Audio Engineering Society

Convention Paper 10013

Presented at the 144th Convention
2018 May 23 – 26, Milan, Italy

This paper was peer-reviewed as a complete manuscript for presentation at this convention. This paper is available in the AES E-Library (http://www.aes.org/e-lib) all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Multi-track crosstalk reduction using spectral subtraction

Fabian Seipel[1] and Alexander Lerch[2]

[1]*Audio Communication Group, Technical University Berlin*
[2]*Center for Music Technology, Georgia Institute of Technology*

Correspondence should be addressed to Fabian Seipel (`fabian.seipel@gmail.com`)

ABSTRACT

While many music-related blind source separation methods focus on mono or stereo material, the detection and reduction of crosstalk in multi-track recordings is less researched. Crosstalk or 'bleed' of one recorded channel in another is a very common phenomenon in specific genres such as jazz and classical, where all instrumentalists are recorded simultaneously. We present an efficient algorithm that estimates the crosstalk amount in the spectral domain and applies spectral subtraction to remove it. Randomly generated artificial mixtures from various anechoic orchestral source material were employed to develop and evaluate the algorithm, which scores an average SIR-Gain result of 15.14 dB on various datasets with different amounts of simulated crosstalk.

1 Introduction

In many real-world music performance recordings, bands and ensembles are recorded without perfect acoustic separation. As a result, microphones which were intended to capture a particular instrument also record nearby signals; these additional signals are referred to as crosstalk, spill, or bleed. Although mixing practices allow to integrate crosstalk into the final mix, there are many instances where a better separation of these tracks is desirable for mixing. Other applications such as correctly annotating audio data with activation values, crosstalk suppression in speech audio, or collecting audio data for classification, could benefit from this approach as well.

Established blind source separation methods for music are typically focused on mono or stereo content while employing techniques based on factorization algorithms like NMF [1] [2], ICA or PCA [3], Hidden-Markov-Models [4], or spatial correlation [5]. More recent approaches also apply deep neural networks to train separation models [6] [7]. However, if there is single source material from recording scenes in form of multitrack data available, none of the methods mentioned above take advantage of this additional information since they are tailored for mono/stereo content.

The method presented in this paper focuses on cases where multi-track recordings of, e.g., classical ensembles are available. Clifford and Reiss [8] have investigated a method for crosstalk cancellation for multiple sources by using delay estimation and centered adaptive filters. Both Kokkinis et al. [9] and Prätzlich et al. [10] estimate the spectral power density of each voice and then apply a Wiener filter for crosstalk reduction.

In the proposed system, the crosstalk on a particular track is modeled as a weighted sum of the remaining tracks of this recording. The amount of crosstalk between each pair of tracks is estimated by minimizing a cost function based on spectral energy content through gradient descent with momentum [11]. As an alternative to Wiener filtering, the crosstalk is then removed through spectral subtraction [12].

For evaluation purposes, various datasets of anechoic orchestral multi-track recordings [13] are used to create artificial mixtures with different amounts of crosstalk (-18 dB, -12 dB and -6 dB). These artificial mixtures will be referred to as mixture tracks. Results are evaluated in two ways: by comparing the mixing matrix to the estimated spectral subtraction weights via correlation and by computing the standard blind source separation performance metrics SDR, SIR, and SAR [14].

2 Method

The main processing steps of the presented method are displayed in the flowchart in Fig. 1. First, a frequency domain representation of the multi-track data is computed by STFT. The crosstalk estimation algorithm models the crosstalk on a particular mixture track as a weighted sum of the other tracks. This estimation process employs an optimization technique based on gradient descent to minimize a spectral energy cost function. After the spectral subtraction, the crosstalk-reduced magnitude spectrogram is recombined with its original phase information to obtain the crosstalk-reduced audio data by inverse STFT.

