SOURCE-FILTER BASED CLUSTERING FOR MONAURAL BLIND SOURCE SEPARATION

Martin Spiertz,
Institut für Nachrichtentechnik,
RWTH Aachen University
Aachen, Germany
spiertz@ient.rwth-aachen.de

Volker Gnann,
Institut für Nachrichtentechnik,
RWTH Aachen University
Aachen, Germany
gnann@ient.rwth-aachen.de

ABSTRACT

In monaural blind audio source separation scenarios, a signal mixture is usually separated into more signals than active sources. Therefore it is necessary to group the separated signals to the final source estimations. Traditionally grouping methods are supervised and thus need a learning step on appropriate training data. In contrast, we discuss unsupervised clustering of the separated channels by Mel frequency cepstrum coefficients (MFCC). We show that replacing the decorrelation step of the MFCC by the non-negative matrix factorization improves the separation quality significantly. The algorithms have been evaluated on a large test set consisting of melodies played with different instruments, vocals, speech, and noise.

1. INTRODUCTION

Monaural blind source separation is a task of great importance for many applications. From music transcription to remixing, many audio applications require separated sources for further signal processing. Although in general monaural separation algorithms show lower performance than multichannel separation scenarios, they are of great interest. They can be applied to all separation scenarios, either because only monaural sources are available or as a first processing step for multi-channel source separation.
Separating different sources out of a monaural mixture is usually done in a time-frequency representation, e.g. by non-negative matrix factorization (NMF). NMF approximates a magnitude spectrogram by a sum of entry-wise, non-negative spectrograms, thus modeling the additive mixture of the signals. The basic iterative NMF algorithm introduced by [1] is further specialized to meet the requirements of audio source separation by [2] and [3]. If the algorithm separates the mixture into more channels than active sources, a clustering is needed. In [2] and [3] the channels are mapped onto the active sources with knowledge of the source signals, thus avoiding the blind clustering. In [4], the clustering problem is addressed by manual clustering.
In [5] a separation based on a source-filter model according to Figure 1 is introduced. Each note is modeled by a source signal corresponding to the harmonic structure of the note (pitch). This signal is filtered by an instrument-specific filter (resonance structure) to form the output signal. Because the source-filter model is incorporated in the separation algorithm, no clustering is needed. Good performance is shown for a test set of 40 mixtures. Motivated by this, we propose a clustering method based on source-filter modeling for a NMF separation method based on [2]. We will show that we reach results comparable to [5] but on a larger test set containing noise, pitched instruments, percussion, and speech/singing.

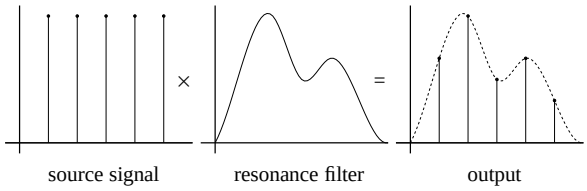

Figure 1: Source-Filter model in frequency domain.

The paper is structured as follows: In Section 2, the general concept of the proposed source separation is described. In Section 3, the new clustering strategy is introduced followed by the experimental setup and a discussion of the results in Section 4. Conclusive remarks are presented in Section 5.

2. FUNDAMENTALS

We assume a linear instantaneous mixture of M time-discrete input signals $s_m(n)$, $1 \leq m \leq M$. In this case, the mixing process can be modeled by a simple addition

$$x(n) = \sum_{m=1}^{M} s_m(n) \, . \tag{1}$$

In the following we will use underlined variables for complex-valued spectrograms. Dropping the underline is equivalent to taking the absolute value of the spectrogram. The mixture is transformed into the spectrogram $\underline{\mathbf{X}}$ by the short-time Fourier transform (STFT). Because the input signals $s_m(n)$ are real, the spectrogram is symmetric, and we can drop the part representing the negative frequency range. Therefore $\underline{\mathbf{X}}$ is a K-by-T matrix. T corresponds to the number of analysis frames transformed by the STFT and K is related to the length l_w of each analysis frame by $K = l_w/2+1$. For further details see also [6].

2.1. Non-negative matrix factorization

In the following, we will use the notation used by [2]. NMF as introduced by [1] approximates a non-negative, real-valued matrix $\mathbf{X}$ of size K-by-T by a product of two matrices $\mathbf{B}$ and $\mathbf{G}$

$$\mathbf{X} \approx \widetilde{\mathbf{X}} = \mathbf{BG} \, . \tag{2}$$

$\mathbf{B}$ is of size K-by-I and $\mathbf{G}$ is of size I-by-T with I as an user-defined parameter. The approximation can be done by a minimiza-

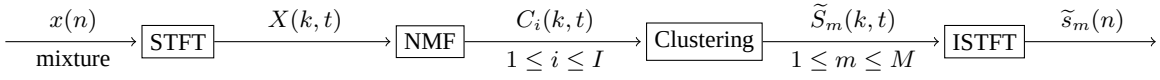

Figure 2: Signal flow of the proposed separation algorithm.

