RE-THINKING SOUND SEPARATION: PRIOR INFORMATION AND ADDITIVITY CONSTRAINT IN SEPARATION ALGORITHMS

Estefanía Cano & Christian Dittmar
Semantic Music Technologies
Fraunhofer IDMT
Ilmenau, Germany
[cano,dmr]@idmt.fraunhofer.de

Gerald Schuller
Institute for Media Technology
Ilmenau University of Technology
Ilmenau, Germany
gerald.schuller@tu-ilmenau.de

ABSTRACT

In this paper, we study the effect of prior information on the quality of informed source separation algorithms. We present results with our system for solo and accompaniment separation and contrast our findings with two other state-of-the art approaches. Results suggest current separation techniques limit performance when compared to extraction process of prior information. Furthermore, we present an alternative view of the separation process where the *additivity constraint* of the algorithm is removed in the attempt to maximize obtained quality. Plausible future directions in sound separation research are discussed.

1. INTRODUCTION

Sound source separation deals with the extraction of independent sound sources from an audio mix. To address this problem, many approaches have been proposed in the literature: filtering and masking techniques, statistical approaches, perceptually motivated systems, time-frequency representation and signal models are some of the techniques used. However even today, sound separation is still considered an unsolved problem. Separation quality of state-of-the-art systems is very limited and dependent on the type of signals used. Currently, there is still no clear direction for a general solution to this problem.

After many years of research, results in the field suggest that separation performance can be improved when prior information about the sources is available. The inclusion of known information about the sources in the separation scheme is referred to as *Informed Sound Source Separation (ISS)* and comprises, among others, the use of MIDI-like musical scores, the use of pitch tracks of one or several sources, oracle sound separation where the original sources are available, and the extraction of model parameters from training data of a particular sound source.

2. PAPER OUTLINE AND MOTIVATION

In Sec. 3 we present a general overview of the state-of-the-art in sound source separation. In an attempt to further understand the potential of current algorithms for informed sound separation, we ask ourselves the questions: How far can we get by using prior information as pitch or musical scores? What could be the expected quality improvement of separation algorithms if we could provide more accurate prior-information of this kind? To address these questions, we discuss in Sec. 5.1 results from three state-of-the-art systems for informed source separation when ground truth (or very accurate) prior information is available. In Sec. 5.2 we use the insights obtained in the previous analyses to propose an alternative to sound source separation. We also address the fundamental goal of sound separation in an attempt to get some insight for future research directions. Concluding remarks are presented in Sec. 6

3. STATE-OF-THE-ART

In general, source separation approaches can be classified according to the processing technique used. Three main categories exist: statistical approaches, classical signal processing approaches, and computational auditory scene analysis (CASA) approaches. Statistical techniques for sound separation generally assume certain statistical properties of the sound sources. Systems based on Independent Subspace Analysis (ISA) [1], [2], Non-Negative Matrix Factorization (NMF) [3], [4], [5], tensor factorization[6], [7], and sparse coding [8], [9], [10] have been proposed. In the case of signal processing approaches for sound separation, different forms of masking and filtering techniques to extract the desired sources are used [11], [12]. Computational auditory scene analysis (CASA) techniques have also been proposed [13], [14].

Many systems for sound source separation have attempted to use pitch as prior information.These systems are based on the assumption that every sound source follows a defined pitch sequence over time. The system described in [15] proposes an invertible mid-level representation of the audio signal which gives access to some semantically rich salience functions for pitch and timbre content analysis. The system uses an instantaneous signal model (IMM) which represents the audio signal as the sum of a signal of interest, i.e., the lead instrument, and a residual, i.e., accompaniment. A source-filter model is used to represent the signal of interest. Information from the *source* is related to the pitch of the lead instrument and information from the *filter* is related to the timbre of the instrument. The residual is modeled using non-negative matrix factorization (NMF). The mid-level representation is used to separate lead instrument form accompaniment in conjunction with a Wiener masking approach. In [16], an approach for singing voice extraction in stereo recordings that uses panning information and a probabilistic pitch tracking approach is presented. Some approaches have been proposed for supervised pitch extraction with a subsequent separation scheme [17], [18], [19], [20].

