EXTRACTING SINGING VOICE FROM MUSIC RECORDINGS BY CASCADING AUDIO DECOMPOSITION TECHNIQUES

Jonathan Driedger and Meinard Müller

International Audio Laboratories Erlangen

ABSTRACT

The problem of extracting singing voice from music recordings has received increasing research interest in recent years. Many proposed decomposition techniques are based on one of the following two strategies. The first approach is to directly decompose a given music recording into one component for the singing voice and one for the accompaniment by exploiting knowledge about specific characteristics of singing voice. Procedures following the second approach disassemble the recording into a large set of fine-grained components, which are classified and reassembled afterwards to yield the desired source estimates. In this paper, we propose a novel approach that combines the strengths of both strategies. We first apply different audio decomposition techniques in a cascaded fashion to disassemble the music recording into a set of mid-level components. This decomposition is fine enough to model various characteristics of singing voice, but coarse enough to keep an explicit semantic meaning of the components. These properties allow us to directly reassemble the singing voice and the accompaniment from the components. Our objective and subjective evaluations show that this strategy can compete with state-of-the-art singing voice separation algorithms and yields perceptually appealing results.

Index Terms— Singing voice extraction, audio decomposition, music processing.

1. INTRODUCTION

In recent years a lot of effort has been put into the development of algorithms for extracting singing voice from music recordings. This interest emerged from both scientific curiosity for better understanding the characteristics of human singing [1] as well as the commercial need for such techniques in applications such as music remixing, remastering, and production [2]. There exists a large variety of algorithmic approaches to this problem. Although a classification of singing voice extraction methods into specific categories is difficult, many of them tend to follow either one of two basic strategies, see Figure 1. Approaches employing a *direct decomposition* strategy aim to decompose a given audio recording directly into one component that contains the singing voice and one that contains the accompaniment. These methods are usually based on some specific characteristic of the singing voice. Examples for such characteristics are the clear and strong harmonic structure of singing voice [3], its spectral sparseness [4], the high variance of singing voice in contrast to the repeating structure of accompanying music [5, 6, 7], the presence of vibrato and glissando in singing voice [8], or the occurrence of specific spectral patterns [9]. Rather than explicitly extracting the singing voice, these decomposition procedures are designed to extract the specific characteristic from a given recording. This usually goes along with extracting large portions of the singing voice. While the resulting decompositions have an explicit semantic meaning, the procedures are usually not designed to also flexibly incorporate knowledge about additional characteristics.

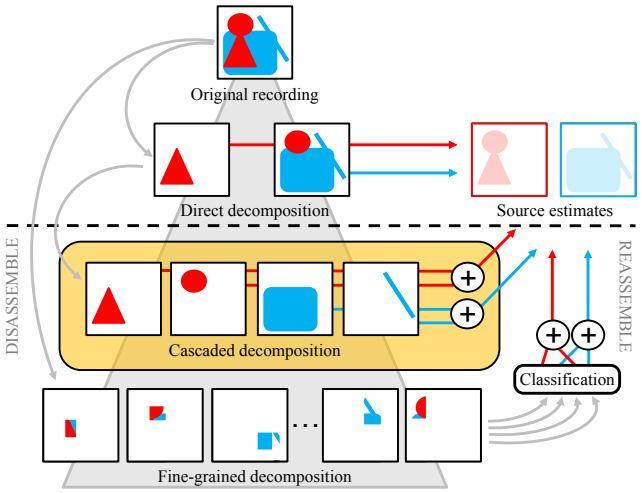

Fig. 1. Schematic overview of different decomposition strategies. Colors encode sources, shapes depict characteristics.

Approaches which follow a *disassemble and reassemble* strategy first decompose the given music recording into a large set of fine-grained components. Afterwards, all components are classified to belong to either the singing voice or the accompaniment and reassembled accordingly. Common techniques to perform this kind of decomposition are Non-negative Matrix Factorization (NMF) and related formulations [10, 11, 12, 13, 14] or time-frequency decompositions [15]. Although the fine-grained decomposition yields a high degree of flexibility when reassembling the sources, the correct classification of the components constitutes a challenging problem. Depending on the chosen decomposition technique, the components may not even have a semantic interpretation anymore. The classification can either be done in an unsupervised fashion [13, 15], in a supervised way [11, 14, 16], or it can be derived from the decomposition process itself [10, 12, 13].