tion of the squared Euclidean distance

$$\|\mathbf{X} - \widetilde{\mathbf{X}}\|_2^2 \,, \tag{3}$$

or the divergence

$$\sum_{k,t} \mathbf{X}_{k,t} \log \frac{\mathbf{X}_{k,t}}{\widetilde{\mathbf{X}}_{k,t}} - \mathbf{X}_{k,t} + \widetilde{\mathbf{X}}_{k,t} \,. \tag{4}$$

After the matrices **B** and **G** are initialized with the absolute values of Gaussian noise, the multiplicative update rules

$$\mathbf{B} \leftarrow \mathbf{B}.\times \frac{\mathbf{X}\mathbf{G}^\mathbf{T}}{\mathbf{B}\mathbf{G}\mathbf{G}^\mathbf{T}} \text{ and} \tag{5}$$

$$\mathbf{G} \leftarrow \mathbf{G}.\times \frac{\mathbf{B}^\mathbf{T}\mathbf{X}}{\mathbf{B}^\mathbf{T}\mathbf{B}\mathbf{G}} \,, \tag{6}$$

minimize Equation 3, with $x.\times y$ and $\frac{x}{y}$ corresponding to element-wise multiplication and division. Equation 4 is minimized by

$$\mathbf{B} \leftarrow \mathbf{B}.\times \frac{\frac{\mathbf{X}}{\mathbf{B}\mathbf{G}}\mathbf{G}^\mathbf{T}}{\mathbf{1}\mathbf{G}^\mathbf{T}} \text{ and} \tag{7}$$

$$\mathbf{G} \leftarrow \mathbf{G}.\times \frac{\mathbf{B}^\mathbf{T}\frac{\mathbf{X}}{\mathbf{B}\mathbf{G}}}{\mathbf{B}^\mathbf{T}\mathbf{1}} \,, \tag{8}$$

with **1** corresponding to a K-by-T matrix containing only ones. As mentioned in [2], Equation 4 is more sensitive to small values in the case of large dynamic range than Equation 3.

2.2. Separation

The NMF separates the magnitude spectrogram **X** of the mixture into I channels with their corresponding spectrograms $\mathbf{C}_i$, $1 \leq i \leq I$. The motivation for using the NMF for blind source separation is the structure of pitched music in the spectrogram representation. A single note of a pitched instrument can be closely approximated by a constant frequency basis vector $\mathbf{B}_i$ and a time varying gain $\mathbf{G}_i$ corresponding to the envelope of the single note [6]. The i-th column of **B** and the i-th row of **G** can be multiplied to form the spectrogram $\mathbf{C}_i$ of the i-th channel

$$\mathbf{C}_i = \mathbf{B}_i\mathbf{G}_i \,, \tag{9}$$

with the $\mathbf{C}_i$ being of rank one. In [2], it is explained that the rows $\mathbf{G}_i$ of matrix **G** should have lowpass characteristics, due to the continuous nature of music. It is shown that an additional cost function c_t considering temporal continuity improves the separation quality for NMF algorithms in the case of music separation:

$$c_t = a \sum_i \frac{\sum_{t=2}^{T} \big(G(i,t) - G(i,t-1)\big)^2}{\sum_{t=1}^{T} G^2(i,t)} \,. \tag{10}$$

The only difference to [2] is the dropped factor T in the cost function, because we use mixtures of different lengths, see also Section 4. Thus the linear weighting a is set to a higher value: $a = 10^4$. In our work, we use the multiplicative update rules proposed there. **B** is updated according to Equation 7. **G** is updated by

$$\nabla c_r^+ = \mathbf{B}^\mathbf{T}\mathbf{1} \tag{11}$$

$$\nabla c_r^- = \mathbf{B}^\mathbf{T}.\times \frac{\mathbf{X}}{\mathbf{B}\mathbf{G}} \tag{12}$$

$$\nabla c_t^+(i,t) = 4a \frac{\mathbf{G}(i,t)}{\sum_{n=1}^{T} \mathbf{G}^2(i,n)} \tag{13}$$

$$\nabla c_t^-(i,t) = 2a \frac{\mathbf{G}(i,t-1) + \mathbf{G}(i,t+1)}{\sum_{n=1}^{T} \mathbf{G}^2(i,n)} + 2a \frac{\sum_{n=2}^{T} (\mathbf{G}(i,n) - \mathbf{G}(i,n-1))^2}{\left(\sum_{n=1}^{T} \mathbf{G}^2(i,n)\right)^2} \tag{14}$$

$$\mathbf{G} \leftarrow \mathbf{G}.\times \frac{\nabla c_r^- + \nabla c_t^-}{\nabla c_r^+ + \nabla c_t^+} \,. \tag{15}$$

For numerical stability we normalize the $\mathbf{B}_i$ and $\mathbf{G}_i$ after each iteration to ensure equal energy

$$A_i = \sqrt{\frac{\|\mathbf{G}_i\|_2}{\|\mathbf{B}_i\|_2}} \tag{16}$$

$$\mathbf{G}_i \leftarrow \mathbf{G}_i / A_i \tag{17}$$

$$\mathbf{B}_i \leftarrow \mathbf{B}_i A_i \,. \tag{18}$$

In case of over-separation ($I > M$), a clustering into M clusters has to be performed.