Score-informed source separation extracts audio sources from the mix by using a MIDI-like score representation of the desired source(s). A score-informed source separation algorithm is presented in [21]. This system attempts to separate solo instruments from their accompaniment using MIDI-like scores of the lead instrument as prior information. The approach uses chroma-based

dynamic time warping (DTW) to address global misalignments between the score and the audio signal. Furthermore, a MIDI confidence measure is proposed to deal with small-scale misalignments. In [22] a score-informed separation algorithm is described, which is based on Probabilistic Latent Component Analysis (PLCA) and the use of synthesized versions of the score as prior distributions in the PLCA decomposition of the original mix.

4. PROPOSED ALGORITHM

In this section, we describe two algorithms used for the experiments described in sections 5.1 and 5.2. The results from the different experiments are used to address the questions posed in the motivation in Sec. 2.

In [23], we propose a pitch-informed method to separate solo instruments from accompaniment. Pitch information from the solo instrument is extracted with an approach described in [24]. The rough pitch estimates are refined using a linear interpolation approach where the energy of the fundamental frequency and its harmonic components is calculated for each interpolation step. The maximum energy is taken as an indicator of the new fundamental frequency. A harmonic component refinement stage iteratively constructs a harmonic series for each fundamental frequency using known acoustical characteristics of musical instruments. Initial binary masks are created based on the iterative estimation and a post-processing stage is used to take care of attack frames and reduce interference from percussive sources. After post-processing, masks are no longer binary. Solo and accompaniment sources are re-synthesized using the obtained masks. For the remainder of this paper, this algorithm will be referred to as ***Cano1***.

Furthermore, we present a basic modification that makes our algorithm more suited for vocal extraction. This algorithm will be referred to as ***Cano2*** and simply modifies the estimation stage by including a noise spectrum in the *Harmonic Refinement* stage to capture characteristic noise-like sounds in vocal signals. Similar approaches have also been used in [17].

5. EXPERIMENTS & DISCUSSION

For the experiments conducted in this paper, a dataset of 10 multitrack recordings was used. These recording are part of the PEASS [1] and BSS [2] datasets and are freely available for download under CC license. All the signals in the dataset are vocal tracks (male or female) with accompaniment. For all signals, the multitrack recordings were mixed to obtain accompaniment tracks, solo tracks, and a final monaural mix. The signals are described in Table 1.

Recognizing the importance of including perceptual aspects in the evaluation of sound separation results, we use the PEASS Toolkit- Perceptual Evaluation Methods for Audio Source Separation [25] to measure quality of the separated signals.The PEASS Toolkit presents a set of four objective perceptual measures for separation quality assessment, i.e., Overall Perceptual Score (OPS), Target Perceptual Score (TPS), Interference Perceptual Score (IPS), Artifact Perceptual Score (APS). For reference purposes, we also present common objective scores based on energy ratios, i.e., Signal to Distortion Ratio (SDR), Image to Spatial Distortion Ratio (ISR), Signal to Interference Ratio (SIR), Signal to Artifact Ratio (SAR).

5.1. Prior Information in Separation Algorithms

In this section, we evaluate the effects of prior information on the quality of the proposed approach and contrast our finding with two other state-of-the-art algorithms.

We refer to our previous work on pitch-informed sound separation (***Cano1***) to separate audio recordings into solo and accompaniment tracks. It is to be noted that both the pitch extraction and separation stages in ***Cano1*** are completely automatic. To create ground truth information, the Songs2See Editor [19] was used to manually correct and refine the pitch extraction of the solo instrument. We used the corrected pitch sequences to feed our separation algorithm and obtain solo and accompaniment tracks by bypassing the automatic pitch detection stage. The goal of this experiment is to assess the potential of the separation algorithm when accurate prior information is available. The algorithm that uses ground truth pitch information will be referred to as ***CanoU***.

