MONAURAL BLIND SOURCE SEPARATION IN THE CONTEXT OF VOCAL DETECTION

Bernhard Lehner, Gerhard Widmer
Department of Computational Perception
Johannes Kepler University of Linz
{bernhard.lehner,gerhard.widmer}@jku.at

ABSTRACT

In this paper, we evaluate the usefulness of several monaural blind source separation (BSS) algorithms in the context of vocal detection (VD). BSS is the problem of recovering several sources, given only a mixture. VD is the problem of automatically identifying the parts in a mixed audio signal, where at least one person is singing. We compare the results of three different strategies for utilising the estimated singing voice signals from four state-of-the-art source separation algorithms. In order to assess the performance of those strategies on an internal data set, we use two different feature sets, each fed to two different classifiers. After selecting the most promising approach, the results on two publicly available data sets are presented. In an additional experiment, we use the improved VD for a simple post-processing technique: For the final estimation of the source signals, we decide to use either silence, or the mixed, or the separated signals, according to the VD. The results of traditionally used BSS evaluation methods suggest that this is useful for both the estimated background signals, as well as for the estimated vocals.

1. INTRODUCTION

Monaural Blind Source Separation (BSS) is a technique for the separation of at least two components from a single-channel signal without using additional information, like the instrumentation or the notation of a musical piece. It is extremely challenging, since we have to deal with the fact, that less mixtures than sources are at hand.

The result of BSS could be useful for many tasks like remixing, creating karaoke songs, manipulate isolated instruments, and so on. Certain Music Information Retrieval (MIR) tasks could also benefit from a BSS as a pre-processing step, e.g. vocalist similarity, pitch detection, automatic transcription, keyword spotting, ...

Unfortunately, it is hard to estimate the usefulness of a certain BSS algorithm for a specific task beforehand. Metrics usually used for evaluating BSS (see Section 4.1) are certainly useful for comparison purposes, but have only limited meaningfulness when it comes to the ultimate question if BSS is actually useful for a specific task.

To give an example, Schuller et al. had no success with achieving better results for tempo and key detection by utilising drum beat separation in [20], despite the fact that the audible results seemed good enough to be used for music remixing. On the other hand, Weninger et al. achieved a significant performance gain in the 3-class task of detecting singing voice segments and simultaneously recognising the vocalist gender in [21].

Therefore, we evaluate the usefulness of several state-of-the-art BSS algorithms in the context of vocal detection (VD), also referred to as singing voice detection. For this task, usually several features are extracted frame-wise from the audio signal and fed to a classifier [9, 13, 14, 19, 21,22], or even to a speech-recogniser [1] in order to obtain the vocal/non-vocal decision. Given this use case, we are mainly interested in separating the signal into two sources: vocals and background. In order to find the best usage of the BSS results, we discuss the outcome of three different strategies.

Furthermore, we investigate if the quality of the separated sources can be improved by using VD as a post-processing technique: For the estimated vocals we mute the parts which are not classified as such by our VD to reduce non-voiced artifacts. For the estimated background, we replace the parts which are classified as non-vocal by our VD with the original mixed audio signal.

2. SELECTED BSS ALGORITHMS

We selected four state-of-the-art BSS algorithms, all of them were already used to extract the singing voice from a mixed audio signal, and reference implementations are provided by the authors. Due to limited space, we can discuss the methods only briefly, and refer to the original papers.

The adaptive REpeating Pattern Extraction Technique (aREPET) is a method, where repeating patterns (background) are identified and used to separate non-repeating (foreground) elements. Those elements are often the varying vocals, and it was shown in [18], that this technique can be used for Music/Voice Separation. There are several variants of the REPET algorithm [11, 16–18], whereas according to the results from Liutkus et al. in [12] the aREPET yielded the best Δ_{SDR} for vocals out of three variants. Therefore, we consider the aREPET the most promising variant and choose this for our comparison.

The FASST toolbox by Ozerov et al. [15] allows to specify prior information and implement arbitrary separation problems. Therefore, it is not merely a method, but more a general framework. However, a baseline implementation is included in the toolbox, which separates a song into the four sources drums, bass, main melody, and the rest. It comes with pre-trained models for several sources, incl. singing voice, which is in our case used to extract the main melody source.

The Kernel Additive Modelling (KAM) approach [10, 12] uses source-dependent proximity kernels to describe local dynamics like periodicity (similar to REPET), smoothness, stability over time or frequency, and more. The different sources are then separated by an algorithm called iterative kernel backfitting.