2.3. Signal Synthesis

We define clustering as a vector **a**, $1 \leq \mathbf{a}(i) \leq M$ with I elements and the mapping on the clusters by the Kronecker delta:

$$\delta_{m\mathbf{a}(i)} = \begin{cases} 1 & \text{if } m = \mathbf{a}(i) \,, \\ 0 & \text{otherwise.} \end{cases} \tag{19}$$

After clustering, we retrieve the complex-valued spectrograms $\underline{\widetilde{\mathbf{S}}}_m$ corresponding to source m by filtering the mixture spectrogram with the clustered channel spectrograms:

$$\underline{\widetilde{\mathbf{S}}}_m(k,t) = \underline{\mathbf{X}}(k,t) \frac{\sum_i \mathbf{C}_i(k,t)\delta_{m\mathbf{a}(i)}}{\sum_i \mathbf{C}_i(k,t)} \,. \tag{20}$$

Thus, signal parts that can not be approximated by spectrograms of rank one, are still present in the output of the separation algorithm. We assume no spectral overlapping of the sources so that we can use the phase of the mixture spectrogram $\underline{\mathbf{X}}$ as phase information for the separated spectrograms $\underline{\widetilde{\mathbf{S}}}_m$. Due to the modifications in the time-frequency representation it is not guaranteed that the single frames of the spectrogram $\underline{\widetilde{\mathbf{S}}}_m$ have smooth overlapping regions in the time domain. This can lead to audible distortions after the overlap-add procedure. To avoid these distortions, the square root of the Hann window is used as analysis and synthesis window for the STFT and the inverse STFT (ISTFT) [3].

2.4. Performance Measurement

Objective quality measures reflect human perception usually not in a good way. But for a large test set, listening tests for each mixture and each combination of parameters are very time-consuming. Therefore we use widely used quality measures for evaluation of separation quality. For their evaluation the knowledge of the input signals s_m and their corresponding magnitude spectrograms $\mathbf{S}_m$ is necessary. The first measure we use is the SER [2]

$$\text{SER}_m = 10\log_{10}\frac{\sum_{k,t}\mathbf{S}_m^2(k,t)}{\sum_{k,t}\big(\mathbf{S}_m(k,t)-\widetilde{\mathbf{S}}_m(k,t)\big)^2}\ \text{[dB]}. \quad (21)$$

For a more detailed discussion of the effects of the dynamic differences between the active sources we evaluate the ΔSER as the difference between the SER after separation and the SER before separation:

$$\Delta\text{SER}_m = 10\log_{10}\frac{\sum_{k,t}\big(\mathbf{S}_m(k,t)-\mathbf{X}(k,t)\big)^2}{\sum_{k,t}\big(\mathbf{S}_m(k,t)-\widetilde{\mathbf{S}}_m(k,t)\big)^2}\ \text{[dB]}. \quad (22)$$

Additionally, we evaluate the source-to-distortion ratio (SDR), the source-to-interference ratio (SIR) and the source-to-artifacts ratio (SAR) as proposed in [7]. If not otherwise mentioned, we use in the following the mean value over all separated sources as evaluation criterion, e.g.

$$\text{SER} = \frac{1}{M}\sum_{m=1}^{M}\text{SER}_m\ . \quad (23)$$

3. PROPOSED CLUSTERING ALGORITHMS

3.1. Reference Clustering

Reference clustering finds iteratively a clustering vector **a** which is a local optimum for the SER. First, **a** is initialized by

$$\mathbf{a}(i) = \arg\max_m \frac{\sum_{k,t}\mathbf{S}_m(k,t)\mathbf{C}_i(k,t)}{\sqrt{\sum_{k,t}\mathbf{S}_m^2(k,t)}\sqrt{\sum_{k,t}\mathbf{C}_i^2(k,t)}}\ . \quad (24)$$

After that, the corresponding SER is maximized by *hill-climbing* [8]. For each channel i, we define M neighboring cluster results by setting temporally $\mathbf{a}(i) = m$ and evaluating the corresponding $\text{SER}(m,i)$. This way we evaluate IM adjacent cluster results $\text{SER}(m,i)$. We define the highest adjacent SER as

$$\text{SER}_{\max} = \text{SER}(m_{\max}, i_{\max})\ , \quad (25)$$

with the corresponding channel $i_{\max}$ and target cluster $m_{\max}$. If $\text{SER}_{\max}$ is higher than the current SER we set

$$\mathbf{a}(i_{\max}) \leftarrow m_{\max}\ , \quad (26)$$

and the algorithm starts again with evaluating the new IM neighboring cluster results $\text{SER}(m,i)$. Otherwise the algorithm stops. Although it is not guaranteed that this algorithm finds the highest possible SER, we will use this simple clustering strategy as the ground truth for clustering.

