EXTENDING HARMONIC-PERCUSSIVE SEPARATION OF AUDIO SIGNALS

Jonathan Driedger[1], **Meinard Müller**[1], **Sascha Disch**[2]
[1]International Audio Laboratories Erlangen
[2]Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany
{jonathan.driedger,meinard.mueller}@audiolabs-erlangen.de, sascha.disch@iis.fraunhofer.de

ABSTRACT

In recent years, methods to decompose an audio signal into a harmonic and a percussive component have received a lot of interest and are frequently applied as a processing step in a variety of scenarios. One problem is that the computed components are often not of purely harmonic or percussive nature but also contain noise-like sounds that are neither clearly harmonic nor percussive. Furthermore, depending on the parameter settings, one often can observe a leakage of harmonic sounds into the percussive component and vice versa. In this paper we present two extensions to a state-of-the-art harmonic-percussive separation procedure to target these problems. First, we introduce a *separation factor* parameter into the decomposition process that allows for tightening separation results and for enforcing the components to be clearly harmonic or percussive. As second contribution, inspired by the classical sines+transients+noise (STN) audio model, this novel concept is exploited to add a third *residual* component to the decomposition which captures the sounds that lie *in between* the clearly harmonic and percussive sounds of the audio signal.

1. INTRODUCTION

The task of decomposing an audio signal into its harmonic and its percussive component has received large interest in recent years. This is mainly because for many applications it is useful to consider just the harmonic or the percussive portion of an input signal. Harmonic-percussive separation has been applied, for example, for audio remixing [9], improving the quality of chroma features [14], tempo estimation [6], or time-scale modification [2, 4]. Several decomposition algorithms have been proposed. In [3], the percussive component is modeled by detecting portions in the input signal which have a rather noisy phase behavior. The harmonic component is then computed by the difference of the original signal and the computed percussive component. In [10], the crucial observation is that harmonic sounds have a horizontal structure in a spectrogram representation of the input signal, while percussive sounds form vertical structures. By iteratively diffusing the spectrogram once in horizontal and once in vertical direction, the harmonic and percussive elements are enhanced, respectively. The two enhanced representations are then compared, and entries in the original spectral representation are assigned to either the harmonic or the percussive component according to the dominating enhanced spectrogram. Finally, the two components are transformed back to the time-domain. Following the same idea, Fitzgerald [5] replaces the diffusion step by a much simpler median filtering strategy, which turns out to yield similar results while having a much lower computational complexity.

A drawback of the aforementioned approaches is that the computed decompositions are often not very *tight* in the sense that the harmonic and percussive components may still contain some non-harmonic and non-percussive residues, respectively. This is mainly because of two reasons. First, sounds that are neither of clearly harmonic nor of clearly percussive nature such as applause, rain, or the sound of a heavily distorted guitar are often more or less

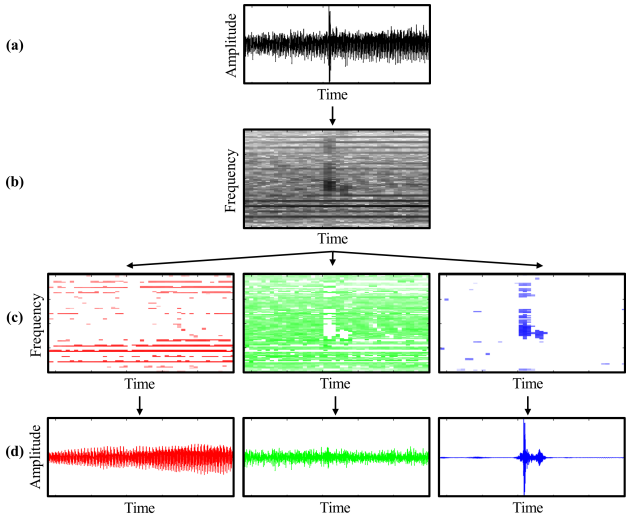

Figure 1. **(a):** Input audio signal x. **(b):** Spectrogram X. **(c):** Spectrogram of the harmonic component X_h (left), the residual component X_r (middle) and the percussive component X_p (right). **(d):** Waveforms of the harmonic component x_h (left), the residual component x_r (middle) and the percussive component x_p (right).

randomly distributed among the two components. Second, depending on the parameter setting, harmonic sounds often leak into the percussive component and the other way around. Finding suitable parameters which yield satisfactory results often involves a delicate trade-off between a leakage in one or the other direction.