For this study, signals 1, 8, and 9 from our dataset were selected to obtain ground truth pitch information and perform separation with the two algorithms (***Cano1***, ***CanoU***). Results are presented in Table 2. Perceptual measures tend to evidence a quality improvement when accurate pitch information is provided to the algorithm. However, the more interesting observation is the fact that even though there is a quality improvement, results are far from reaching maximum quality scores. The maximum Overall Perceptual Score (OPS) obtained was 32.02 for accomp9 and high Interference Perceptual Scores (IPS) are obtained in general, reaching the highest score of 74.8 in solo1. The highest scores obtained for the four perceptual measures are written in bold font in Table 2.

Given that these results only represent the particular case of our algorithm and cannot be generalized, we revisit the results presented by [15] and [21] to get a wider view of the performance of separation algorithms. The details of these algorithms were briefly described in Sec. 3. For the remainder of this paper, these algorithms will be referred to as ***Durrieu*** and ***Bosch*** respectively. In both cases, a system for solo/accompaniment separation is presented. Similarly, the authors present results comparing the performance of the fully automatic algorithm with the one obtained when ground truth information is available. In [17], Durrieu presents an user-assisted algorithm where the pitch track of the solo instrument can be refined using a specially designed user interface. This algorithm will be referred to as ***DurrieuU***. In the case of [21], the authors manually align the scores to the audio tracks and use them as ground truth information. The user-assisted version of this algorithm will be referred to as ***BoschU***.

In Table 3 we show some of the results presented by the authors related to these four algorithms. In the case of ***Durrieu*** and ***DurrieuU***, we present the average scores for solo extraction obtained in the 2011 Signal Separation Evaluation Campaign (SiSEC11), which can be accessed in the campaign's website [3]. In this case, both the PEASS measures and the energy ratios are available. In the case of ***Bosch*** and ***BoschU***, the results for solo and accompaniment extraction for the dataset S1-D3 using a Wall-N mask are

[1] http://sisec.wiki.irisa.fr/tiki-index.php?page=Professionally+produced+music+recordings

[2] http://bass-db.gforge.inria.fr/BASS-dB/?show=browse&id=mtracks

[3] http://www.irisa.fr/metiss/SiSEC11/professional/test_eval2011.htm

Table 1: *Data set used: In the text, signal numbers are used to refer to each of the signals.*

Signal Num.	**Name**	**Segment [sec]**
1	Bearlin Roads	85 - 99
2	Tamy que pena tanto faz	6 - 19
3	Another Dreamer	69 - 94
4	Ultimate nz tour	43 - 61
5	Dreams	0 - 35
6	Life as a disturbed infobeing	0 - 57
7	Mix Tape	7 - 53
8	The ones we love	32 - 48
9	We weren't there	0 - 32
10	Wreck	15 - 34

Table 2: *Signal version obtained with the different masking and post-processing approaches*

Sig. Num	**Source**	**Alg.**	**OPS**	**TPS**	**IPS**	**APS**	**SDR**	**ISR**	**SIR**	**SAR**
1	solo	Cano1	20.01	2.41	72.25	6.52	-5.19	-4.15	5.26	9.85
		CanoU	22.09	3.17	**74.8**	9.21	-6.71	-5.99	8.38	11.58
	accomp	Cano1	24.99	34.38	57.30	36.97	-3.37	-3.25	12.66	15.55
		CanoU	27.08	39.63	54.31	40.64	-3.544	-3.42	13.02	17.13
8	solo	Cano1	19.31	2.60	51.16	9.36	-4.91	-3.70	3.17	11.368
		CanoU	23.23	3.20	72.73	8.49	-4.43	-3.72	6.47	12.67
	accomp	Cano1	26.60	43.46	62.41	42.53	-3.54	-3.28	10.73	13.46
		CanoU	30.73	**56.13**	55.33	**48.41**	-3.623	-3.41	12.65	14.66
9	solo	Cano1	25.66	0.96	67.7	2.23	-3.64	-3.11	8.47	10.23
		CanoU	18.05	3.8	61.31	10.83	-3.72	-3.01	5.81	11.95
	accomp	Cano1	30.84	29.03	59.74	33.41	-3.44	-3.15	9.87	14.48
		CanoU	**32.02**	30.44	53.96	33.69	-2.81	-2.55	9.59	15.12

presented. In this case only the Signal to Distortion Ratio (SDR) is available. We refer the reader to [21] for details. It is important to bear in mind that the three algorithms presented in this Section (Cano, Durrieu, Bosch) make use of different datasets and their results cannot be used for direct comparison. The goal of presenting these results is to describe a similar phenomenon occurring in different algorithms but not to perform a direct comparison between them.