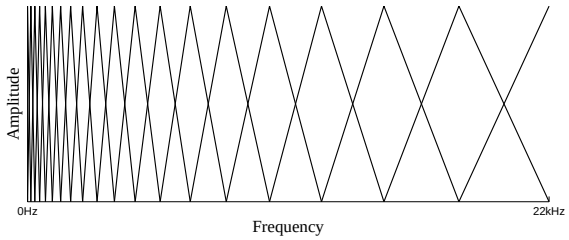

Figure 3: Weighting functions for Mel scale filtering.

3.2. Clustering by MFCC

In [5], [9] and [10], a source-filter model for sound generation is discussed, see Figure 1. According to this model, each frequency basis vector $\mathbf{B}_i$ of the mixture is a harmonic source/excitation signal $\mathbf{E}_i$ multiplied with an instrument-specific resonance filter $\mathbf{H}_m$ which is mainly responsible for formants. MFCC-based instrument classification use this instrument-specific weighting function. Although the calculation of MFCC is well-known, we explain the details in the following because these details are important for understanding the improvements applied to the general MFCC-based clustering algorithm.

The evaluation of the k-th Mel Frequency Cepstrum Coefficient $\text{mfcc}_i(k)$ corresponding to the channel i is done in three steps, see Figure 4. First the element-wise squared input vector $\mathbf{B}_i^2$ is filtered by a Mel filterbank with N_{Mel} filters to form the basis vectors $\mathbf{F}_i$ in Mel frequency domain. Each filter of the Mel filterbank is a triangular shaped weighting function with center frequencies being equidistant in Mel scale, see also Figure 3. According to [5], the filtering can be interpreted as a multiplication with a matrix $\mathbf{R}$ of size N_{Mel}-by-K with each row containing one Mel filter

$$\begin{aligned}\mathbf{F}_i &= \mathbf{R}\mathbf{B}_i^2 &(27)\\ &\approx \mathbf{R}\big[\mathbf{E}_i^2.\times\mathbf{H}_m^2\big] &(28)\\ &= \mathbf{R}\mathbf{E}_i^2.\times\mathbf{R}\mathbf{H}_m^2\ . &(29)\end{aligned}$$

After that, the logarithm is applied to the outputs of the Mel filterbank $\mathbf{F}_i(n)$, $1 \le n \le N_{\text{Mel}}$. Therefore, the order of magnitude between the outputs of the filterbank is reduced and the multiplication of the source signal with a spectral filter becomes an addition:

$$\log\big(\mathbf{F}_i+1\big) \approx \log\big(\mathbf{R}\mathbf{E}_i^2\big) + \log\big(\mathbf{R}\mathbf{H}_m^2\big)\ . \quad (30)$$

The offset $+1$ for the logarithm in Equation 30 leads to strictly positive logarithms. Therefore a continuous mapping of the values is guaranteed. If there is no offset, the logarithm maps filterbank outputs with very low amplitudes ($\mathbf{F}_i(n) \ll 1$) to large negative amplitudes. This causes unwanted influences in the following steps. To reduce the effect of this offset, we normalize all $\mathbf{F}_i(n)$ to a maximum amplitude of $A_{\max}$ before applying the logarithm:

$$\begin{aligned}\mathbf{F}_i &\leftarrow \frac{A_{\max}}{\max\{\mathbf{F}_i(n)\}}\mathbf{F}_i &(31)\\ &= c\mathbf{F}_i\ . &(32)\end{aligned}$$

This normalization has two effects on the logarithm. First, a constant offset is added, second, the non-linearity introduced by the offset is reduced:

$$\log(c\mathbf{F}_i+1) = \log(c) + \log(\mathbf{F}_i) + \log\big(1+\frac{1}{c\mathbf{F}_i}\big)\ . \quad (33)$$

The higher the factor $A_{\max}$ is set, the smaller is the non-linearity $\log\big(1+1/(c\mathbf{F}_i)\big)$. Unfortunately, the higher factor $A_{\max}$ also introduces a constant signal $\log(c)$. For higher $A_{\max}$ this constant signal overlays the wanted signal $\log(F_i)$ and therefore deteriorates the results of the following steps.
Finally the *Discrete Cosine Transform* (DCT) is applied for decorrelation of the both signals

$$\mathrm{mfcc}_i(k) = \sum_{n=1}^{N_{\mathrm{Mel}}} \log\big(c\mathbf{F}_i(n)+1\big)\cos\left(\frac{\pi(n-1/2)k}{N_{\mathrm{Mel}}}\right), \quad (34)$$

with $0 \le k \le N_{\mathrm{Mel}}-1$. According to the source-filter model, the $\mathbf{H}_m$ are assumed to have lowpass characteristics and the $\mathbf{E}_i$ have wideband characteristics. For this reason the DCT coefficients that correspond to high frequencies are dropped. The first coefficient represents mainly the signal energy and is therefore also dropped [9].
The $\mathrm{mfcc}_i(k)$ are then used as features for a *k-means clustering* [8]. First, we normalize each coefficient by subtracting the mean value and scaling the variance to unity:

$$\mathrm{mfcc}_i(k) \quad \leftarrow \quad \mathrm{mfcc}_i(k) - \frac{1}{I}\sum_i \mathrm{mfcc}_i(k) \quad (35)$$

$$\mathrm{mfcc}_i(k) \quad \leftarrow \quad \mathrm{mfcc}_i(k)\frac{1}{\sqrt{\sum_i \mathrm{mfcc}_i^2(k)}}\,. \quad (36)$$