Let $X(j,n,k)$ denote the matrix containing the magnitude spectrogram of the j-th mixture track, calculated with a Hamming window (framesize 4096 samples, hopsize 2048 samples) with k representing the frequency bin index and n the time frame. The analysis window length is approx. 85 ms (at 48 kHz). The rationale behind using comparably long analysis frames is that short time delays have less effect on the crosstalk reduction method and can therefore be neglected. $X_{\mathrm{mix},l}(n,k) = X(j=l,n,k)$ represents the spectrum of the l-th track with crosstalk.

2.1 Crosstalk estimation

This section outlines the process of estimating the amount of crosstalk from the multi-track data. For each target instrument l, the amount of crosstalk from all other tracks is estimated to derive the weighting factor $\lambda_{l,j}$ via gradient descent on a cost function $\Theta(\lambda)$ that aims to minimize the spectral crosstalk energy in the target mixture spectrum $X_{\mathrm{red},l}(n,k)$, summed over all time frames N as well as all frequency bins K:

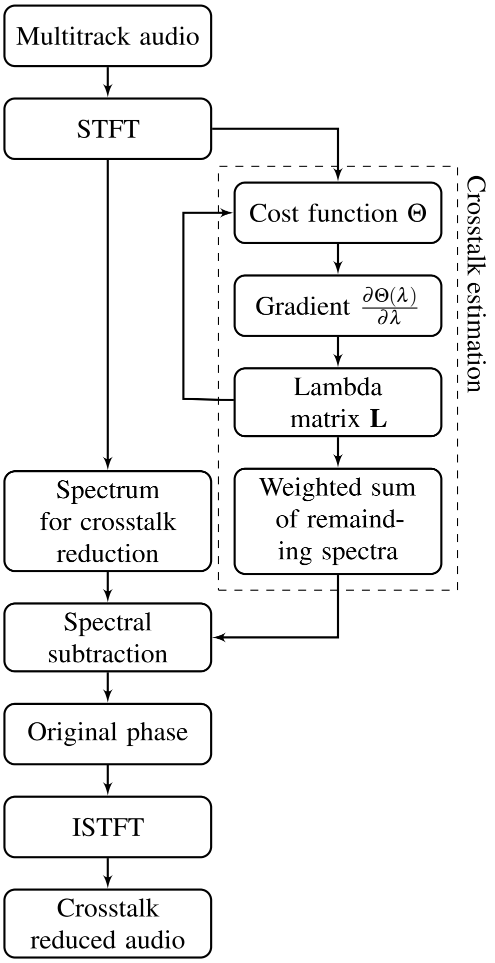

Fig. 1: Processing steps of the crosstalk reduction algorithm

$$\Theta(\lambda) = \frac{1}{N} \cdot \sum_{n=1}^{N} \sum_{k=1}^{K} \left[X_{\mathrm{mix},l}(n,k) - \sum_{j=1,\, j \neq l}^{J} \lambda_{l,j} \cdot X(j,n,k) \right]^2 \quad (1)$$

A correction factor $1/N$ accounts for different track lengths. The gradient for $\lambda_{l,j=i}(m)$ in iteration step m is then given by:

$$\frac{\partial\Theta(\lambda_{l,i}(m))}{\partial\lambda_{l,i}(m)} = -\frac{2}{N} \cdot \sum_{n=1}^{N} \sum_{k=1}^{K} \Big[X_{\mathrm{mix},l}(n,k) - \sum_{j=1,\, j \neq l}^{J} \lambda_{l,j}(m) \cdot X(j,n,k) \Big] \cdot X(i,n,k). \quad (2)$$

$\lambda_{l,j}$	bassoon	clarinet	bass	flute	f_horn	sopran	viola	violin	cello
bassoon	0	0.208	0	0.141	0.314	0.046	0.086	0.147	0.058
clarinet	0.182	0	0.07	0.119	0.18	0.065	0	0.105	0.072
bass	0	0.073	0	0	0.298	0.114	0.03	0.18	0.287
flute	0.065	0.062	0	0	0.204	0.017	0	0.112	0
f_horn	0.095	0.066	0.086	0.14	0	0.031	0.141	0	0.059
sopran	0.053	0.085	0.118	0.054	0.118	0	0.102	0.084	0.017
viola	0.059	0	0.022	0	0.4	0.078	0	0.2	0.157
violin	0.102	0.089	0.122	0.179	0	0.054	0.166	0	0.065
cello	0.047	0.062	0.185	0	0.135	0.011	0.128	0.063	0