Combining ideas from both strategies, we propose in this paper a novel approach for singing voice extraction. Inspired by the disassemble and reassemble strategy, a given music signal is first split into a set of components. However, contrary to other procedures following this methodology, we decompose the recording on a coarser granularity level by cascading different direct decomposition procedures, see Figure 1. This yields several advantages. On the one hand, the resulting *mid-level components* have an explicit semantic meaning, inherited directly from the sequence of applied decomposition procedures. On the other hand, the cascaded decomposition is flexible enough to account for various characteristics of the singing voice

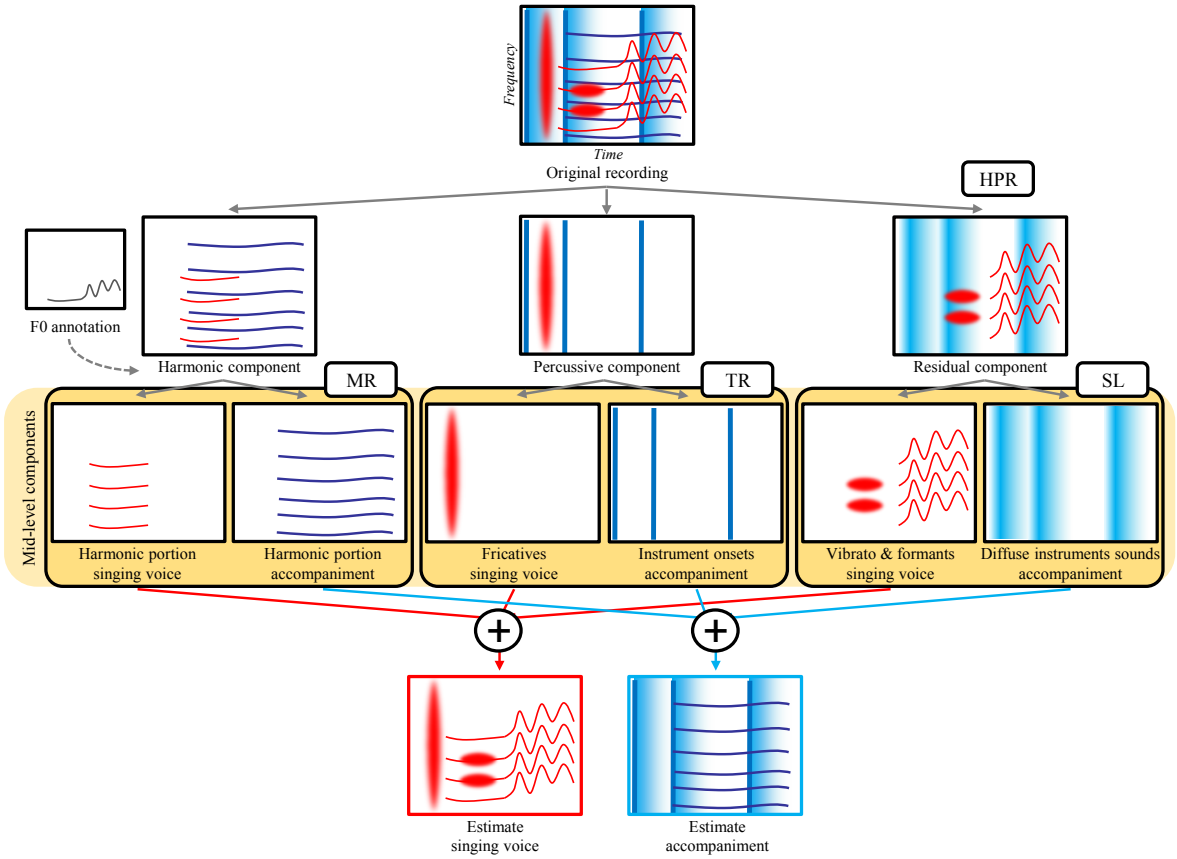

Fig. 2. Overview of our disassemble and reassemble approach. Audio material is visualized as stylized spectrograms. Spectral portions of the singing voice are depicted in red, spectral portions of the accompaniment in blue.

and the accompaniment. Mid-level components hold for example the fricatives in the singing voice or the harmonic portion of the accompaniment. Because of the explicit interpretation of all mid-level components, their classification can be done on the semantic level based on the characteristics which are assumed to be captured by the components. Estimates of the singing voice and the accompaniment are then reassembled by adding up the respective components. A review of the proposed procedure is given in Section 2, see also Figure 2. Our objective and subjective evaluation is then described in Section 3 and we conclude the paper in Section 4.