In this paper, we propose two extensions to [5] that lead towards more flexible and refined decompositions. First, we introduce the concept of a *separation factor* (Section 2). This novel parameter allows for *tightening* decomposition results by enforcing the harmonic and percussive component to contain just the clearly harmonic and percussive sounds of the input signal, respectively, and therefore to attenuate the aforementioned problems. Second, we exploit this concept to add a third *residual* component that captures all sounds in the input audio signal which are neither clearly harmonic nor percussive (see Figure 1). This kind of decomposition is inspired by the classical *sines+transients+noise* (STN) audio model [8, 11] which aims at resynthesizing a given audio signal in terms of a parameterized set of sine waves, transient sounds, and shaped white noise. While a first methodology to compute such a decomposition follows rather straightforward from the concept of a separation factor, we also propose a more involved iterative decomposition procedure. Building on concepts proposed in [13], this procedure allows for a more refined adjustment of the decomposition results (Section 3.3). Finally, we evaluate our proposed procedures based on objective evaluation measures as well as subjective listening tests (Section 4). Note that this paper has an accompanying website [1] where you can find all audio examples discussed in this paper.

2. TIGHTENED HARMONIC-PERCUSSIVE SEPARATION

The first steps of our proposed decomposition procedure for tightening the harmonic and the percussive component are the same as in [5], which we now summarize. Given an input audio signal x, our goal is to compute a harmonic component x_h and a percussive component x_p such that x_h and x_p contain the clearly harmonic and percussive sounds of x, respectively. To achieve this goal, first a spectrogram X of the signal x is computed by applying a short-time Fourier transform (STFT)

$$X(t,k) = \sum_{n=0}^{N-1} w(n)\, x(n+tH)\, \exp(-2\pi i k n/N)$$

with $t \in [0:T-1]$ and $k \in [0:K]$, where T is the number of frames, $K = N/2$ is the frequency index corresponding to the Nyquist frequency, N is the frame size and length of the discrete Fourier transform, w is a sine-window function and H is the hopsize (we usually set $H = N/4$). A crucial observation is that looking at one frequency band in the magnitude spectrogram $Y = |X|$ (one row of Y), harmonic components stay rather constant, while percussive structures show up as peaks. Contrary, in one frame (one column of Y), percussive components tend to be equally distributed, while the harmonic components stand out. By applying a median filter to Y once in horizontal and once in vertical direction, we get a harmonically enhanced magnitude spectrogram $\tilde{Y}_h$ and a magnitude spectrogram $\tilde{Y}_p$ with enhanced percussive content

$$\begin{aligned} \tilde{Y}_h(t,k) &:= \text{median}(Y(t-\ell_h,k),\ldots,Y(t+\ell_h,k)) \\ \tilde{Y}_p(t,k) &:= \text{median}(Y(t,k-\ell_p),\ldots,Y(t,k+\ell_p)) \end{aligned}$$

for $\ell_h, \ell_p \in \mathbb{N}$ where $2\ell_h + 1$ and $2\ell_p + 1$ are the lengths of the median filters, respectively.