Table 1: Example matrix for estimated $\lambda_{l,j}$ values, computed by the gradient descent algorithm

The update rule with momentum [11] is defined by:

$$\lambda_{l,i}(m+1) = \lambda_{l,i}(m) - \gamma(m) \cdot v(m+1). \tag{3}$$

The adaptation v is computed as:

$$v(m+1) = \beta \cdot v(m) + \frac{\partial \Theta(\lambda_{l,i}(m))}{\partial \lambda_{l,i}(m)} \tag{4}$$

with β as the momentum parameter, sometimes called friction, set to 0.8. The learning rate γ slightly decreases in each iteration step m:

$$\gamma(m+1) = 0.99 \cdot \gamma(m) \tag{5}$$

with an initial value of $\gamma(0) = 0.001$. Convergence is reached if the stepwise optimization of cost function $\Theta(\lambda(m))$ falls below a certain threshold δ:

$$\Theta(\lambda(m+1)) - \Theta(\lambda(m)) < \delta. \tag{6}$$

The gradient descent algorithm aims to find the $\lambda_{l,j}$ that guarantee the lowest overall power within the spectrum $X_{\mathrm{red},l}$ according to the cost function $\Theta(\lambda)$. By utilizing this approach, $\lambda_{l,j}$ adapts to the relative crosstalk amount of the different mixture tracks during the crosstalk estimation. All weighting factors $\lambda_{l,j}$ smaller than zero are automatically set to zero during the gradient descent process.

Table 1 shows an entire set of $\lambda_{l,j}$ values for a dataset with nine tracks. The first row displays all estimated weighting factors $\lambda_{\mathrm{bassoon},j}$ that have to be subtracted from the bassoon track to minimize crosstalk.

2.2 Crosstalk reduction

After estimating the weight factors for all the bleeding instruments for each track as outlined in Sect. 2.1, a sum of their weighted magnitude spectra can be subtracted from $X_{\mathrm{mix},l}(n,k)$ by simple spectral subtraction [12]. The result with reduced crosstalk $X_{\mathrm{red},l}(n,k)$ is therefore calculated as

$$X_{\mathrm{red,l}}(n,k) = X_{\mathrm{mix,l}}(n,k) - \sum_{j=1,\, j \neq l}^{J} \lambda_{l,j} \cdot X(j,n,k), \tag{7}$$

where J represents the total number of tracks. The weighting factor $\lambda_{l,j}$ is estimated from the spectrograms as explained below. If the subtraction results in negative spectrum values in $X_{\mathrm{red},l}(n,k)$, they will be set to zero. Finally, the reduced magnitude spectrum is combined with the original phase information from the STFT analysis to obtain the crosstalk-reduced audio file by inverse Fourier transform.

Figure 2 displays example results of the process described for one excerpt from the dataset. The left graphic shows the magnitude spectrogram of a single clean bassoon signal $X_{\mathrm{dry,bassoon}}$, the plot in the middle represents the magnitude spectrogram of the same signal with crosstalk $X_{\mathrm{mix,bassoon}}$, and the result with reduced crosstalk $X_{\mathrm{red,bassoon}}$ is shown on the right. All spectrograms are in logarithmic dB scale.