Huang et al. use in [8] Robust Principal Component Analysis (RPCA) for the separation of singing voice. Their basic assumptions are, that singing voice is relatively sparse within songs, and accompaniment is in a low-rank subspace due to its repetitive structure. Their method uses solely the spectrogram as input, and neither training nor particular features are required.

Another interesting approach, which we didn't include in our experiments (because of the results reported in [12]), is suggested by Durrieu et al. in [3], where a source-filter model is used for the vocals, and non-negative matrix factorisation (NMF) for the background.

3. EXPERIMENTS

In this Section, we discuss the outcome of three different strategies to utilise the results of the selected BSS algorithms.

3.1 Internal Data Set

For the first experiments, we use a set of 149 annotated rock songs by 149 different artists. All songs are recorded at a sampling rate of 22 kHz with 16-bit resolution and converted to mono. Background and vocal tracks are separately available to allow for a more complex evaluation of the results. Approximately 52% of the frames are annotated as vocal, and the amount of pure singing, i.e. without instrumental accompaniment, is negligible. This set is split into a 75 song train set, and a 74 song test set, approximately 5h each. It is challenging for BSS algorithms, because it contains lots of guitar soli, where singing voice characteristics are mimicked.

3.2 Feature Sets and Classifier

For the following experiments we choose the features from [9], which we refer to as *IC14* for the remainder of this paper. The IC14 feature vector comprises 116 attributes in total. This method was already compared to several others in [9], and turned out to deliver the best VD results in almost every testing scenario.

For new insights, we compare this feature set to the one used by Weninger et al. in [22], henceforth referred to as *OS11*. This feature vector comprises 46 attributes. It was used along with a BLSTM-RNN classifier to achieve state-of-the-art performance for several singing voice related classification tasks, among them gender recognition and VD.

In our implementation, both feature sets are extracted with a fixed frequency of five observations per second (200 ms frames). Therefore, the units of audio to be classified are 200 ms frames. In the original implementation of OS11 in [22], the features were extracted beat-wise, hence using a variable framesize. As classifier, we choose the Random Forest (RF) as well as the Support Vector Machine (SVM) implementations of the Weka toolkit [7]. To be able to focus on the performance of feature set and classifier, no post-processing is applied.

3.3 Foreground Separation Evaluation

In this Section we present the results of the first strategy, were we extract the features (IC14 from [9], OS11 from [22]) from just the separated foreground audio signals.

The results are presented in Table 1, where we first see the performance of a model trained from the original audio, and tested with the original audio (row MIX). To simulate a perfect BSS, we additionally extract the features from the real vocal track (containing only vocals and silence) of the song, and test with the same model as before (row VOC). Clearly, the results improve by just using pure vocals as test data, e.g. from 83.7% to 91.6% accuracy for the IC14 feature set and the Random Forest (col. RF-accuracy-IC14).

For the upcoming results, we use the placeholder METHOD to refer to the four BSS algorithms {aREPET, FASST, KAM, RPCA} in general. For $\text{METHOD}_{\text{mix}}$ classification, we always use the model trained from the mixed audio signals. For $\text{METHOD}_{\text{sep}}$ classification, we use the model trained from the separated vocal signals to incorporate BSS characteristics.

The test data presented to the classifier is extracted from the separated vocals. As can be seen in Table 1, consistently and regardless of the feature set and classifier, both accuracy and F-measure improve when the model is trained with the separated vocals instead of the mixed audio data.

Nevertheless, there is quite some room for improvement, since all methods show a substantial performance decrease relative to testing with pure vocals (row VOC).

Compared to the results of training and testing with the mixed audio data (row MIX), only the $\text{aREPET}_{\text{sep}}$ and RPCA_{sep} methods, where both training and testing is done with the separated foreground, yields slightly better results (e.g. for RF-accuracy-IC14: 83.7% vs. 84.1% and 84.5% respectively).