2. PROPOSED METHOD

In this section we briefly review the decomposition techniques which we use in order to disassemble a given music recording into mid-level components. As shown in Figure 2, we start by applying a *harmonic-percussive-residual* (HPR) decomposition procedure. The resulting three components are further processed with a fundamental frequency informed *melody-residual* (MR) decomposition procedure, a *transient-residual* (TR) decomposition procedure, and a *sparse-low rank* (SL) decomposition procedure, respectively. Finally, the resulting components are reassembled to form the estimates of the singing voice and the accompaniment.

2.1. Harmonic-Percussive-Residual Decomposition (HPR)

This procedure decomposes an audio signal into a harmonic component that corresponds to horizontal spectral structures, a percussive component that corresponds to vertical spectral structures, and a residual component that captures sounds whose spectral structure is neither clearly horizontal nor vertical. Applied to a music recording with singing voice, the harmonic component usually contains most of the tonal portion of the singing voice as well as of the accompaniment. The percussive component holds sounds like fricatives in the sung lyrics, drum hits, or pronounced instrument onsets. In the residual component, one can often hear strong vibrato passages and sounds resulting from strong formants in the singing voice as well as noise-like instrument sounds as for example the decaying sound of a snare drum or an open hi-hat. For a detailed description and further information about the HPR decomposition method, we refer to [17].

2.2. Melody-Residual Decomposition (MR)

This procedure, initially proposed in [3] as a singing voice extraction technique on its own, is based on the observation that singing voice usually has a clear and strong harmonic structure. Given the fundamental frequency track of the sung melody, the desired source is extracted from a spectrogram of the music recording by considering all time-frequency instances that correspond to the fundamental frequency track or one of its harmonics. This estimate, which may still contain portions of spectrally overlapping sources, is then refined using NMF-based techniques, see [3]. By subtracting the resulting

melody component from the original recording, the accompaniment can be estimated as well. Applied to the harmonic component of the previous decomposition step, the resulting mid-level components hold the harmonic portion of the singing voice and the harmonic portion of the accompaniment, respectively.

2.3. Transient-Residual Decomposition (TR)

Initially designed for the extraction of transient noise from speech signals in [18], the core observation of this decomposition procedure is that transients produced by a specific instrument, like for example a drum, usually occur many times in a given recording. In a spectral representation, these transients are similar to each other while the spectral structure of speech is usually more diverse. Given a spectrogram representation of an audio recording, spectral frames of similar structure are identified as transient candidates, and a prototype transient is computed by averaging over all candidates. This prototype is then subtracted from the spectrogram at the identified transient positions, yielding the spectrogram of the residual component. The transient component is then computed by subtracting the residual component from the original recording. Since a music recording usually contains different kinds of frequently occurring transients, as for example the ones produced by the bass drum, the snare, and the hi-hat, we apply this technique to the percussive component of the HPR decomposition in an iterative fashion. Hereby, we decompose it into one mid-level component that typically holds the instrument transients as well as a second one typically holding fricatives of the singing voice.

2.4. Sparse-Low Rank Decomposition (SL)

This decomposition method, which has been used in the context of singing voice extraction in [4], is based on Robust Principle Component Analysis [19]. This technique splits a given matrix into the sum of two matrices, one being sparse and the other having a low rank. In our cascade, we apply this procedure to the spectrogram of the HPR decomposition's residual component. In this component, the contained formant and vibrato sounds tend to have a sparse spectral structure. The diffuse instrument sounds usually occur many times in a spectrally similar way and can be represented by a spectrogram having a low rank. Therefore, the sparse-low rank decomposition technique is well-suited to split the residual component of the first decomposition stage into a mid-level component that contains formant and vibrato sounds of the singing voice as well as a second mid-level component that contains diffuse instrument sounds.

3. EVALUATION

We evaluate our proposed procedure in three ways. First, we compare the performance of our approach with state-of-the-art singing voice extraction methods on a standard dataset by means of objective source separation evaluation measures. Then, we discuss the results of a subjective listening experiment, which shows that the singing voice estimates of our procedure have a high perceptual quality. Finally, we also provide an accompanying website for this paper where one can find all audio files used in the objective and subjective evaluation as well as many further audio examples.