A similar behavior is observed in both of the comparison algorithms. The use of ground truth prior information tends to result in higher quality measures. However, two important observations can be made: (1) Scores with ground truth information are still far from reaching maximum levels, and (2) Quality differences between ground truth and automatically extracted pitch are marginal.

In general, results reveal that there is still much room for improvement when it comes to informed sound separation algorithms. Including prior information obviously benefits performance but it is hard to envision a generalized and robust solution given current results. This leads us to two possible paths: **(1)** On the one hand we could consider the possibility that the type of prior information that we are currently using does not carry enough signal details to allow robust separation and consider possibilities to enhance the information used. Taking into account the great diversity not only of musical signals, but also of playing styles, genres, and recording conditions which a separation algorithm can encounter, the expectation that such general information as pitch could suffice to guide the separation schemes, comes short. We could then consider including, besides pitch information, other types of information that allow better characterization of sound sources. This would naturally lead to the development of target-designed algorithms optimized for the extraction of a particular class of signals. The use of instrument-specific information and instrument models within the separation schemes could be an option as for example in [26]. **(2)** On the other hand we could consider the possibility to further improve our current sound analysis and synthesis methods so that more accurate information can be extracted. Works on this topic include developments of reassignment and derivative methods among others [27]. In [28] and [29], the authors already recognize the limitations of current analysis techniques and propose an informed-analysis front-end to improve sound source separation. In these approaches prior information is included directly in the analysis in the form of watermarks and bits of ground truth information, respectively. In both cases, separation results evidence an improvement. However, theses approaches have a limitation in the sense that the original signals need to be available to extract the prior information inserted in the analysis. Having the original,

Table 3: *Comparison between two automatic algorithms and their corresponding user-assisted versions*

Alg.	OPS	TPS	IPS	APS	SDR	ISR	SIR	SAR
Durrieu **[solo]**	22.4	28.8	59.0	30.8	3.8	6.2	Inf	3.1
DurrieuU **[solo]**	26.0	28.4	61.1	29.9	5.4	8.3	Inf	5.4
Bosch **[Accom/solo]**	-	-	-	-	10.35/6.20	-	-	-
BoschU **[Accom/solo]**	-	-	-	-	10.46/6.31	-	-	-

unmixed signals is not always possible.

There is however an alternative possibility to be explored. Results have shown that not only is the available information about the sources of critical importance for separation performance, but also the mechanisms used to include such information in the separation scheme. Including prior information in the time-frequency domain (like pitch or MIDI scores) has proven to contribute to the quality of sound separation. Including information in the analysis stage (like watermarks and bits of information) has also proven to improve separation. This option with the difficulty of requiring the original sources to extract the prior information. The third option is then naturally, including prior information about the sources in the synthesis stage. Here again, separation algorithms would be optimized to deal with a particular class of signals. Instrument synthesis models, developed and trained off-line, could be used to recreate the original signals as closely as possible. This option leaves open the possibility to include information in the time-frequency transform coming from domains as pitch, timbre, scores, etc. Furthermore, having the original signals would not be required. Another important characteristic of such an approach would be that a strict constraint to exactly reconstruct the original mix from the extracted sources, could not be set. In the remainder of this paper, the hard constraint imposed to most separation algorithms to exactly reconstruct the mix from the extracted sources is referred to as *additivity constraint*. With the third option, the analysis and synthesis stages of separation algorithms would most likely use different signal processing techniques and such constraint would be difficult to impose.

5.2. Redefining Sound Separation

After the discussion presented in Sec. 5.1, there is still an open question that we wish to address: Is there any possibility to obtain a performance gain with our current separation approaches without fundamentally changing them?

