applied to the percussive parts while NMF was applied to the harmonic parts. Both were initialised with the corresponding parts of the note instances as follows. 56 instances of each MIDI note were taken from the training set and split into harmonic and percussive parts. Together with the harmonic and percussive parts of the mixtures, they were transformed using STFT with a 1024 samples long Hamming window and 50% overlap. Then, the 56 instances of each note were averaged into one (harmonic and percussive parts separately). 6 NMF bases per MIDI note were learnt by factorizing the averaged percussive parts of each note instance. Although the number of bases per averaged note instance could have been reduced to 1, multiple bases were learnt in order to represent the TF non-stationarity more accurately. The averaged percussive parts of note instances were used as bases for NMFD.

Updating both the activations and the bases did not lead to satisfactory results, hence, only the former were updated for both NMF and NMFD. The activations were normalized at each iteration.

The resulting factors of the harmonic parts were then added to the corresponding factors of the percussive parts. Finally, the separated notes were sorted into strings relying on the annotations.

[1] https://github.com/daliasen/GuitarStringSeparation-MF-NMF-NMFD

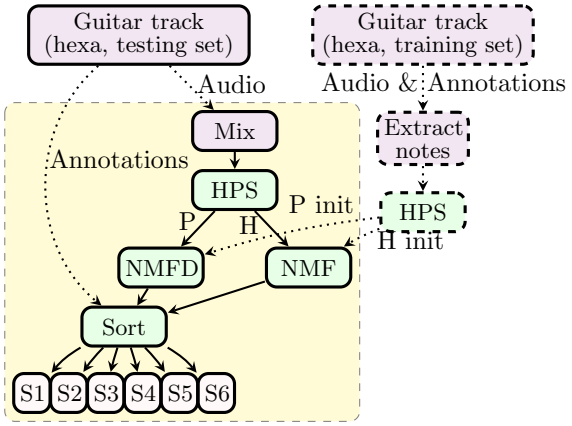

Figure 2: Guitar string separation flow. P are percussive while H are harmonic parts of the mixtures. P init are percussive while H init are harmonic parts of the note instances used for the initialisation of the bases. Dashed boxes show the training procedure. S1 - S6 are the separated strings.

5 EVALUATION, RESULTS AND DISCUSSION

Three metrics from the BSS Eval toolbox were used for objective evaluation: Source to Distortion Ratio (SDR), Source to Interferences Ratio (SIR) and Sources to Artifacts Ratio (SAR) [19].

The Performance of TF Masks

Masks created using the ground truth were used to extract the sources from the mixtures in order to evaluate the mask performance independently of estimation. As expected, BMs gave higher mean SIR but lower SAR and SDR compared to continuous masks ($p = 1, 2$) while the performance of CMs ($\zeta = 2$) was closer to that of the BMs in terms of SIR but closer to the performance of the continuous masks in terms of SAR and SDR (see Tab. 1). SMs gave better results than RMs. However, they performed worse than the CMs in terms of SIR. This illustrates that a CM provides a good compromise between interference suppression and distortion/artefacts.

Table 1: The results of masking the mixtures with ground truth masks (averaged over strings and test tracks).

	SDR, dB	SIR, dB	SAR, dB
Binary	6.2 ±2.5	**23.2** ±4.4	6.4 ±2.5
Ratio	6.3 ±2.5	16.2 ±3.8	7.2 ±2.4
Sigmoid, $p = 2$	**7.0** ±2.4	20.1 ±3.7	**7.4** ±2.5
Combined, $\zeta = 2$	6.9 ±2.4	21.9 ±4.1	7.2 ±2.5

Separation Evaluation

An experiment where the HPS was omitted and only NMF was used was carried out for comparison. Surprisingly, NMF gave the best overall results. It slightly outperformed our proposed method in terms of SIR & SDR before the final masking (see Tab. 2).

Focusing on the effects of the different mask types used for HPS in our proposed method, as anticipated, BMs outperformed continuous masks and CMs with regard to SIR while they were outperformed by the continuous masks and the CMs with regard to SDR & SAR. The CMs and SMs gave similar results but the CMs gave higher SIR than the RMs while the SMs gave similar SIR to the RMs. Both types of mask performed worse than the RMs and better than the BMs in terms of SDR & SAR.

When applied to the estimated signals BMs gave higher SIR but at the expense of SDR & SAR as expected. Note that when the BMs were used for HPS the improvement in SIR was much less significant. As anticipated, RMs gave the best results in terms of SDR & SAR but at the expense of SIR. When the RMs were used for HPS the improvement in SDR & SAR was much less significant.