Now, extending [5], we introduce an additional parameter $\beta \in \mathbb{R}$, $\beta \geq 1$, called the *separation factor*. We assume an entry of the original spectrogram $X(t,k)$ to be part of the clearly harmonic or percussive component if $\tilde{Y}_h(t,k)/\tilde{Y}_p(t,k) > \beta$ or $\tilde{Y}_p(t,k)/\tilde{Y}_h(t,k) \geq \beta$, respectively. Intuitively, for a sound to be included in the harmonic component it is required to stand out from the percussive portion of the signal by at least a factor of β, and vice versa for the percussive component. Using this principle, we can define binary masks M_h and M_p

$$\begin{aligned} M_h(t,k) &:= \left(\tilde{Y}_h(t,k)/(\tilde{Y}_p(t,k)+\epsilon)\right) > \beta \\ M_p(t,k) &:= \left(\tilde{Y}_p(t,k)/(\tilde{Y}_h(t,k)+\epsilon)\right) \geq \beta \end{aligned}$$

where ϵ is a small constant to avoid division by zero, and the operators $\geq$ and $>$ yield a binary result from $\{0,1\}$. Applying these masks to the original spectrogram X yields the spectrograms for the harmonic and the percussive component

$$\begin{aligned} X_h(t,k) &:= X(t,k) \cdot M_h(t,k) \\ X_p(t,k) &:= X(t,k) \cdot M_p(t,k)\ . \end{aligned}$$

These spectrograms can then be brought back to the time-domain by applying an "inverse" short-time Fourier transform, see [7]. This yields the desired signals x_h and x_p. Choosing a separation factor $\beta > 1$ tightens the separation result of the procedure by preventing sounds which are neither clearly harmonic nor percussive to be included in the components. In Figure 2a, for example, you see the spectrogram of a sound mixture of a violin (clearly harmonic), castanets (clearly percussive), and applause (noise-like, and neither harmonic nor percussive). The sound of the violin manifests itself as clear horizontal structures, while one clap of the castanets is visible as a clear vertical structure in the middle of the spectrogram. The sound of the applause however does not form any kind of directed structure and is spread all over the spectrum. When decomposing this audio signal with a separation factor of β=1, which basically yields the procedure proposed in [5], the applause is more or less equally distributed among the harmonic and the percussive component, see Figure 2b. However, when choosing β=3, only the clearly horizontal and vertical structures are preserved in X_h and X_p, respectively, and the applause is no longer contained in the two components, see Figure 2c.

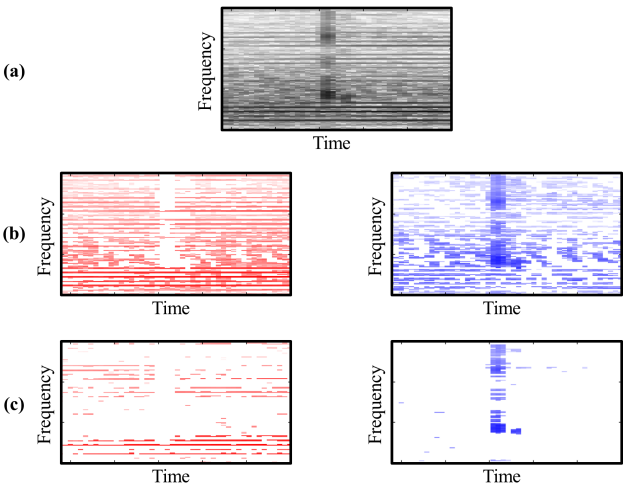

Figure 2. **(a):** Original spectrogram X. **(b):** Spectrograms X_h (left) and X_p (right) for $\beta = 1$. **(c):** Spectrograms X_h (left) and X_p (right) for $\beta = 3$.

3. HARMONIC-PERCUSSIVE-RESIDUAL SEPARATION

In Section 3.1 we show how harmonic-percussive separation can be extended with a third *residual* component. Afterwards, in Section 3.2, we show how the parameters of the proposed procedure influence the decomposition results. Finally, in Section 3.3, we present an iterative decomposition procedure which allows for a more flexible adjustment of the decomposition results.