Interestingly, the feature set from [22] in combination with the SVM (col. SVM-accuracy-OS11) is only in the pure vocals scenario (row VOC) superior to the feature set from [9] (col. SVM-accuracy-IC14) (94.9% vs. 93.9% accuracy). It seems that the feature set from [22] is quite ca-

Internal Data Set (framesize=200ms)

	RF				SVM			
	accuracy		F-measure		accuracy		F-measure	
	IC14	OS11	IC14	OS11	IC14	OS11	IC14	OS11
MIX	.837	.795	.846	.814	.855	.807	.863	.819
VOC	.916	.910	.920	.905	.939	.949	.943	.951
aREPET$_{\text{mix}}$	.768	.756	.800	.781	.783	.742	.797	.789
aREPET$_{\text{sep}}$	.841	.796	.850	.810	.861	.811	.866	.822
FASST$_{\text{mix}}$	.732	.670	.682	.603	.751	.711	.756	.686
FASST$_{\text{sep}}$	.826	.778	.835	.795	.845	.791	.854	.803
KAM$_{\text{mix}}$	.752	.736	.773	.738	.631	.577	.728	.709
KAM$_{\text{sep}}$	.826	.786	.835	.798	.849	.805	.855	.815
RPCA$_{\text{mix}}$	.752	.691	.788	.763	.620	.563	.704	.703
RPCA$_{\text{sep}}$	.845	.797	.851	.809	.861	.820	.867	.828

Table 1. Results of Foreground Separation. F-measure relates to the class *vocal*. MIX: trained and tested with mixed audio; VOC: trained with mixed audio, tested with pure vocals. METHOD$_{\text{mix}}$: trained with mixed audio, tested with separated vocals; METHOD$_{\text{sep}}$: trained and tested with separated vocals. The columns IC14 and OS11 refer to the feature sets used in [9] and [22].

pable to model singing voice, but less robust to background noise.

Generally, comparing the performance of the classifiers, SVM delivers better results than the Random Forest. Regarding the feature set, IC14 seems to be the better choice. This can also be observed in the following experiments.

3.4 Foreground Concatenation Evaluation

Here, we concatenate the features extracted from the mix to the features extracted from the separated foreground into a single feature vector, hence doubling its size.

In Table 2 we can see that this strategy leads to better results regardless of BSS method, classifier and feature set. In order to assess the upper bound of this strategy, we include the results when using the real vocals also (row MIX+VOC), simulating perfect separation. Similar to Section 3.3, the results from utilising RPCA are the best, even though the absolute differences between the BSS methods are within 1 percentage point (ppt).

Compared to the previous strategy (see Section 3.3), the computational effort is much higher. This is especially true for training the SVM, due to the increased size of the feature vector. Therefore, we evaluate another strategy in the next Section, where the size of the feature vector stays the same instead of being doubled.

3.5 Foreground Enhancement Evaluation

In this Section we present the results of the third strategy to improve VD. In order to enhance the vocals (i.e. increase the SNR), we remix the separated foreground with the original signal. The mixes were made with different levels of the separated track, ranging from -6 dB to 6 dB in 3 dB steps. The results from 0 dB indicate, that the remix was done without any gain changes.

Training as well as testing was done by using the features extracted from the remixed signals. Again, we include the results when using the real vocals also. In Table 3 we can see that different gain changes for remixing do not make a big difference for the results, regardless of the BSS method, classifier and feature set, except when using the real vocals (rows VOC). However, only the feature set from [9] allows for results at least as good as for the previous experiment in Section 3.4. Since those results are achieved without the additional computational burden due to a two-fold feature extraction, and the increased size of the feature vector, the enhancing-by-remixing strategy seems to be the best choice.

Internal Data Set (framesize=200ms)

	RF				SVM			
	accuracy		F-measure		accuracy		F-measure	
	IC14	OS11	IC14	OS11	IC14	OS11	IC14	OS11
MIX	.837	.795	.846	.814	.855	.807	.863	.819
MIX+VOC	.960	.985	.962	.986	.976	.984	.977	.985
MIX+aREPET	.845	.800	.853	.817	.865	.825	.872	.834
MIX+FASST	.842	.798	.850	.816	.863	.825	.871	.835
MIX+KAM	.844	.800	.853	.815	.871	.830	.877	.839
MIX+RPCA	.850	.806	.858	.822	.870	.833	.877	.841

Table 2. Results of Foreground Concatenation. MIX: trained and tested with mixed audio. For training and testing of the methods aREPET, FASST, KAM, and RPCA, the classifier is given a double-sized vector containing the features from the mixed and the separated audio signal. MIX+VOC: concatenating features from the real vocals to simulate perfect separation.

Again, RPCA based results are slightly better compared to the other source separation methods.