After that, the vector **a** is initialized randomly. The *k-means clustering* then iteratively finds a clustering by evaluating the cluster center $\mathbf{center}_m$ and the new clustering vector **a**:

$$\mathbf{center}_m(k) \quad = \quad \frac{1}{\sum_i \delta_{m\mathbf{a}(i)}}\sum_i \mathrm{mfcc}_i(k)\delta_{m\mathbf{a}(i)} \quad (37)$$

$$\mathbf{a}(i) \quad = \quad \arg\min_m \sum_k \big(\mathbf{center}_m(k) - \mathrm{mfcc}_i(k)\big)^2 \quad (38)$$

The algorithm stops when the vector **a** does not change from one iteration to another.

3.3. Clustering by NMF

As mentioned in Section 3.2, the DCT decorrelates the two signals $\log(\mathbf{RE}_i^2)$ and $\log(\mathbf{RH}_m^2)$ of Equation 30. Unfortunately the DCT decorrelates the spectral envelope for each channel without utilizing knowledge of the other channels. Therefore, we rearrange the input of the decorrelation step in a matrix $\mathbf{Y}$ of size N_{Mel}-by-I

$$\mathbf{Y}(n,i) = \log\big(c\mathbf{F}_i(n)+1\big)\,. \quad (39)$$

Each column of $\mathbf{Y}$ consists approximately of an addition corresponding to Equation 30. Therefore it is generally possible to extract M basis functions $\mathbf{H}_m$ as constant parts in this matrix by the NMF algorithm, see also Figure 5. We initialize two matrices $\mathbf{W}$ of size N_{Mel}-by-M and $\mathbf{V}$ of size M-by-I with absolute values of Gaussian noise. After that, either the cost function in Equation 3 or the cost function in Equation 4 are minimized by the update rules introduced in Section 2.1. The algorithm stops after 100 iterations. Additionally the k-means clustering step is not required. Because of Equation 17 and 18, the clustering can simply be defined by

$$\mathbf{a}(i) = \arg\max_m \mathbf{V}(m,i)\,. \quad (40)$$

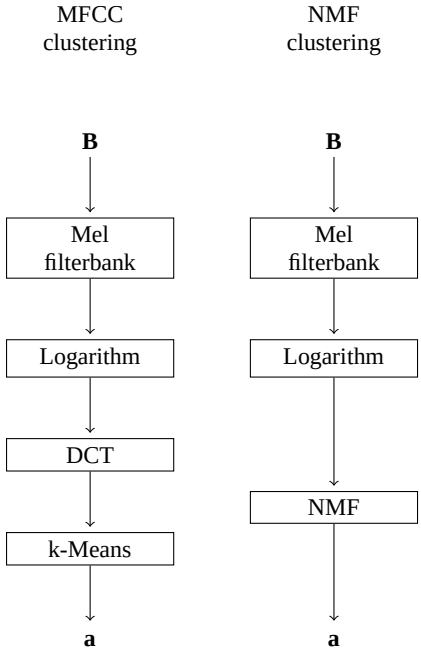

Figure 4: Signal flow of the proposed blind clustering algorithms.

3.4. Hierarchical Clustering

The clustering algorithms as proposed in Section 3.2 and 3.3 could directly be applied for any number M of target cluster. In the case of more than two active sources ($M > 2$), alternatively a hierarchical clustering strategy could be applied. In a first step we cluster all I channels into two clusters $\widetilde{m}$, $\widetilde{m} \in \{1,2\}$ by the clustering vector $\widetilde{\mathbf{a}}$, $\widetilde{\mathbf{a}}(i) \in \{1,2\}$. We define the estimated energy $\widetilde{E}_{\widetilde{m}}$ of the spectrograms of both clusters as

$$\widetilde{E}_{\widetilde{m}} = \sum_i \sum_{k,t} \mathbf{C}_i^2(k,t)\delta_{\widetilde{m}\widetilde{\mathbf{a}}(i)}\,. \quad (41)$$

If we assume uncorrelated sources, the energies of the different channels sum up to the energy of the mixture signal [3]. Further we assume that one cluster corresponds to one source, and the other cluster contains the remaining sources. Therefore we expect that the first separated source $\widetilde{s}_{m_1}$ corresponds to the cluster with lowest energy because the other cluster corresponds to more sources than one:

$$m_1 = \arg\min_{\widetilde{m}} \widetilde{E}_{\widetilde{m}}\,. \quad (42)$$

In the next iteration all remaining channels with $\delta_{m_1\widetilde{\mathbf{a}}(i)} = 0$ are clustered again into two clusters. The algorithm stops if the number of clusters equals the number of active sources. Hierarchical clustering could be used in combination with both clustering algorithms proposed in Section 3.2 and 3.3.