To address this question, we created different versions of our separation algorithm which are described in Table 4. For each version, the pitch detection and spectral component estimation is kept unchanged, but different spectral masking techniques are used to obtain the resulting signals. For the cases where Wiener filtering is used, p denotes the power to which each individual element of the spectrograms are raised. In versions 1 and 6, the *Post-Processing* stage of the algorithm is bypassed to allow binary and Wiener masking respectively. We advise the reader to refer to [23] for algorithm details. As can be seen, the modifications performed to the algorithm are rather basic and do not fundamentally change the original system. However, each one of them has clear effects on its performance.

For this experiment we use the ten tracks from the dataset described in Table 1 and for each one of them, solo and accompaniment tracks are extracted using each of the seven algorithm versions. As in Sec. 5.1, the PEASS Toolkit is used for quality evaluation and measures based on energy ratios are presented for reference. Separation results for the solo and accompaniment signals using the PEASS Toolkit are presented in Figures 1 and 2 respectively. Similarly, results for the solo and accompaniment signals using the energy ratios are presented in Figures 3 and 4. In all figures, the mean performance over the ten signals is presented for the seven algorithm versions. The whiskers indicate the standard deviation of each version and the highest score among all versions, is shown with an inner (blue) dot in the marker.

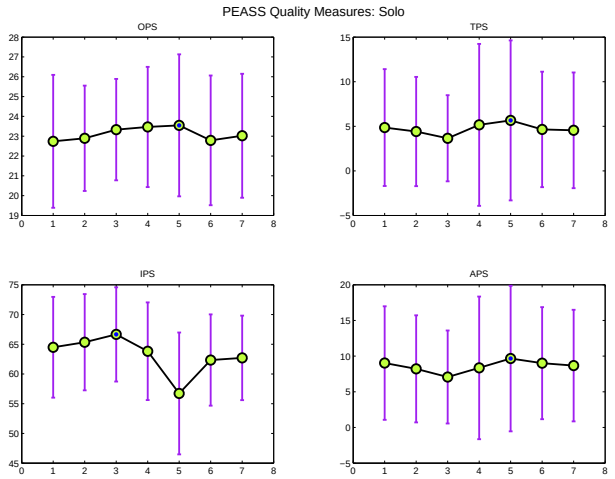

Figure 1: *Mean perceptual scores for the solo signals : Overall Perceptual Score (OPS), Target Perceptual Score (TPS), Interference Perceptual Score (IPS), Artifact Perceptual Score (APS).*

As can be seen in Fig. 1, for three of the four perceptual scores (OPS, TPS, APS) for the solo tracks, the highest mean performance is obtained with the ***Cano2*** algorithm (version 5). This is however an expected result, as the algorithm was specifically modified to better handle vocal signals. On the other hand, results for the accompaniment tracks differ. As shown in Fig. 2, the highest scores are obtained for three of the measures (OPS, TPS, APS) with different versions of the ***Cano1*** algorithm. For the Interference Perceptual Score (IPS), the highest mean performance is obtained with the ***Cano2*** algorithm. This result further evidences that

Table 4: *Signal version obtained with the different masking and post-processing approaches*

Version	Description
1	Binary
2	Cano1
3	Cano1 + Wiener [p=2]
4	Cano2 + Wiener [p=2]
5	Cano2
6	Wiener [p=0.3]
7	Cano1 + Wiener [p=0.3]

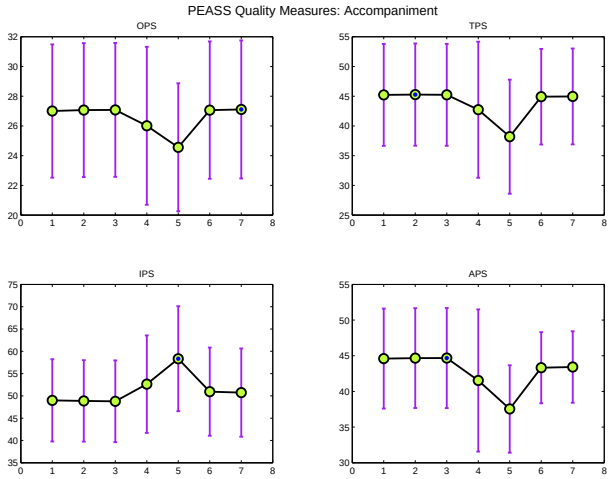

Figure 2: *Mean perceptual scores for the accompaniment signals: Overall Perceptual Score (OPS), Target Perceptual Score (TPS), Interference Perceptual Score (IPS), Artifact Perceptual Score (APS).*