3.6 Final Method

Considering the results from the previous experiments, we choose the following setting for the upcoming experiments: For source separation, we choose the RPCA method, and we use the result to enhance the singing voice by remixing it with the original signal with an increased gain of 6 dB. The 116-attribute feature set IC14 as suggested in [9] is used, and fed to a SVM classifier with a radial basis function (RBF) kernel ($C = 2, \gamma = 0.35$). The remixed audios are used both for training and testing.

A very simple post-processing, where we use a median filter (order=5) for majority voting is also applied, which improves the results slightly from 87.3% to 87.8% accuracy.

3.7 Results on Public Data Sets

In this Section we compare the results of our suggested method as described in Section 3.6 to previously published results.

3.7.1 Jamendo

In [19], the authors presented results on a precisely defined split of the Jamendo data set, where the training set comprises 61 songs, and validation and test sets comprise 16 songs each. This allows for a fair comparison.

Table 4 lists the results reported by Lehner et al. in [9], compared with our new method. While the untouched output of the classifier (col. NEW) is on par with the (post-processed) baseline (col. LEH), the simple post-processing

Internal Data Set (framesize=200ms)

	RF				SVM			
	accuracy		F-measure		accuracy		F-measure	
	IC14	OS11	IC14	OS11	IC14	OS11	IC14	OS11
MIX	.837	.795	.846	.814	.855	.807	.863	.819
VOC $-6dB$	.880	.861	.886	.868	.907	.869	.911	.874
VOC $-3dB$	.895	.883	.900	.888	.922	.888	.925	.892
VOC $0dB$	.909	.905	.914	.907	.937	.907	.939	.911
VOC $3dB$	.923	.926	.926	.927	.949	.927	.951	.929
VOC $6dB$	.937	.943	.940	.944	.960	.945	.961	.946
aREPET $-6dB$	.844	.792	.852	.809	.862	.807	.869	.818
aREPET $-3dB$	.843	.792	.852	.809	.863	.807	.871	.818
aREPET $0dB$	.845	.794	.853	.811	.865	.809	.872	.820
aREPET $3dB$	.845	.795	.854	.811	.866	.812	.873	.822
aREPET $6dB$	.845	.799	.854	.813	.867	.813	.874	.823
FASST $-6dB$	.844	.795	.852	.812	.861	.805	.868	.817
FASST $-3dB$	.842	.796	.851	.813	.862	.807	.870	.819
FASST $0dB$	.844	.799	.852	.816	.864	.809	.871	.820
FASST $3dB$	.843	.799	.851	.816	.865	.811	.872	.822
FASST $6dB$	.844	.799	.852	.815	.864	.811	.871	.822
KAM $-6dB$	.845	.801	.854	.816	.866	.815	.873	.825
KAM $-3dB$	.846	.803	.855	.817	.868	.818	.874	.827
KAM $0dB$	.847	.804	.855	.817	.868	.819	.874	.828
KAM $3dB$	.846	.803	.855	.816	.870	.820	.875	.829
KAM $6dB$	.845	.803	.854	.816	.870	.821	.876	.829
RPCA $-6dB$	.847	.803	.855	.817	.868	.817	.874	.826
RPCA $-3dB$	.848	.805	.856	.819	.870	.818	.876	.827
RPCA $0dB$	.851	.806	.858	.819	.871	.819	.877	.828
RPCA $3dB$	.850	.807	.858	.819	.872	.820	.877	.829
RPCA $6dB$	.850	.809	.858	.821	.873	.821	.878	.829

Table 3. Results of Foreground Enhancement. MIX: trained and tested with mixed audio. For training and testing of the methods aREPET, FASST, KAM, and RPCA, the classifier is given the features extracted from a signal, where the separated vocals are remixed with the original audio signal. VOC: using the real vocals instead of the separated.

	LEH	NEW	NEW+
accuracy	.882	.882	.896
recall	.862	.873	.892
precision	.880	.872	.884
F-measure	.871	.873	.888

Table 4. Jamendo corpus results. LEH: results reported in [9]. NEW: our new classifier (SVM) with RPCA based vocal enhancement. NEW+: incl. post-processing with median filter.

(see Section 3.6) helps to reach better results, with an accuracy of 89.6% (col. NEW+).

3.7.2 RWC

In [13], Mauch et al. report 87.2% accuracy with a 5-fold cross validation (CV) on a 102 song data set that is composed of 90 songs from the RWC music database [5], and 12 additional songs. Since we had just access to the 100 RWC songs, our results are only comparable to a certain extent. Therefore, we also include the (post-processed) results reported from Lehner et al. in [9] (col. LEH), where we could use exactly the same splits for the 5-fold CV.