3.1. Objective Evaluation

To evaluate our proposed method in an objective way and to compare it to other procedures, we applied it to the well-known dataset of the *Signal Separation Evaluation Campaign* (SiSEC) [20, 21]. This dataset consists of five pop music multitrack recordings. For algorithms that participated in previous rounds of the campaign, separation results along with objective evaluation measures are available online at [20]. The reported evaluation measures were computed using the *Perceptual Evaluation methods for Audio Source Separation* toolkit (PEASS) [22] and consist of the Signal to Distortion Ratio (SDR), the source Image to Spatial distortion Ratio (ISR), the Signal to Interference Ratio (SIR), the Signal to Artifacts Ratio (SAR), the Overall Perceptual Score (OPS), the Target-related Perceptual Score (TPS), the Interference-related Perceptual Score (IPS), and the Artifacts-related Perceptual Score (APS). To examine the influence of the fundamental frequency track that is needed for the MR decomposition step (see Section 2.2), we applied our cascaded decomposition procedure to all recordings in the dataset twice: Once "informed" (CD-I) with a manually annotated fundamental frequency track, and once "blind" (CD-B) with a track automatically extracted using the MELODIA vamp plug-in [23]. Table 1 shows the computed evaluation measures for our singing voice estimates together with those of several state-of-the-art singing voice extraction algorithms. The measures for the oracle ideal binary mask (IBM), the REpeating Pattern Extraction Technique (REPET) [5], as well as the Voiced+Unvoiced Instantaneous Mixture Model technique (VUIMM) [10] were taken directly from [20]. The results for the Sparse-Low rank decomposition (SL) were computed by ourselves. REPET and SL are representatives of direct decomposition approaches, while VUIMM employs a disassemble and reassemble strategy. The first observation is that our proposed procedure yields evaluation measures in the same order of magnitude as REPET and VUIMM. The IBM can be seen as an upper limit for separation quality when using binary masking. The performance of SL which is also part of our decomposition cascade, falls slightly behind. This indicates that our proposed approach can actually improve on the separation performance of its individual decomposition procedures. Finally, we can observe that using a manually annotated fundamental frequency track goes along with slight improvements of the objective evaluation measures.

	IBM*	CD-I	CD-B	SL	REPET*	VUIMM*
SDR [dB]	7.9	4.9	3.7	-0.9	4.1	5.6
ISR [dB]	14.9	7.8	6.1	7.2	7.7	7.9
SIR [dB]	11.4	4.6	2.7	-0.8	5.1	9.4
SAR [dB]	14.1	15.2	14.3	13.5	13.1	13.1
OPS	37.9	34.1	32.2	28.2	34.0	31.5
TPS	66.4	53.0	47.6	38.7	56.5	42.8
IPS	74.1	45.2	45.2	54.7	52.8	63.0
APS	30.3	51.0	48.1	37.8	49.2	37.8

Table 1. Average PEASS measures for singing voice estimates on the SiSEC dataset. Results marked with (*) were reported on the website [20]. Higher numbers indicate better results.

3.2. Listening Experiment

In order to analyze the subjective quality of the singing voice extracted by our procedure, we conducted a listening experiment. In order to be able to compare objective and subjective evaluation results, we considered the same procedures and the same dataset as in the objective evaluation. For each of the five recordings, the mixture of all sources as well as the separate singing voice, which consti-

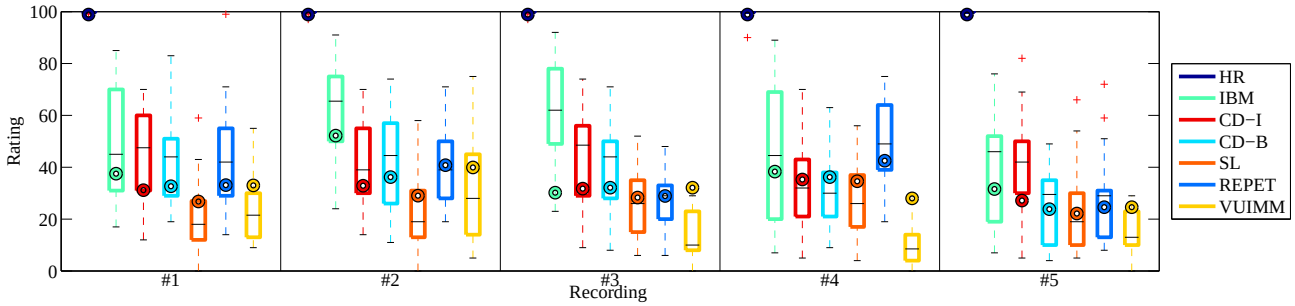

Fig. 3. Results of the performed listening experiment. Boxes indicate the interquartile range, black bars the median, and discs the objective OPS measures. Recordings: BEARLIN (#1), TAMY (#2), ANOTHER DREAMER (#3), FORT MINOR (#4), ULTIMATE NZ TOUR (#5).