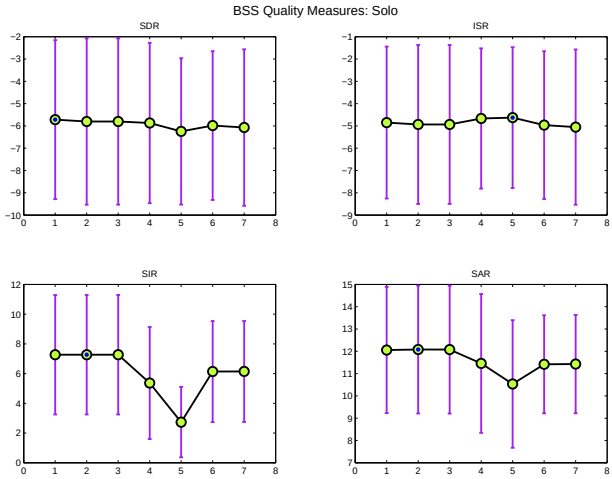

Figure 3: *Mean energy ratios for the solo signals: Signal to Distortion Ratio (SDR), Image to Spatial Distortion Ratio (ISR), Signal to Interference Ratio (SIR), Signal to Artifact Ratio (SAR).*

better solo extraction is obtained with the ***Cano2*** algorithm as for the backing track, the vocal signal is, in this case, the only source of interference.

Results suggest that for the particular task of solo and accompaniment separation, the highest perceptual scores can be obtained differently for each of the desired sources. Algorithm modifications that might benefit solo extraction can potentially have a negative effect in the performance for accompaniment extraction. In this line of thought, we conducted informal tests using the ***Cano2*** algorithm to extract solo tracks from different musical instruments (clarinet, trumpet, and saxophone). In all cases, performance of solo extraction suggested a performance decrease. This further confirms the idea that for our current approach, performance might be maximized if different versions of the algorithm are used for the solo and accompaniment tracks. This brings us back to the concept of *additivity constraint* presented in Sec.5.1. Bearing in mind the possibilities and limitations of our current sound analysis techniques, and the fact that theoretical bound exist for them, allowing separation algorithms to extract sources with the goal of maximizing perceptual quality instead of reconstructing the original mix, might bring us better final results.

Following this line of thought, the idea of moving from separation to understanding presented in a keynote presentation by Smaragdis [4], becomes relevant. In most cases, source separation is not the final goal but most likely, an intermediate step to further types of processing: more accurate music transcription, re-mixing audio tracks, audio classification, etc. In that sense, extracting the exact original source might not even be necessary for the final application. Different quality requirements for different applications might be needed: music transcription, for example, would most likely require high Interference Perceptual Scores (IPS) for robust performance, as pitch tracks of the independent sources are the final goal. On the other hand, IPS requirements might not be so strict when it comes to re-mixing audio tracks. Minimizing arti-

[4] `http://www.cs.illinois.edu/~paris/pubs/smaragdis-LVAICA12.pdf`

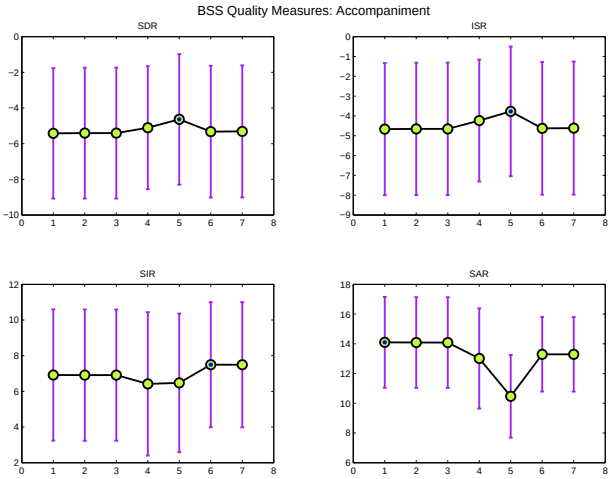

Figure 4: *Mean energy ratios for the accompaniment signals: Signal to Distortion Ratio (SDR), Image to Spatial Distortion Ratio (ISR), Signal to Interference Ratio (SIR), Signal to Artifact Ratio (SAR).*

facts and preserving the sources are probably more relevant and consequently higher APS and TPS might be required. Thus, considering the final processing goal and its quality requirements, instead of focusing on the separation task only, might bring better overall results and open possibilities for further analyses.