In Table 5 we can see an improvement of 2.3 ppt accuracy by comparing LEH and NEW (87.5% vs. 89.8%), despite the lack of any post-processing. The post-processing (col. NEW+) did not improve the accuracy on this data set. However, the increased recall (0.928 vs. 0.939) could still be desired for certain use cases, even when it comes with reduced precision (0.905 vs. 0.898).

	MODE	MAUCH	MODE	LEH	NEW	NEW+
accuracy	.654	.872	.604	.875	.898	.898
recall	1.00	.921	1.00	.926	.928	.939
precision	.654	.887	.604	.875	.905	.898
F-measure	.791	.904	.753	.900	.917	.918

Table 5. Results on the RWC data set. MAUCH: results reported in [13]. NEW: our new classifier (SVM) with RPCA based vocal enhancement. NEW+: incl. post-processing with median filter. LEH and NEW were trained on the 100 RWC songs, MAUCH on 90 RWC + 12 additional (unknown) songs. MODE: baseline achievable by always predicting the majority class (*vocals*); MODE of classification accuracy thus tells the percentage of vocals in the data set.

4. IMPROVING BACKGROUND AND VOCAL ESTIMATES

In this Section, we discuss the results of BSS algorithms in more detail regarding the amount of non-vocal artifacts in the estimated vocals, and vocal artifacts in the estimated background.

All of the four presented BSS methods have one characteristic in common: they do not incorporate VD results. In [18], the authors even state that their REPET method does not require any explicit handling of singing voice segments. Although, by listening to the results of all presented BSS algorithms in this paper, we believe there is nevertheless room for improvement. Our internal data set contains a lot of instrumental soli, played by a guitarist. Considering the basic principle of e.g. the REPET method, it comes as no surprise that the estimates of the vocals have passages containing those solo instruments only, and no vocals whatsoever. This is especially troublesome for use cases like artist recognition. On the other hand, the estimates of the instrumental background often contain artifacts from the singing voice. This is problematic for tasks like automatic karaoke track creation.

In [2, 16], the vocal frames were already successfully used to improve the results of the source separation, but according to the annotated ground truth, and not to an automatic classification. Therefore, we investigate the impact of VD on the results of the BSS with respect to metrics traditionally used to evaluate BSS algorithms. Even though we consider only RPCA henceforth, the remaining three BSS methods show a very similar characteristic in that matter. Concerning the VD, we use the one improved by RPCA as described in Section 3.6.

We suggest a simple post-processing strategy to improve the estimates: Regarding the estimated vocals, we simply filter out (i.e. mute) the non-vocal frames. In other words, for the final estimates of the separated vocals, we decide whether to use the vocal estimates from RPCA or silence – according to our VD.

Figure 1 illustrates this principle, where we can see in the upper plot a time signal of vocals (dark) embedded in the mixture (bright). The lower plot shows the estimated

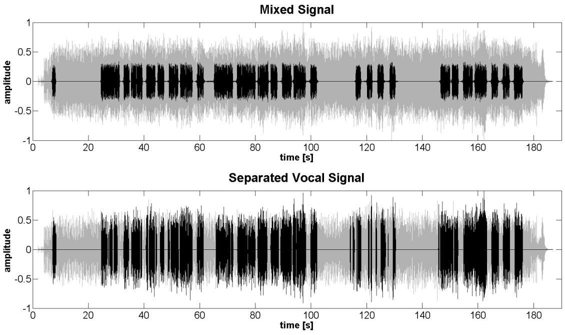

Figure 1. Example of RPCA separated singing voice. In the upper subplot we can see the mixed signal (bright) and the embedded singing voice (dark). In the lower subplot we can see the result from RPCA (bright) and the same result combined with singing voice detection (dark). Clearly, the latter approach is closer to the true singing voice.

vocals (bright) from the RPCA method, and the vocals after the VD based post-processing (dark). Obviously, the amount of non-vocal artifacts in the vocal estimate is reduced by applying this simple post-processing.

The same principle is applied in order to improve the estimates of the background. Here, we decide for the final estimates, whether to use the separated background or the original mix. This means, the separated background is only chosen, where the VD classifies the audio signal as vocal.