tuted the *reference*, was given to the test participants. Their task was then to rate the *overall quality* of the different singing voice estimates with respect to the reference on a scale from 0 (poor) to 100 (excellent). The estimates were presented in a blind test along with a *hidden reference* (HR). Overall, 24 persons participated in the experiment from which six were excluded from the final evaluation during post screening since they were not able to detect the hidden reference reliably. The results are visualized in Figure 3 for all five recordings. In each plot, one finds the respective OPS measure indicated by a colored disk. This objective measure, also having a value range from 0 to 100, is designed to predict the overall quality rating of the test participants. However, you can see that IBM, CD-I, CD-B, and REPET tend to be underrated by the objective measure, while SL and VUIMM tend to be overrated. This indicates that the objective evaluation measures only vaguely correspond to human perception and that listening experiments are still necessary to obtain reliable measurements.

The subjective evaluation gives various insights. First, one can observe that VUIMM was rated rather low for all five recordings. Listening to the respective estimates reveals that here, although the accompaniment is usually suppressed well, the singing voice often has an unnatural, synthetic sound. This demonstrates that the reassembling of the singing voice from fine-grained components is a very difficult task. Also SL falls behind the remaining procedures. It shows that the direct decomposition of a recording into a sparse and a low rank component leaves the sparse singing voice estimate with a lot of musical noise which is reflected in the rating. In comparison to these two procedures, the singing voice estimates of CD-I, CD-B, and REPET are perceived to have a clearly better quality for recordings #1 and #2. For recording #3, CD-I and CD-B demonstrate the benefit of not focusing on a single characteristic of singing voice. Here, REPET's assumption of repeating patterns in the accompaniment is not satisfied which leads to many accompaniment residues in the singing voice estimate and therefore to lower ratings. CD-I and CD-B do not rely on a specific musical structure and therefore receive good ratings for this recording as well. For recording #4, rapped lyrics over a looped beat, REPET excels all other approaches since its assumption of a repeating accompaniment is met perfectly. The decomposition cascade of CD-I and CD-B was however not optimized to capture the spoken "non-singing voice" in rap music. In particular the MR decomposition fails and yields meaningless decomposition results what explains the lower ratings. Looking at the results for recording #5, the first observation is that here even IBM receives rather low ratings. This indicates that the extraction of the singing voice from recording #5 can be considered difficult. However, while CD-I and CD-B were rated similarly for recordings #1 to #4, CD-I performs much better than all other procedures on this recording, even being close to IBM. It turns out that for this recording the blindly estimated fundamental frequency track has an octave error. Therefore, in the MR decomposition step only every second harmonic of the singing voice is extracted which leads to a thin sound of the singing voice estimate of CD-B. However, octave errors can be corrected easily by manual inspection. This example shows that it is possible to stabilize the performance of our procedure, even for difficult recordings, with very little user interaction.

3.3. Accompanying Website

The objective and subjective evaluation showed that our proposed method yields good estimates of the singing voice for the pop music recordings of the SiSEC dataset. One of the advantages of this method lies in its flexibility in the sense that, contrary to approaches like REPET, it does not make strong assumptions about the accompanying music material. To demonstrate that our procedure works on a wide range of musical styles, including genres like classical opera music, romantic piano music with singing, or even metal, we prepared an accompanying website for this paper at [24]. On this website, one can find many illustrative audio examples of decomposition results of our procedure along with the results of the intermediate decomposition steps.

4. CONCLUSION AND FUTURE WORK

In this paper, we have shown how different direct decomposition techniques can be cascaded to disassemble a given music recording into a set of semantically interpretable mid-level components. These components can be easily reassembled to yield estimates of the singing voice and the accompaniment in the recording. Objective and subjective evaluation on a standard dataset suggest that this approach yields singing voice estimates which are comparable to state-of-the-art methods. Furthermore, to demonstrate that our procedure works for a large variety of musical styles and genres going beyond the tested dataset, we also provide an accompanying website with additional audio material. In future work, we will investigate how further decomposition procedures, like for example *center channel extraction* methods [25], can be incorporated into our proposed cascade to further improve the quality of the extracted singing voice.

Acknowledgments:
This work has been supported by the German Research Foundation (DFG MU 2686/6-1). The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer IIS.

5. REFERENCES

[1] Youngmoo Edmund Kim, "Singing voice analysis, synthesis, and modeling," in *Handbook of Signal Processing in Acoustics*, David Havelock, Sonoko Kuwano, and Michael Vorländer, Eds., pp. 359–374. Springer New York, 2008.