6. CONCLUDING REMARKS

We have addressed two defining topics in informed sound separation research: (1) The effects of pitch and score information in the performance of separation algorithms were studied showing that attainable quality, when accurate prior information is available, still fails to reach maximal scores. We review three possibilities to include prior information in separation approaches; namely in the analysis , time-frequency transform, and synthesis stages. Due to the processing possibilities and flexibility that it provides to the separation scheme, we see great potential in including information directly in the synthesis stage. Future work will be conducted in this direction. (2) We propose the possibility to remove the *additivity constraint* to improve quality of separation and as a future direction where algorithms could have completely independent analysis and synthesis approaches. In this sense we redefine our goal with separation research from extracting the original sources from the mix, to obtaining sources that meet perceptual quality requirements imposed for different applications and that allow the use of separation schemes as intermediate processing steps.

7. REFERENCES

[1] Derry FitzGerald, R Lawlor, and Eugene Coyle, "Sub-band independent subspace analysis for drum transcription," in *5th International Conference on Digital Audio Effects (DAFx-02)*, Hamburg, Germany, 2002, number 5, pp. 1–5.

[2] Christian Uhle, C Dittmar, and T Sporer, "Extraction of drum tracks from polyphonic music using independent subspace analysis," in *4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA 2003)*, Nara, Japan, 2003, pp. 843–848.

[3] Romain Hennequin, Bertrand David, and Roland Badeau, "Score informed audio source separation using a parametric model of non-negative spectrogram," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011)*, Prague, Czech Republic, 2011, number 1, pp. 45–48.

[4] S. Kirbiz and Paris Smaragdis, "Adaptive time-frequency resolution for single channel sound source separation based on non-negative matrix factorization," in *IEEE 19th Signal Processing and Communications Applications Conference (SIU 2011)*, Antalya, Turkey, 2011, pp. 964–967.

[5] A. Lefevre, Francis Bach, and C Févotte, "Itakura-Saito non-negative matrix factorization with group sparsity," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011)*, Prague, Czech Republic, 2011, number 1, pp. 21–24.

[6] Derry Fitzgerald, Matt Cranitch, and Eugene Coyle, "Using tensor factorisation models to separate drums from polyphonic music," in *12th International Conference on Digital Audio Effects (DAFx-09)*, Como, Italy, 2009, pp. 1–5.

[7] U. Simsekli and A. Cemgil, "Score guided musical source separation using generalized coupled tensor factorization," in *20th European Signal Processing Conference (EUSIPCO 2012)*, Bucharest, Romania, 2012, number Eusipco, pp. 2639–2643.

[8] Mark D. Plumbley, Thomas Blumensath, Laurent Daudet, Rémi Gribonval, and Mike Davies, "Sparse representations in audio and music: from coding to source separation," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 995–1005, 2010.

[9] Samer a Abdallah and Mark D Plumbley, "Unsupervised analysis of polyphonic music by sparse coding.," *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, vol. 17, no. 1, pp. 179–96, Jan. 2006.

[10] Rémi Gribonval and Synlvain Lesage, "A survey of sparse component analysis for blind source separation: principles, perspectives, and new challenges," in *European Symposium on Artificial Neural Networks (ESANN 2006)*, Bruges, Belgium, 2006, number April.

[11] Derry Fitzgerald, "Harmonic/percussive separation using median filtering," in *13th International Conference on Digital Audio Effects (DAFx-10)*, Graz, Austria, 2010, number 1, pp. 10–13.

[12] David Gunawan and Deep Sen, "Separation of Harmonic Musical Instrument Notes Using Spectro-Temporal Modeling of Harmonic Magnitudes and Spectrogram Inversion with Phase Optimization," *Journal of the Audio Engineering Society (AES)*, vol. 60, no. 12, 2012.