Nevertheless, it is not certain, if metrics traditionally used to evaluate BSS algorithms also reflect any improvement. A recall of vocals below an unknown – depending on the current situation – threshold would cause too much of the vocal estimates to be muted. At the same time, the estimated background would suffer from too much presence of vocals, since we would often wrongly opt for the original mixture instead of the separated background. Therefore, a thorough evaluation of the aforementioned post-processing is necessary in order to shed light on how useful it actually is.

4.1 Evaluation Metrics

In order to get meaningful evaluation results, we use the measurements proposed by Gribonval et al. in [6], where the overall estimation error is decomposed into *target distortion*, *interference*, and *artifacts*. Based on this components, the following energy ratios are defined: source Image to Spatial distortion Ratio (ISR), Source to Interference Ratio (SIR), and Source to Artifacts Ratio (SAR). Source to Distortion Ratio (SDR) is based on the three aforementioned measures, and serves as a global measure of distortion. For all metrics applies, that higher values indicate better performance.

Additionally, a set of measures that was proposed by Emiya et al. in [4] is used. Compared to the previously presented set, they better correlate with the perceived audio quality judged by human listeners. The overall distortion is also decomposed into the same three components, and based on them, the following measures are defined: Target-related Perceptual Score (TPS), Interference-related Perceptual Score (IPS), and Artifacts-related Perceptual Score (APS). The Overall Perceptual Score (OPS) is based on the three aforementioned scores, and serves as a global measure of perceived audio quality. Similar to the aforementioned metrics, higher values indicate better performance. All measures were extracted with the PEASS toolkit [4].

4.2 Evaluation Results

In this Section, we present box plots of the evaluation results on our internal data set (see Section 3.1) regarding the VD based post-processing method, which we described in Section 4. Audio examples are available at `http://www.cp.jku.at/misc/ismir2015bss`.

4.2.1 Background

In Figure 2, we can see the evaluation results of the background, separated with RPCA. For each metric, we can see three results: raw RPCA output (*A*), RPCA output post-processed with VD (*B*), and RPCA output post-processed with ground-truth annotations (*C*). By adding the results from a ground-truth based post-processing, we assess the potential benefit of the suggested post-processing method, and how far away we are from this optimum.

Compared to the raw RPCA outputs *A*, the post-processed results *B* and *C* improve for all metrics, except IPS. The median of the global measure of distortion (SDR) improves by 2 dB for post-processing *B*, and 1.9 dB for post-processing *C* (*A*: 1.3 dB, *B*: 3.3 dB, *C*: 3.2 dB). This suggests, that our VD performs on par with using ground-truth.

The median of the global measure of perceived audio quality (OPS) improves by 6.5 points for post-processing *B*, and 8.3 points for post-processing *C* (*A*: 26.5, *B*: 33.0, *C*: 34.8). Even though the median OPS is approximately the same for post-processing *B* and *C*, we can see still room for improvement, since the distribution of the ground-truth based results *C* has a tendency towards higher values.

Interestingly, compared to the raw RPCA output *A*, the median of the IPS results drops for both post-processing methods. For the VD based results *B*, we assume, this is due to some missed vocals, where the original mix is chosen instead of the separated background. This causes the vocals to be moved back into the final background estimation, and deteriorates the result. For the ground-truth based results *C* we assume, this is due to the fact, that the vocal track that we use for evaluation, contains not complete silence, but rather some noise. But our final estimation of the vocals replaces non-vocal segments with silence.

Based on the results, we consider it useful to incorporate VD in order to yield better estimations of the background. This could be especially useful for generating karaoke tracks, where for the non-vocal segments the original mixture can be used, without any loss in quality due to BSS characteristics. Obviously, for songs with high vocal content, the impact will be rather small.

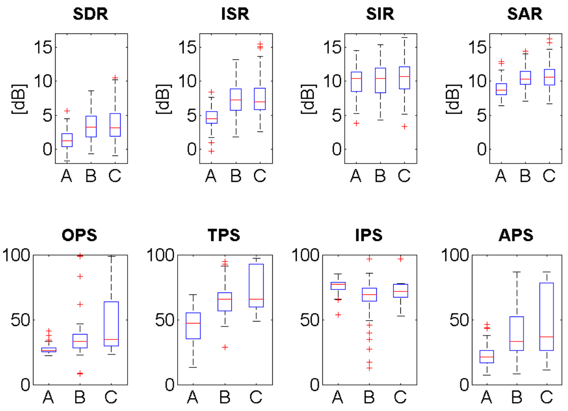

Figure 2. RPCA background estimation evaluation results. A: raw RPCA output; B: VD post-processed output; C: post-processed using ground truth. Higher values indicate better performance. In general, the performance increases for all metrics, except IPS. We assume, this is due to some missed vocals from our VD, where the original mix (incl. vocals) is chosen instead of the separated background.