4. EXPERIMENTAL RESULTS

4.1. Test Set and Parameter Setting

The test set consists of all melodious phrases except the full organ, all singers except the quartet, the English and French female/male

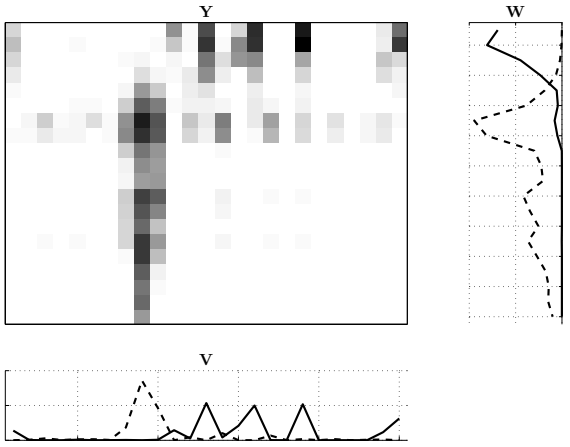

Figure 5: Decorrelation and clustering by NMF for a mixture of castanet and double bass. The matrix **Y** is separated into the matrix **W** with the 2 dominant resonance filters and their corresponding activity matrix **V**. The dashed lines correspond to the castanet, the solid lines to the double bass.

speech, and the pink noise from the Sound Quality Assessment Material of the EBU [11]. For adding more percussive instruments, the castanet, the roll of the side drum with snares, and the cymbal roll are included. The instrument classification scheme of [12] leads to 7 percussive instruments, 7 string instruments, 12 wind instruments, and 8 signals produced by humans. Additionally we add the bass, guitar, drums and keyboard of the BASS-dB [13]. As a last signal, Gaussian white noise is added to the test set. This is a total of 40 input signals of roughly 5 to 15 seconds length with a sampling frequency of 44.1 kHz and a resolution of 16 bit. In case of stereo signals, the right channel is dropped. The mixture is shortened to the length of the shortest input signal.

After clustering and signal synthesis, the separation quality is evaluated with knowledge of the input signals s_m using the measures SDR, SIR, SAR [7], and SER [2]. N_{Mel} is set to 20 for all experiments. In the case of MFCC clustering 9 coefficients remain as input signals for the k-means clustering step. The other coefficients are dropped according to Section 3.2.

The input signals are normalized to a defined dynamic difference: 0 dB, ± 3 dB, ± 6 dB, ± 10 dB, or ± 20 dB. After normalization we add the input signals according to Equation 1. The resulting mixture is transformed into a time-frequency representation by the STFT. The length of the analysis and synthesis windows of the STFT is 4096 samples with 50% overlap. The NMF algorithm separates the magnitude spectrogram into $I = 25$ channels with a maximum number of 300 iterations. In the following, the three proposed clustering algorithms are called P_{MFCC} (clustering by MFCC), $P_{\text{NMF,Div}}$ (clustering by NMF with divergence cost function), and $P_{\text{NMF,Euc}}$ (clustering by NMF with Euclidean cost function). For comparison, two other clustering methods are presented as lower and upper bound for clustering performance. P_{rand} corresponds to random clustering and P_{ref} to reference clustering.

Table 1: *Separation results for $M = 2$ with dynamic differences from 0 dB to ± 20 dB. The best clustering algorithm for each dynamic difference is marked bold. The results are shown in dB.*

	0 dB	± 3 dB	± 6 dB	± 10 dB	± 20 dB
SER					
P_{rand}	2.37	2.25	1.79	0.71	-3.62
P_{MFCC}	6.02	5.96	5.82	5.06	1.66
$P_{\text{NMF,Div}}$	6.89	7.02	7.08	**6.75**	**4.39**
$P_{\text{NMF,Euc}}$	**7.77**	**7.65**	**7.17**	5.95	0.52
P_{ref}	12.01	12.09	12.27	12.73	14.54
SDR					
P_{rand}	-0.82	-1.21	-2.04	-3.32	-7.73
P_{MFCC}	3.81	3.21	2.51	1.31	-2.46
$P_{\text{NMF,Div}}$	4.56	3.80	3.41	2.52	$-$**0.41**
$P_{\text{NMF,Euc}}$	**6.09**	**5.60**	**4.79**	**3.11**	-3.27
P_{ref}	10.98	11.02	11.10	11.31	11.54
SIR					
P_{rand}	4.51	4.12	3.54	3.16	2.57
P_{MFCC}	13.63	12.54	11.87	10.85	9.35
$P_{\text{NMF,Div}}$	15.42	13.87	13.40	12.49	**11.28**
$P_{\text{NMF,Euc}}$	**16.27**	**15.52**	**14.70**	**13.35**	9.31
P_{ref}	21.01	21.11	21.28	21.59	22.40
SAR					
P_{rand}	3.12	3.11	3.07	3.00	2.84
P_{MFCC}	6.54	6.37	6.21	5.73	4.53
$P_{\text{NMF,Div}}$	7.54	7.50	7.43	**7.11**	**6.19**
$P_{\text{NMF,Euc}}$	**8.07**	**7.90**	**7.49**	6.65	4.28
P_{ref}	11.82	11.87	12.03	12.48	14.25