For SMs, SIR was always higher than that of the RMs but always lower than that of the BMs while SDR & SAR were always higher than that of the BMs but always lower than that of the RMs. Although we anticipated similar trends for CMs, all three metrics were decreased in most cases. Moreover, the SMs gave better results than the CMs in all cases but one. The latter was when no HPS was used. In this case, the CMs provided an alternative compromise between interference suppression and distortion/artefacts. Overall, these results showed that the optimality of the final stage masking is dependent on the source estimation algorithm.

6 CONCLUSIONS AND FUTURE WORK

We introduced an approach to guitar string separation. While it gave reasonable results it did not outperform the basic NMF based method. This suggests that, in this scenario, NMF may be more adaptive than NMFD since in NMF each note is constructed from multiple bases. Therefore, it may be better at representing different articulations of the same note. Updating the NMFD bases with constraints for adapting them to each track may improve the results. Deep Learning based approaches could also be considered, however, the amount of available training data is limited.

In addition, our proposed method has many parameters that could be explored further. We showed that the mask type used for HPS stage has an effect on the source estimation quality. Hence, different HPS methods, for instance, [3, 5], should be investigated. We showed that the mask type at

Table 2: The final stage masking results (averaged over strings and test tracks), $p = 2$, $\zeta = 2$.

HPS mask	Final mask	SDR, dB	SIR, dB	SAR, dB
None (NMF)	Estimated	3.9 ±2.7	14.8 ±4.8	4.8 ±2.7
	Binary	3.1 ±2.9	**16.2** ±5.0	3.8 ±3.1
	Ratio	**4.0** ±2.7	13.6 ±4.4	**5.1** ±2.9
	Sigmoid	3.9 ±2.8	15.1 ±4.7	4.8 ±3.0
	Combined	3.8 ±2.9	15.9 ±4.9	4.5 ±3.1
Binary	Estimated	2.5 ±3.3	14.6 ±5.1	3.3 ±3.1
	Binary	2.2 ±2.9	**14.7** ±4.9	3.0 ±3.1
	Ratio	**3.0** ±2.7	12.8 ±4.4	**4.2** ±2.9
	Sigmoid	2.9 ±2.8	13.8 ±4.6	3.9 ±3.0
	Combined	2.2 ±3.1	12.7 ±4.1	3.2 ±3.1
Ratio	Estimated	**3.8** ±2.6	14.1 ±4.9	4.8 ±2.4
	Binary	3.0 ±2.9	**15.8** ±5.2	3.7 ±3.1
	Ratio	**3.8** ±2.7	13.5 ±4.5	**4.9** ±2.9
	Sigmoid	3.7 ±2.7	14.7 ±4.9	4.6 ±3.0
	Combined	3.5 ±2.7	14.1 ±4.5	4.5 ±2.7
Sigmoid	Estimated	3.2 ±3.0	14.1 ±5.0	4.2 ±2.7
	Binary	2.8 ±2.9	**15.5** ±5.1	3.5 ±3.1
	Ratio	**3.6** ±2.7	13.4 ±4.4	**4.8** ±2.9
	Sigmoid	3.5 ±2.8	14.5 ±4.8	4.4 ±3.0
	Combined	2.7 ±3.0	12.3 ±4.3	3.9 ±2.8
Combined	Estimated	2.9 ±3.1	14.3 ±5.1	3.9 ±2.8
	Binary	2.7 ±2.9	**15.4** ±5.1	3.4 ±3.1
	Ratio	**3.5** ±2.7	13.3 ±4.4	**4.7** ±2.9
	Sigmoid	3.4 ±2.8	14.5 ±4.8	4.4 ±3.0
	Combined	2.4 ±3.1	12.2 ±4.5	3.6 ±2.8

the final masking stage needs to be chosen to optimise for a given source estimation algorithm.

ACKNOWLEDGMENTS

This work has been funded by the UK Engineering and Physical Sciences Research Council (EPSRC) grant EP/L019981/1.

REFERENCES

[1] Isabel Barbancho, Lorenzo J. Tardon, Simone Sammartino, and Ana M. Barbancho. 2012. Inharmonicity-Based Method for the Automatic Generation of Guitar Tablature. *IEEE Transactions on Audio, Speech, and Language Processing* 20, 6 (Aug. 2012), 1857–1868.