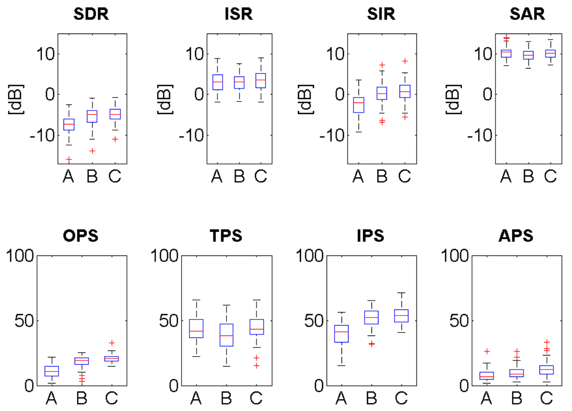

Figure 3. RPCA vocal estimation evaluation results. A: raw RPCA output; B: VD post-processed output; C: post-processed using ground truth. The global measures SDR and OPS indicate better performance for the post-processed output. The higher performance regarding interferences SIR and IPS are caused by the parts, that are muted, when our VD classifies them as non-vocal.

4.2.2 Vocals

In Figure 3, we can see the evaluation results of the vocals, separated with RPCA. Compared to the raw RPCA outputs *A*, the median of SDR indicate better performance for the post-processed output *B* (-7.2 dB vs. -4.9 dB), and no improvement comparing post-processing *B* to *C*.

The impact of silencing all non-vocal segments for the final vocal estimates can be seen in the interference related SIR (*A*: -2.0 dB, *B*: 0.2 dB, *C*: 0.6 dB). The perceptually motivated IPS reveals this relationship even better, where we can see an improvement of 11.2 points for post-processing *B* and 12.5 points for post-processing *C* (*A*: 41.2, *B*: 52.4, *C*: 53.7).

The median of the OPS improves by 8.3 points for post-processing *B*, and 9.7 points for post-processing *C* (*A*: 10.9, *B*: 19.2, *C*: 20.6).

Similar to the background estimates, the results of the metrics indicate improvement, when VD based post-processing is applied. Especially for tasks like artist recognition it could be useful to only use the parts which are classified as vocals, even when some are missed by the VD.

5. CONCLUSION AND OUTLOOK

In this paper we first presented the outcome of three strategies of utilising different monaural BSS techniques to improve VD: foreground separation, foreground concatenation, and foreground enhancement. According to the results on an internal data set, foreground enhancement is the best strategy. The difference of the usefulness between the four techniques aREPET, FASST, KAM, and RPCA is relatively small, and the latter usually performs best. We compared the results achieved with the best approach on publicly available data sets, and could show an improvement of 2.3 ppt relative to the baseline, reaching an accuracy of 89.8% on the RWC data set. Compared to the same baseline, the results on the Jamendo data set have also improved by 1.4 ppt, with an accuracy of 89.6%. However, approximately half of the improvement is due to using a SVM instead of a Random Forest. Depending on the use case, the effort of employing a BSS might therefore not always be justified. Nevertheless, by adding the results obtained by using the real vocals, we could show that VD would principally benefit from better separation results.

Our second contribution addressed the issue, that all of the four separation techniques produce vocal estimates, where many segments contain only instrumental background, and no singing voice at all. We suggested to use the results of the VD to simply mute the non-singing parts. Regarding the vocal estimates, we could see an improvement of 2.3 dB SDR when applying this post-processing (-7.2 dB vs. -4.9 dB).

For the final background estimates, we suggested to use the original mixed audio signal, where the VD classifies the signal as non-vocal. Regarding the background estimates, we could see an improvement of 2.0 dB SDR when applying this post-processing (1.3 dB vs. 3.3 dB).

We think it is safe to conclude that VD based post-processing improves the results of BSS vocal and background estimates, although not by much regarding traditional evaluation metrics. However, in the context of vocalist recognition, it could be helpful to only use the classified vocal parts, especially when solo instruments like guitars cause the BSS algorithm to produce lots of non-vocal artifacts in the vocal estimates. As one of the next steps, we plan to investigate the usefulness of our approach in this topic.