4.2. Blind Source Separation with 2 Sources

In a first experiment we set the number of active sources $M = 2$. With 40 input signals, this corresponds to a total of 780 mixing scenarios. All of them are mixed at the 9 different dynamic differences as described in Section 4.1. The maximum amplitude of the Mel filter outputs is normalized to $A_{\text{max}} = 10^4$. The performance of the proposed clustering algorithms is shown in Table 1. We can make the following observations: The SER, SDR, SIR and SAR behave very similar so that we concentrate on the SER in the following. The NMF clustering with divergence outperforms the clustering with MFCC over the complete dynamic range for mixtures. Further, for nearly equal loud mixtures ($-6... + 6$ dB) the clustering $P_{\text{NMF,Euc}}$ leads to better results than $P_{\text{NMF,Div}}$. The most probable reason is that the divergence is easier distorted by small values than the squared Euclidean distance, see also Section 2.1. If the expected dynamic difference between two source signals is known, the appropriate clustering algorithm can be chosen to maximize the expected separation quality.

The dynamic differences have different effects on the quieter and the louder source, as shown in Table 2. The quieter source, here defined as $\widetilde{s}_1$, is separated with lower SER. This could be simply explained by the high energy of the interfering source $\widetilde{s}_2$. The ΔSER behaves in the opposite way for the same reason. For the quieter signal even few separated interfering parts of $\widetilde{s}_2$ leads to large improvements for the ΔSER because of the low energy of $\widetilde{s}_1$.

In Table 3 the influence of the normalization A_{max} of Equation 31 is shown for the clustering $P_{\text{NMF,Euc}}$. Similar results are obtained

Table 2: *SER and ΔSER for $M = 2$ with dynamic differences from 0 dB to ±20 dB. Results are shown in dB and $P_{NMF,Div}$ is used as clustering strategy. $\widetilde{s}_1$ is the quieter source, $\widetilde{s}_2$ the louder source.*

	0 dB	±3 dB	±6 dB	±10 dB	±20 dB
SER					
$\widetilde{s}_1$	6.89	5.48	4.03	1.70	−5.65
$\widetilde{s}_2$	6.89	8.56	10.13	11.79	14.42
mean	6.89	7.02	7.08	6.75	4.39
ΔSDR					
$\widetilde{s}_1$	6.55	8.20	9.81	11.54	14.28
$\widetilde{s}_2$	6.54	5.14	3.62	1.15	−6.63
mean	6.54	6.67	6.71	6.34	3.83

for the clustering algorithms P_{MFCC} and $P_{\text{NMF,Div}}$. As mentioned in Section 3.2, the normalization has critical influence on the hit probability of the clustering and therefore on the final separation quality. For large dynamic differences, the instrument filter $\mathbf{H}_m$ of the quieter instrument has much lower amplitudes than the louder instrument. In this case, the non-linearity of the offset (see Equation 33) has to be small for error free detection of both instrument filters. This leads to the requirement of large values for A_{max}. In the case of nearly equal loudness for both instruments, the influence of the constant signal in Equation 33 should be as low as possible, and therefore low values for A_{max} are preferable. For $A_{max} = 10^2$ and 0 dB dynamic difference, the separation quality decreases compared to the case with $A_{max} = 10^3$. This shows that a certain range of values is necessary for successful clustering. The authors of [5] reported for their best algorithm a SDR of roughly 8 dB, a SIR of roughly 22.5 dB, and a SAR of roughly 8.1 dB. We can see in Table 1 that the clustering $P_{\text{NMF,Euc}}$ for a dynamic difference of 0 dB results in a worse SIR and an identical SAR. Therefore the SDR is slightly worse, because it evaluates the overall distortion by interferences and artifacts [7]. This could be partly confirmed by our significant larger test set. Unfortunately in [5] dynamic differences for the input signals are all set to zero, although our results show, that dynamic differences have significant influence on source-filter based source separation. In opposite to [5], our proposed clustering algorithm can be adjusted to an expected dynamic difference by the parameter A_{max}. Furthermore, no additional information like lowest pitch of each instrument is necessary for our clustering algorithms. The additional complexity by a clustering as proposed in our separation scheme is very low compared to the separation by the NMF. In informal complexity tests evaluated on a small number of mixtures, the clustering is calculated in less than 0.2% of the time needed for the NMF[1]. Therefore the additional complexity is insignificant compared to the STFT, the NMF and the signal synthesis step.

4.3. Blind Source Separation with 3 Sources

In a second experiment, we set the number of active sources $M = 3$, the dynamic difference to 0 dB, and $A_{\text{max}} = 10^3$. With 40 input signals, this leads to a total of 9880 mixtures. Table 4 shows the results. First we discuss the mean value for all sources and all

[1] All algorithms are implemented in Matlab. Evaluation is done on a P4 with 3200 MHz.