[2] Website, "Audionamix, ADX Trax, commercial singing voice extraction software," `http://www.audionamix.com/`.

[3] Tuomas Virtanen, Annamaria Mesaros, and Matti Ryynänen, "Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music," in *Proceedings of the ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition (SAPA)*, Brisbane, Australia, 2008, pp. 17–22.

[4] Po-Sen Huang, Scott Deeann Chen, Paris Smaragdis, and Mark Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012.

[5] Zafar Rafii and Bryan Pardo, "Repeating pattern extraction technique (REPET): A simple method for music/voice separation.," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 21, no. 1, pp. 71–82, 2013.

[6] Derry Fitzgerald, "Vocal separation using nearest neighbours and median filtering," in *Irish Signals and Systems Conference (ISSC 2012)*, June 2012, pp. 1–5.

[7] Antoine Liutkus, Zafar Rafii, Roland Badeau, Bryan Pardo, and Gaël Richard, "Adaptive filtering for music/voice separation exploiting the repeating musical structure," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 53–56.

[8] Hideyuki Tachibana, Nobutaka Ono, and Shigeki Sagayama, "Singing voice enhancement in monaural music signals based on two-stage harmonic/percussive sound separation on multiple resolution spectrograms," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 228–237, January 2013.

[9] Antoine Liutkus, Zafar Rafii, Bryan Pardo, Derry Fitzgerald, and Laurent Daudet, "Kernel additive models for source separation," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4298–4310, August 2014.

[10] Jean-Louis Durrieu, Bertrand David, and Gaël Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1180–1191, 2011.

[11] Shankar Vembu and Stephan Baumann, "Separation of vocals from polyphonic audio recordings," in *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR)*, 2005, pp. 337–344.

[12] Paris Smaragdis, Bhiksha Raj, and Madhusudana Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proceedings of the International Conference on Independent Component Analysis and Signal Separation (ICA)*, London, UK, 2007, pp. 414–421.

[13] Michael A. Casey and Alex Westner, "Separation of mixed audio sources by independent subspace analysis," in *Proceedings of the International Computer Music Conference*, Berlin, Germany, 2000, pp. 154–161.

[14] Alexey Ozerov, Ngoc Q. K. Duong, and Louis Chevallier, "Weighted nonnegative tensor factorization: on monotonicity of multiplicative update rules and application to user-guided audio source separation," Tech. Rep., October 2013.

[15] Chao-Ling Hsu and Jyh-Shing Roger Jang, "On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 18, no. 2, pp. 310–319, February 2010.

[16] Marko Helén and Tuomas Virtanen, "Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine," in *Proceedings of EUSIPCO 2005*, 2005.

[17] Jonathan Driedger, Meinard Müller, and Sascha Disch, "Extending harmonic-percussive separation of audio signals," in *Proceedings of the 15th International Conference on Music Information Retrieval (ISMIR)*, Taipei, Taiwan, 2014.

[18] Ronen Talmon, Israel Cohen, and Sharon Gannot, "Transient noise reduction using nonlocal diffusion filters," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 19, no. 6, pp. 1584–1599, August 2011.

[19] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright, "Robust principal component analysis?," *Journal of the ACM*, vol. 58, no. 3, pp. 11:1–11:37, June 2011.

[20] Website, "Signal separation evaluation campaign (SiSEC)," `http://www.onn.nii.ac.jp/sisec13/evaluation_result/MUS/devMUS2013.htm`.

[21] Shoko Araki, Francesco Nesta, Emmanuel Vincent, Zbynek Koldovský, Guido Nolte, Andreas Ziehe, and Alexis Benichoux, "The 2011 signal separation evaluation campaign (SiSEC2011): Audio source separation," in *LVA/ICA*, 2012, pp. 414–422.

[22] Valentin Emiya, Emmanuel Vincent, Niklas Harlander, and Volker Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.

[23] Justin Salamon and Emilia Gómez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 20, no. 6, pp. 1759–770, 2012.

[24] Jonathan Driedger and Meinard Müller, "Accompanying website: Extracting singing voice from music recordings by cascading audio decomposition techniques," `http://www.audiolabs-erlangen.de/resources/MIR/2015-ICASSP-SVECD`.

[25] Christian Uhle and Emanuël Habets, "Subband center scaling using power ratios," in *Proceedings of the AES 53rd International Conference*, London, UK, 2014.