3 Evaluation

There exist no standardized datasets or evaluation methods for the tasks of crosstalk estimation and reduction for multi-track data. A dataset for these tasks should allow for full control over parameters in the mixing process such as the amount and combination of crosstalk. The multi-track recordings used to create the dataset

	bassoon	clarinet	bass	flute	f_horn	sopran	viola	violin	cello
bassoon_mix	1	0.097	0.051	0.085	0.088	0.066	0.074	0.097	0.026
clarinet_mix	0.097	1	0.094	0.062	0.073	0.06	0.031	0.073	0.07
bass_mix	0.051	0.094	1	0.024	0.103	0.101	0.084	0.117	0.079
flute_mix	0.085	0.062	0.024	1	0.111	0.029	0.011	0.073	0.004
f_horn_mix	0.088	0.073	0.103	0.111	1	0.063	0.079	0.002	0.077
sopran_mix	0.066	0.06	0.101	0.029	0.063	1	0.083	0.015	0.046
viola_mix	0.074	0.031	0.084	0.011	0.079	0.083	1	0.109	0.006
violin_mix	0.097	0.073	0.117	0.073	0.002	0.015	0.109	1	0.062
cello_mix	0.026	0.07	0.079	0.004	0.077	0.046	0.006	0.062	1

Table 2: Randomly generated mixing matrix for the -18 dB Mozart dataset, maximal values of 0.126

used in this study are from anechoic symphonic recordings.

Two different metrics are used to evaluate the two main processing blocks of the presented algorithm, respectively. First, the correlation coefficient is computed between the estimated lambda values and a transformed mixing matrix to evaluate the crosstalk estimation. Second, established blind source separation measures (SDR, SIR, and SAR, see below) are used on the audio results to evaluate the overall system.

3.1 Dataset

The dataset is created from excerpts from four orchestral anechoic multi-track recordings [13]:

- Beethoven: Symphony no. 7, I mov. (3:11 min):
  11 Parts: Flutes, Oboes, Clarinets, Bassoon, French horns, Trumpets, Timpani, Violin, Viola, Cello, Contrabass
- Bruckner: Symphony no. 8, II mov. (1:27 min):
  13 Parts: Flutes, Oboes, Clarinets, Bassoon, French horns, Trumpets, Trombones, Tuba, Timpani, Violin, Viola, Cello, Contrabass
- Mahler: Symphony no. 1, IV mov. (2:12 min):
  14 Parts: Flutes, Oboes, Clarinets, Bassoon, French horns, Trumpets, Trombones, Tuba, Timpani, Percussions, Violin, Viola, Cello, Contrabass
- Mozart: An aria of Donna Elvira from the opera Don Giovanni (3:47 min):
  9 Parts: Flute, Clarinet, Bassoon, French horns, Violin, Viola, Cello, Contrabass, Soprano

For each of these four pieces, three different mixture sets are constructed with a randomly generated mixing matrix. This mixing matrix has ones on the diagonal and positive values elsewhere that are limited to a defined maximum crosstalk value for the remaining elements. The three mixture sets have a different maximum crosstalk amount: -6 dB, -12 dB, and -18 dB, which relates to maximum mix factors of 0.5, 0.25, and 0.126, respectively. Table 2 shows an example mixing matrix (-18 dB dataset of the Mozart piece). Every row shows the contributions of each input track to one mixture track. The first row, for example, contains all fractions of the solo anechoic instrument tracks that are combined to the artificial bassoon mixture. A symmetric mixing matrix ensures that, for example, the scaling factor of the bassoon instrument on the clarinet mixture track equals the factor of the clarinet instrument on the bassoon mixture track. It is important to note that these mixing values are scaling factors that are independent of the individual track's acticity and loudness; thus, the actual amount of crosstalk is not necessarily reflected through the mixing matrix value.

Time delays are then calculated according to the reciprocal quadratic relation of distance and amplitude in the free field and accounted for in the mixing process. The reference amplitudes $A = 1$ (diagonal elements in the mixing matrix) correspond to a a distance of 1 m. Finally, all mixture tracks are normalized to the maximum amplitude of the loudest mixture to preserve the mixing matrix relations.