[13] L. Drake, J. Rutledge, J. Zhang, and a Katsaggelos, "A Computational Auditory Scene Analysis-Enhanced Beamforming Approach for Sound Source Separation," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, no. 1, pp. 403681, 2009.

[14] Yipeng Li, John Woodruff, and D. Wang, "Monaural musical sound separation based on pitch and common amplitude modulation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 17, no. 7, pp. 1361–1371, 2009.

[15] 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.

[16] Ricard Marxer, Jordi Janer, and Jordi Bonada, "Low-Latency instrument separation in polyphonic audio using timbre models," *Latent Variable Analysis and Signal Separation*, vol. 7191, pp. 314–321, 2012.

[17] Jean-Louis Durrieu and Jean-Philippe Thiran, "Musical audio source separation based on user-selected f0 track," *Latent Variable Analysis and Signal Separation*, , no. 1, pp. 1–8, 2012.

[18] Derry FitzGerald, "User assisted separation using tensor factorisation," in *20th European Signal Processing Conference (EUSIPCO 2012)*, Bucharest, Romania, 2012, pp. 2412–2416.

[19] Estefanía Cano, Christian Dittmar, and Sascha Grollmisch, "Songs2See: Learn to Play by Playing," in *12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, Miami, USA, 2011.

[20] Benoit Fuentes, Roland Badeau, and Gaël Richard, "Blind Harmonic Adaptive Decomposition Applied to Supervised Source Separation," in *2009 Ninth IEEE International Conference on Advanced Learning Technologies*, Bucharest, Romania, 2012, number Eusipco, pp. 2654–2658.

[21] Juan J. Bosch, K. Kondo, R. Marxer, and J. Janer, "Score-informed and timbre independent lead instrument separation in real-world scenarios," in *20th European Signal Processing Conference (EUSIPCO 2012)*, Bucharest, Romania, 2012, vol. 25, pp. 2417–2421.

[22] Joachim Ganseman, Paul Scheunders, Gautham J. Mysore, and Jonathan S. Abel, "Evaluation of a score-informed source separation system," in *11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, Utrecht, Netherlands, 2010.

[23] Estefanía Cano, Christian Dittmar, and Gerald Schuller, "Efficient Implementation of a System for Solo and Accompaniment Separation in Polyphonic Music," in *20th European Signal Processing Conference (EUSIPCO 2012)*, Bucharest, Romania, 2012, pp. 285–289.

[24] Karin Dressler, "Pitch Estimation by pair-wise Evaluation of Spectral Peaks," in *Proceedings of the AES 42nd Conference on Semantic Audio*, Ilmenau, 2011, pp. 278–290.

[25] 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, Sept. 2011.

[26] Mathieu Coïc and Juan José Burred, "Bayesian Non-negative Matrix Factorization with Learned Temporal Smoothness Priors," in *International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, Tel-Aviv, Israel, 2012, pp. 280–287.

[27] Sylvain Marchand and Philippe Depalle, "Generalization of the derivative analysis method to non-stationary sinusoidal modeling.," in *11th International Conference on Digital Audio Effects (DAFx-08)*, Espoo, Finland, 2008, pp. 1–8.

[28] Mathieu Parvaix, Laurent Girin, and Jean-Marc Brossier, "A watermarking-based method for single-channel audio source separation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2009)*, Taiper, Taiwan, 2009, pp. 101–104.

[29] Sylvain Marchand and Dominique Fourer, "Breaking the bounds: Introducing informed spectral analysis," in *13th International Conference on Digital Audio Effects (DAFx-10)*, Graz, Austria, 2010, pp. 1–8.

[30] MA Casey and A Westner, "Separation of mixed audio sources by independent subspace analysis," in *International Computer Music Conference (ICMC 2000)*, Berlin, Germany, 2000.

[31] Niall Cahill, Rory Cooney, and Robert Lawlor, "An Enhanced Implementation of the ADRess (Azimuth Discrimination and Resynthesis) Music Source Separation Algorithm," in *121th Convention of the Audio Engineering Society (AES)*, San Francisco, USA, 2006.