6. ACKNOWLEDGEMENTS

This research is supported by the Austrian Science Fund (FWF) under grants TRP307-N23 and Z159.

7. REFERENCES

[1] A. L. Berenzweig and D. P. W. Ellis. Locating singing voice segments within music signals. In *Workshop on the Applications of Signal Processing to Audio and Acoustics*, pages 119–122. IEEE, 2001.

[2] T-S Chan, T-C Yeh, Z-C Fan, H-W Chen, L. Su, Y-H Yang, and R. Jang. Vocal activity informed singing voice separation with the ikala dataset. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014)*. IEEE, 2014.

[3] J-L Durrieu, B. David, and G. Richard. A musically motivated mid-level representation for pitch estimation and musical audio source separation. *Journal of Selected Topics in Signal Processing*, 5(6):1180–1191, 2011.

[4] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann. Subjective and objective quality assessment of audio source separation. *Transactions on Audio, Speech, and Language Processing*, 19(7):2046–2057, 2011.

[5] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC music database: Popular, classical, and jazz music databases. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, volume 2, pages 287–288, 2002.

[6] R. Gribonval, L. Benaroya, E. Vincent, and C. Févotte. Proposals for performance measurement in source separation. In *4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pages 763–768, 2003.

[7] M. Hall, F. Eibe, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[8] P-S Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson. Singing-voice separation from monaural recordings using robust principal component analysis. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, pages 57–60. IEEE, 2012.

[9] B. Lehner, G. Widmer, and R. Sonnleitner. On the reduction of false positives in singing voice detection. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014)*, pages 7530–7534. IEEE, 2014.

[10] A. Liutkus, D. Fitzgerald, Z. Rafii, B. Pardo, and L. Daudet. Kernel additive models for source separation. *Transactions on Signal Processing*, 62(16):4298–4310, 2014.

[11] A. Liutkus, Z. Rafii, R. Badeau, B. Pardo, and G. Richard. Adaptive filtering for music/voice separation exploiting the repeating musical structure. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, pages 53–56. IEEE, 2012.

[12] A. Liutkus, Z. Rafii, B. Pardo, D. Fitzgerald, L. Daudet, et al. Kernel spectrogram models for source separation. In *4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, pages 6–10. IEEE, 2014.

[13] M. Mauch, H. Fujihara, K. Yoshii, and M. Goto. Timbre and Melody Features for the Recognition of Vocal Activity and Instrumental Solos in Polyphonic Music. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, pages 233–238, 2011.

[14] T. L. Nwe, A. Shenoy, and Y. Wang. Singing voice detection in popular music. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 324–327. ACM, 2004.

[15] A. Ozerov, E. Vincent, and F. Bimbot. A general flexible framework for the handling of prior information in audio source separation. *Transactions on Audio, Speech, and Language Processing*, 20(4):1118–1133, 2012.

[16] Z. Rafii and B. Pardo. A simple music/voice separation method based on the extraction of the repeating musical structure. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011)*, pages 221–224. IEEE, 2011.

[17] Z. Rafii and B. Pardo. Music/voice separation using the similarity matrix. In *Proceedings of the 13th International Conference on Music Information Retrieval (ISMIR 2012)*, pages 583–588, 2012.

[18] Z. Rafii and B. Pardo. Repeating pattern extraction technique (REPET): A simple method for music/voice separation. *Transactions on Audio, Speech, and Language Processing*, 21(1):73–84, 2013.

[19] M. Ramona, G. Richard, and B. David. Vocal detection in music with support vector machines. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2008)*, pages 1885–1888. IEEE, 2008.

[20] B. Schuller, A. Lehmann, F. Weninger, F. Eyben, and G. Rigoll. Blind enhancement of the rhythmic and harmonic sections by nmf: Does it help. In *Proceedings of the International Conference on Acoustics (NAG/DAGA 2009)*, pages 361–364, 2009.

[21] F. Weninger, J-L Durrieu, F. Eyben, G. Richard, and B. Schuller. Combining monaural source separation with long short-term memory for increased robustness in vocalist gender recognition. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011)*, pages 2196–2199. IEEE, 2011.

[22] F. Weninger, M. Wöllmer, and B. Schuller. Automatic assessment of singer traits in popular music: Gender, age, height and race. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, pages 37–42, 2011.