3.2 Correlation results

The mixing procedure may be represented as a system of linear equations $\mathbf{A}\mathbf{x} = \mathbf{b}$ in which $\mathbf{A}$ is the mixing

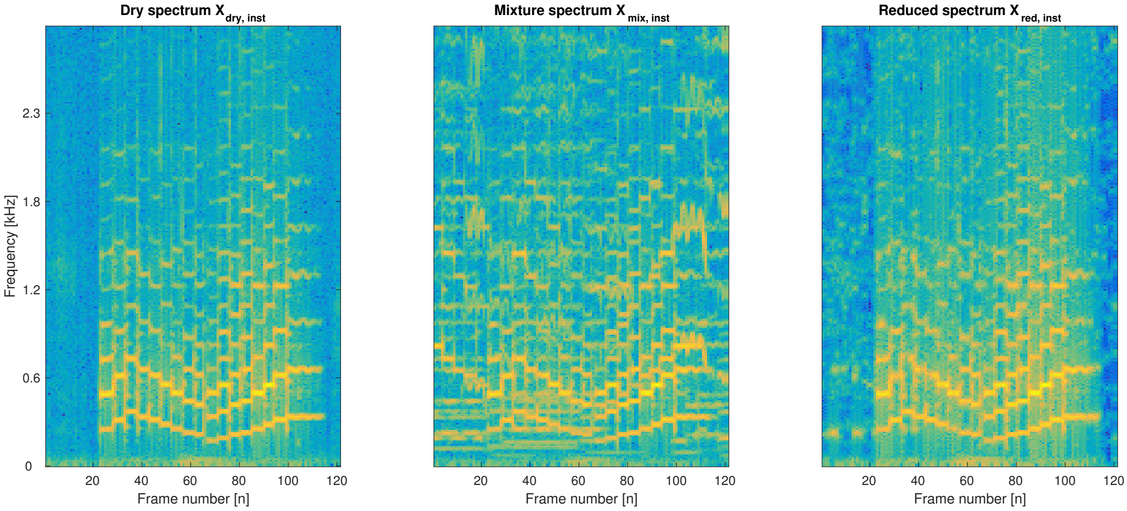

Fig. 2: Three different magnitude spectra excerpts from the unmixed original bassoon track (left), the artificially generated bassoon mixture with 6 dB crosstalk (middle) and the crosstalk reduced track (right)

matrix, $\mathbf{x}$ represents the vector of unmixed instrument tracks and $\mathbf{b}$ describes the mixture vector. Solving this equation for $\mathbf{x}$ leads to the de-mixing matrix $\mathbf{A}^{-1}$. The spectral subtraction Eq. (7) can be seen as related to this system of linear equations, where the $\lambda_{l,j}$ matrix (see Table 1) represents an estimation of this inverted matrix with flipped signs and empty diagonal. For the sake of concise notation, we will refer to $\lambda_{l,j}$ as $\mathbf{L}$ from now on. Similar to the mixing operation, we thus get a similar system of linear equations $\mathbf{x} \approx (-\mathbf{L}+\mathbf{I})\mathbf{b}$ and it follows that $\mathbf{A}^{-1} \approx -\mathbf{L}+\mathbf{I}$.

The matrices cannot be expected to be identical even in the best case as time delays were introduced when creating the dataset, however, the correlation between $\mathbf{A}^{-1}$ and $(-\mathbf{L}+\mathbf{I})$ should be high if the estimation works. Table 3 shows the correlation results for each of the three mixture sets, averaged over all four orchestral pieces. While the -12 dB and -18 dB sets both show very high correlation values of at least 0.9 over all four pieces with barely any variation, the -6 dB sets perform comparably bad for all pieces except the Mozart one. While the Mahler -6 dB set still shows a correlation of about 0.75, the two other sets show only correlation values between 0.6 and 0.437.