[2] Andrzej Cichocki, Rafal Zdunek, Anh Huy Phan, and Shun-Ichi Amari. 2009. *Nonnegative Matrix and Tensor Factorizations - Applications to Exploratory Multi-way Data Analysis and Blind Source Separation.* John Wiley and Sons, Ltd, Chichester.

[3] Derry Fitzgerald, Antoine Liutkus, Zafar Rafii, Bryan Pardo, and Laurent Daudet. 2014. Harmonic/Percussive Separation Using Kernel Additive Modelling. In *25th IET Irish Signals & Systems Conference 2014 and 2014 China-Ireland International Conference on Information and Communities Technologies.*

[4] Jonathan Driedger and Meinard Müller. 2014. TSM Toolbox: MATLAB Implementations of Time-Scale Modification Algorithms. In *Proc. of the 17th Int. Conference on Digital Audio Effects.*

[5] Konstantinos Drossos, Paul Magron, Stylianos Ioannis Mimilakis, and Tuomas Virtanen. 2018. Harmonic-Percussive Source Separation with Deep Neural Networks and Phase Recovery. In *16th International Workshop on Acoustic Signal Enhancement.*

[6] Derry FitzGerald. 2010. Harmonic/Percussive Separation Using Median Filtering. In *Proc. of the 13th Int. Conference on Digital Audio Effects.*

[7] Emad M. Grais and Hakan Erdogan. 2011. Single channel speech music separation using nonnegative matrix factorization and spectral masks. In *2011 17th International Conference on Digital Signal Processing.*

[8] Joshua D. Reiss and Andrew. P. McPherson. 2015. *Audio Effects : Theory, Implementation and Application.* CRC Press, Boca Raton, FL.

[9] Elias Kokkinis, Alexandros Tsilfidis, Thanos Kostis, and Kostas Karamitas. 2013. A New DSP Tool for Drum Leakage Suppression. In *Audio Engineering Society Convention 135.*

[10] Kisoo Kwon, Jong Won Shin, Inkyu Choi, Hyung Yong Kim, and Nam Soo Kim. 2016. Incremental approach to NMF basis estimation for audio source separation. In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference.*

[11] Clément Laroche, Helene Papadopoulos, Matthieu Kowalski, and Gaël Richard. 2017. Drum extraction in single channel audio signals using multi-layer non negative matrix factor deconvolution. In *IEEE International Conference on Acoustics, Speech and Signal Processing.*

[12] Daniel D. Lee and H. Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401 (Oct. 1999).

[13] Zulfadhli Mohamad, Simon Dixon, and Christopher Harte. 2017. Pickup position and plucking point estimation on an electric guitar. In *IEEE International Conference on Acoustics, Speech and Signal Processing.*

[14] Thomas Prätzlich, Rachel M. Bittner, Antoine Liutkus, and Meinard Müller. 2015. Kernel Additive Modeling for interference reduction in multi-channel music recordings. In *IEEE International Conference on Acoustics, Speech and Signal Processing.*

[15] David Ramsay, Ted Burke, Dan Barry, and Eugene Coyle. 2011. A Novel Fourier Approach to Guitar String Separation. In *IET Irish Signals and Systems Conference.*

[16] Paris Smaragdis. 2004. Non-negative Matrix Factor Deconvolution; Extracation of Multiple Sound Sources from Monophonic Inputs. In *International Congress on Independent Component Analysis and Blind Signal Separation.*

[17] Soundararajan Srinivasan, Nicoleta Roman, and DeLiang Wang. 2006. Binary and ratio time-frequency masks for robust speech recognition. *Speech Communication* 48 (2006), 1486–1501.

[18] Tobias Stokes. 2015. *Improving the perceptual quality of single-channel blind audio source separation.* PhD. University of Surrey, Guildford, UK.

[19] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. 2006. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech and Language Processing* 14, 4 (2006), 1462–1469.

[20] Emmanuel Vincent, Rémi Gribonval, and Mark D. Plumbley. 2007. Oracle estimators for the benchmarking of source separation algorithms. *Signal Processing* 87, 8 (Aug. 2007), 1933–1950.

[21] Qingyang Xi, Rachel M Bittner, Johan Pauwels, Xuzhou Ye, and Juan P Bello. 2018. GuitarSet: A dataset for Guitar Transcription. In *International Society for Music Information Retrieval.*

[22] Özgür Yılmaz and Scott Rickard. 2004. Blind Separation of Speech Mixtures via Time-Frequency Masking. *IEEE Transactions on Signal Processing* 52, 7 (July 2004), 1830–1847.