Table 3: *Influence of the normalization factor A_{max} on the SER for different dynamic differences. The best normalization for each dynamic difference is marked bold. Results are shown in dB and $P_{NMF,Euc}$ is used as clustering strategy.*

A_{max}	0 dB	±3 dB	±6 dB	±10 dB	±20 dB
10^2	7.79	7.54	6.69	4.36	−2.44
10^3	**7.88**	7.60	6.95	5.41	−1.24
10^4	7.77	**7.65**	7.17	5.95	0.52
10^5	7.46	7.44	**7.21**	**6.25**	**1.92**

Table 4: *Mean SER in dB for $M = 3$. The mean value over all 3 sources is shown. Additionally the mean values are evaluated individually for the best source defined as $\widetilde{s}_{m_1}$ to the worst source $\widetilde{s}_{m_3}$.*

	$\widetilde{s}_{m_1}$	$\widetilde{s}_{m_2}$	$\widetilde{s}_{m_3}$	mean
P_{rand}	2.21	1.19	0.44	1.28
P_{MFCC}	6.10	2.66	1.63	3.46
$P_{\text{NMF,Div}}$	6.38	2.75	1.71	3.61
$P_{\text{NMF,Euc}}$	6.62	3.23	2.17	4.01
$P_{\text{MFCC,Hier}}$	6.07	3.01	1.94	3.67
$P_{\text{NMF,Div,Hier}}$	6.86	3.20	2.16	4.07
$P_{\text{NMF,Euc,Hier}}$	6.92	3.55	2.49	4.32
P_{ref}	9.83	7.36	6.45	7.88

mixtures. It can be seen that the mean separation quality of P_{ref} degrades by more than 4 dB for $M = 3$, compared with the same scenario with $M = 2$ (see also Table 1). The same degradation of 2.5 dB to 3.5 dB can be observed for all blind clustering algorithms, but the ranking of the different clustering algorithms for a dynamic difference of 0 dB remains the same.
In the following, we define the index of the estimated source $\widetilde{s}_m$ with highest SER as m_1, and the index corresponding to $\widetilde{s}_m$ with lowest SER as m_3. The remaining index is defined as m_2. For a more detailed analysis, the mean values are individually evaluated for $\widetilde{s}_{m_1}$, $\widetilde{s}_{m_2}$, and $\widetilde{s}_{m_3}$ over all mixtures. In general, only one of the three source estimations is separated with acceptable SER values. This motivates us to apply the hierarchical clustering described in Section 3.4 to the test set. For the clustering by MFCC $P_{\text{MFCC,Hier}}$, the separation quality for $\widetilde{s}_{m_2}$ and $\widetilde{s}_{m_3}$ is improved. For both NMF-based clustering methods, the separation quality for all three sources is increased. Again the ranking of the three proposed clustering methods remain the same.

5. CONCLUSIONS

In this paper, we introduced low-complexity clustering algorithms for monaural blind source separation based on NMF. We have shown the disadvantages of decorrelating each channel on its own for MFCC and circumvented this drawback by replacing the DCT by a NMF. We tested the proposed clustering algorithm on a large test set, so that we can consider the results reliable. Furthermore, we discuss the influence of dynamic differences between the input signals. Finally, we show, that even in the case of three active sources the algorithm is in general capable of separating at least

one source properly out of the mixture. We compare our algorithm with a separation algorithm that implements the source-filter model in the separation process, and show that our algorithm leads to comparable results, although it is evaluated on a larger test set. Another important advantage of our algorithm is that it is possible to adjust the clustering algorithm to an expected dynamic difference between the sources. The higher separation quality of the reference clustering shows that there is room for improvements regarding the clustering strategies.

6. REFERENCES

[1] D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2000, pp. 556–562.

[2] T. Virtanen, "Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[3] T. Virtanen, "Monaural sound source separation by perceptually weighted non-negative matrix factorization," Tech. Rep., Tampere University of Technology, Institute of Signal Processing, 2007.

[4] B. Wang and M. D. Plumbley, "Investigating single-channel audio source separation methods based on non-negative matrix factorization," in *Proceedings of the ICA Research Network International Workshop*, 2006, pp. 17–20.

[5] D. FitzGerald, M. Cranitch, and E. Coyle, "Extended nonnegative tensor factorisation models for musical sound source separation," *Computational Intelligence and Neuroscience*, 2008.

[6] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003, pp. 177–180.

[7] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[8] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall, 2 edition, 2003.

[9] A.B. Nielsen, S. Sigurdsson, L.K. Hansen, and J. Arenas-Garcia, "On the relevance of spectral features for instrument classification," in *Proc. IEEE Int. Conference on Acoustic Speech and Signal Processing ICASSP '*, Apr. 2007, vol. 2, pp. 485–488.

[10] Z. Duan, Y. Zhang, C. Zhang, and Z. Shi, "Unsupervised single-channel music source separation by average harmonic structure modeling," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 766–778, May 2008.

[11] EBU, "Sound Quality Assessment Material," 1988, http://www.ebu.ch/en/technical/publications/tech3000_series/tech3253/.

[12] N. H. Fletcher and T. D. Rossing, *The Physics of Musical Instruments*, Springer, 2nd edition, May 2008.

[13] E. Vincent, R. Gribonval, C. Fevotte, and al., "Bass-db: the blind audio source separation evaluation database," http://www.irisa.fr/metiss/BASS-dB/.