	-6dB	-12dB	-18dB
Beethoven	0.437	0.964	0.933
Bruckner	0.595	0.961	0.941
Mahler	0.752	0.931	0.900
Mozart	0.932	0.973	0.976
Average	0.678	0.957	0.937

Table 3: Correlation of $\mathbf{A}^{-1}$ and $(-\mathbf{L}+\mathbf{I})$

3.3 BSS Eval results

In order to evaluate the crosstalk suppression, the blind source separation (BSS) evaluation measures signal-to-distortion-ratio (SDR), signal-to-interference-ratio (SIR), and signal-to-artifacts-ratio (SAR) [14] were computed and investigated. These measures have become standard metrics for the evaluation of blind source separation systems, for example, in the SiSEC campaign[1]. Since the present approach processes multi-track data as opposed to most BSS methods which work with mono or stereo content, the following results cannot be compared directly to other studies. BSS evaluation metrics are highly dependent on the datasets used for separation (or crosstalk reduction). For this reason, a comparison of BSS Eval measures of the mixtures and the actual crosstalk reduced audio files seems more suitable to get a better insight of the algorithm performance. Table 4 shows the BSS Eval measures of mixtures and crosstalk reduced tracks as well as their

[1] http://sisec.inria.fr, last accessed March 6, 2018

	Mixture			Crosstalk-reduced			Difference		
	-6 dB	-12 dB	-18 dB	-6 dB	-12 dB	-18 dB	-6 dB	-12 dB	-18 dB
SDR	0.04	5.79	11.83	7.88	12.95	15.15	7.84	7.16	3.32
SIR	0.04	5.8	11.93	13.3	22.16	27.75	13.26	16.36	15.81
SAR	48.05	39.56	31.53	10.48	13.75	15.49	-37.57	-25.81	-16.05

Table 4: BSS evaluation measures for mixtures, crosstalk reduced tracks and their difference, sorted by dataset and averaged over all four orchestral pieces

difference.

SDR and SIR measures for the mixture tracks show very similar values, ranging from 0 dB to about 12 dB according to the intervals given by the datasets. The SAR values for those vary from 48 dB down to about 31.5 dB in equal distance. For both the -6 dB and -12 dB mixture sets the crosstalk reduced audio tracks show SDR improvements of about 7–8 dB while the -18 dB sets only gain about 3 dB. The difference in terms of SIR measures is very similar, all sets show improvements in the range of 13–16 dB. Changes regarding the SAR values depend more on the mixture set. While the SAR measure drops heavily for the -6 dB set, the -12 dB set shows a negative gain of about 26 dB and the -18 dB SAR value only decreases by 16 dB.

To investigate the variation of the BSS Eval measures for different music pieces, Table 5 displays the SDR/SIR difference of mixtures and crosstalk-reduced tracks as well as the absolute SAR scores for the crosstalk reduced results. SDR values increase nearly uniformly over all four orchestral pieces. While the increase for the -6 dB and -12 dB mixture sets ranges from 6 dB to 9 dB, the difference regarding the -18 dB mixture set only results in about 2–4 dB. SIR values are even more evenly distributed: most results lie in the interval of about 14–16 dB except the Bruckner/Beethoven values for the -6 dB dataset (10–11 dB) and all Mozart scores which are about 4 dB higher.

SAR values, in turn, quantify the musical noise in the crosstalk-reduced audio tracks. Similar to the SDR/SIR gain results, the metrics especially vary for the -6 dB mixture set, where the values range from 9 dB to about 12.5 dB. Results constantly increase for the mixture sets with lower crosstalk, although the improvement from the -6 dB to -12 dB sets is more distinct than from -12 dB to -18 dB. Again, the Mozart pieces have the best results being about 3 dB higher than the other pieces.

4 Discussion

The results of evaluation metrics show generally consistent trends. Very high correlation values of te demixing matrix $\mathbf{A}^{-1}$ and $(-\mathbf{L}+\mathbf{I})$ for both the -12 dB and -18 dB dataset for all pieces validate the success of the presented approach and prove that the gradient descent algorithm finds suitable $\lambda_{l,j}$ values minimizing the cost function. Multiple runs with random initialization further indicate that the detected cost minima are actually global minima.

There exist multiple reasons explaining the variation of the results between the different pieces. First, the total amount of crosstalk is not equal between the pieces. This is shown by the SIR measure of the mixtures given in Table 6 as these values represent the actual amount of crosstalk. While the increase over mixtures is consistently 6 dB as expected, the variation between pieces amounts to up to 4 dB. There are two reasons for that. First, the way the mixing matrix is generated means that the resulting amount of crosstalk depends on the number of instrument tracks. Second, the mixing matrix only signifies the factor but the resulting amount of crosstalk also depends on the track content and distribution; for example, percussion or timpani mixture tracks often show very low SIR and SAR since events only occur rarely during the track. Therefore, the optimal solution for the minimization of the cost function using the spectral energy criterion produces relatively high $\lambda_{l,j}$ values which in turn result in a harsh spectral subtraction and more musical artifacts. Those artifacts are quantified in the SAR score (see Table 4).

Crosstalk reduction for instruments that have overlapping frequency ranges with a similar tonal character is harder than for instruments with a unique spectral signature. Sections where the whole ensemble plays simultaneously are more difficult to manage for the algorithm than solo parts of individual instruments. In

	SDR gain			SIR gain			SAR absolute (results)		
	-6 dB	-12 dB	-18 dB	-6 dB	-12 dB	-18 dB	-6 dB	-12 dB	-18 dB
Beethoven	7.53	7.33	3.5	11.21	15.83	15.52	10.78	14.09	15.54
Bruckner	6.67	6.66	2.41	10.12	14.62	14.29	8.99	12.69	14.12
Mahler	8.04	7.33	3.46	14.66	15.39	14.99	9.71	12.24	14.54
Mozart	9.14	7.31	3.92	17.03	19.59	18.46	12.43	15.97	17.74

Table 5: SDR/SIR gain for each mixture set as well as absolute SAR values of the crosstalk reduced results

all cases, the Mozart piece achieves the best performance scores. The SAR values of the three remaining pieces range about 3 dB in comparison the the Mozart piece, each one following the above mentioned relation so that the absolute SIR values increase in 6 dB steps towards the -18 dB datasets (see Table 6).

In general, the crosstalk reduction method generates promising results. Whether the algorithm is suitable for a specific use case highly depends on the application. For tasks such as annotating audio data with instrument activations, the amount of separation (compare SIR) is crucial while the actual audio quality is irrelevant. The amount of subtraction can be controlled by scaling the $\lambda_{l,j}$ values, which can be an advantage in this scenario. In other cases, such as mixing software or more consumer-oriented products, the amount of musical artifacts can be decreased by utilizing an improved separation approach. Possible such improvements could include filtering approaches [15, 16] or other musical noise suppression techniques [17, 18] to reduce artifacts. Thresholding in the spectral or temporal domain could constitute another post-processing feasibility.

The generated and employed dataset contains numerous different instruments which makes the task more challenging. To explore the applicability for a broader field of possible applications, the algorithm needs to be tested on different musical genres, for example with the MedleyDB dataset [19] or the Mixing Secrets dataset [20].

	-6dB	-12dB	-18dB
Beethoven	0.03	5.92	11.73
Bruckner	-1.35	5.1	11.46
Mahler	-0.91	3.81	10.81
Mozart	2.39	8.39	13.73

Table 6: Absolute SIR values of the mixtures

5 Summary

The present study has introduced a new method for crosstalk reduction applied to multi-track data such as multi-microphone ensemble recordings. In a first step, this approach estimates the amount of crosstalk from a particular mixture track with a weighted sum of the remaining tracks by iteratively minimizing a spectral cost function with gradient descent. Second, this weighted sum of remaining instruments is subtracted from the mixture track to perform the reduction. Combining the resulting magnitude spectrum with its original phase information allows to obtain the crosstalk-reduced audio data via inverse STFT. In order to evaluate the algorithm, various mixtures with different amounts of crosstalk were artificially generated from multiple anechoic orchestral recordings. Results were evaluated in two ways: first, by correlating the mixing matrices and the resulting lambda matrices containing the estimated crosstalk factors to investigate the crosstalk estimation itself, and second by employing the standard blind source separation evaluation metrics SDR, SIR, and SAR to evaluate the suppression. Both evaluation metrics showed promising results. Post-processing techniques such as filtering or noise suppression could further improve the algorithm by reducing artifacts. For possible applications in a wider musical scope, tests with more genres are recommended.

References

[1] Lerch, A., *An introduction to audio content analysis: Applications in signal processing and music informatics*, John Wiley & Sons, 2012.

[2] Müller, M., *Information retrieval for music and motion*, volume 2, Springer, 2007.

[3] Comon, P. and Jutten, C., *Handbook of Blind Source Separation: Independent component analysis and applications*, Academic press, 2010.

[4] Mysore, G. J., Smaragdis, P., and Raj, B., "Non-negative hidden Markov modeling of audio with application to source separation," in *International Conference on Latent Variable Analysis and Signal Separation*, pp. 140–148, Springer, 2010.

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

[6] Chandna, P., Miron, M., Janer, J., and Gómez, E., "Monoaural audio source separation using deep convolutional neural networks," in *International Conference on Latent Variable Analysis and Signal Separation*, pp. 258–266, Springer, 2017.

[7] Uhlich, S., Giron, F., and Mitsufuji, Y., "Deep neural network based instrument extraction from music," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2135–2139, IEEE, 2015.

[8] Clifford, A. and Reiss, J. D., "Microphone interference reduction in live sound," in *Proceedings of the 14th International Conference on Digital Audio Effects (DAFx-11)*, 2011.

[9] Kokkinis, E. K., Reiss, J. D., and Mourjopoulos, J., "A Wiener Filter Approach to Microphone Leakage Reduction in Close-Microphone Applications," *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3), pp. 767–779, 2012.

[10] Prätzlich, T., Bittner, R. M., Liutkus, A., and Müller, M., "Kernel Additive Modeling for interference reduction in multi-channel music recordings," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 584–588, 2015.

[11] Sutskever, I., Martens, J., Dahl, G., and Hinton, G., "On the importance of initialization and momentum in deep learning," in *International conference on machine learning*, pp. 1139–1147, 2013.

[12] Boll, S., "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(2), pp. 113–120, 1979.

[13] Pätynen, J., Pulkki, V., and Lokki, T., "Anechoic recording system for symphony orchestra," *Acta Acustica united with Acustica*, 94(6), pp. 856–865, 2008.

[14] Vincent, E., Gribonval, R., and Févotte, C., "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4), pp. 1462–1469, 2006.

[15] Ephraim, Y. and Malah, D., "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(6), pp. 1109–1121, 1984.

[16] Lukin, A. and Todd, J., "Suppression of musical noise artifacts in audio noise reduction by adaptive 2-D filtering," in *Audio Engineering Society Convention 123*, Audio Engineering Society, 2007.

[17] Goh, Z., Tan, K.-C., and Tan, T., "Postprocessing method for suppressing musical noise generated by spectral subtraction," *IEEE Transactions on Speech and Audio Processing*, 6(3), pp. 287–292, 1998.

[18] Esch, T. and Vary, P., "Efficient musical noise suppression for speech enhancement system," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4409–4412, IEEE, 2009.

[19] Bittner, R. M., Salamon, J., Tierney, M., Mauch, M., Cannam, C., and Bello, J. P., "MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research." in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, volume 14, pp. 155–160, 2014.

[20] Gururani, S. and Lerch, A., "Mixing Secrets: A multitrack dataset for instrument detection in polyphonic music," in *Late Breaking Demo (Extended Abstract), Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Suzhou, 2017.