3.1 Basic Procedure and Related Work

The concept presented in Section 2 allows us to extend the decomposition procedure with a third component x_r, called the *residual component*. It contains the portion of the input signal x that is neither part of the harmonic component x_h nor the percussive components x_p. To compute x_r, we define the binary mask

$$M_r(t,k) \quad := \quad 1 - \big(M_h(t,k) + M_p(t,k)\big),$$

apply it to X, and transform the resulting spectrogram X_r back to the time-domain (note that the masks M_h and M_p are disjoint). This decomposition into three components is inspired by the STN audio model. Here, an audio signal is analyzed to yield parameters for sinusoidal, transient, and noise components which can then be used to approximately resynthesize the original signal [8, 11]. While the main application of the STN model lies in the field of low bitrate audio coding, the estimated parameters can also be used to synthesize just the sinusoidal, the transient, or the noise component of the approximated signal. The harmonic, the percussive, and the residual component resulting from our proposed decomposition procedure are often perceptually similar to the STN components. However, our proposed procedure is conceptually different. STN modeling aims for a *parametrization* of the given audio signal. While the estimated parameters constitute a compact approximation of the input signal, this approximation and

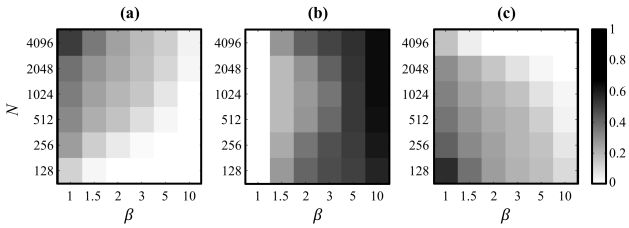

Figure 3. Energy distribution between the harmonic, residual, and percussive components for different frame sizes N and separation factors β. **(a):** Harmonic components. **(b):** Residual components. **(c):** Percussive components.

the original signal are not necessarily equal. Our proposed approach yields a *decomposition* of the signal. The three components always add up to the original signal again. The separation factor β hereby constitutes a flexible handle to adjust the sound characteristics of the components.

3.2 Influence of the Parameters

The main parameters of our decomposition procedure are the length of the median filters, the frame size N used to compute the STFT, and the separation factor β. Intuitively, the length of the filters specify the minimal sizes of horizontal and vertical structures which should be considered as harmonic and percussive sounds in the STFT of x, respectively. Our experiments have shown that the filter lengths actually do not influence the decomposition too much as long as no extreme values are chosen, see also [1]. The frame size N on the other hand pushes the overall energy of the input signal towards one of the components. For large frame sizes, the short percussive sounds lose influence in the spectral representation and more energy is assigned to the harmonic component. This results in a leakage of some percussive sounds to the harmonic component. Vice versa, for small frame sizes the low frequency resolution often leads to a blurring of horizontal structures, and harmonic sounds tend to leak into the percussive component. The separation factor β shows a different behavior to the previous parameters. The larger its value, the clearer becomes the harmonic and percussive nature of the components x_h and x_p. Meanwhile, also the portion of the signal that is assigned to the residual component x_r increases. To illustrate this behavior, let us consider a first synthetic example where we apply our proposed procedure to the mixture of a violin (clearly harmonic), castanets (clearly percussive), and applause (neither harmonic nor percussive), all sampled at 22050 Hertz and having the same energy. In Figure 3, we visualized the relative energy distribution of the three components for varying frame sizes N and separation factors β, while fixing the length of the median filters to be always equivalent to 200 milliseconds in horizontal direction and 500 Hertz in vertical direction, see also [1]. Since the energy of all three signals is normalized, potential leakage between the components is indicated by components that have either more or less than a third of the overall energy assigned. Considering Fitzgerald's procedure [5] as a baseline (β=1), we can investigate

its behavior by looking at the first columns of the matrices in Figure 3. While the residual component has zero energy in this setting, one can observe by listening that the applause is more or less equally distributed between the harmonic and the percussive component for medium frame sizes. This is also reflected in Figure 3a/c by the energy being split up roughly into equal portions. For very large N, most of the signal's energy moves towards the harmonic component (value close to one in Figure 3a for β=1, N=4096), while for very small N, the energy is shifted towards the percussive component (value close to one in Figure 3c for β=1, N=128). With increasing β, one can observe how the energy gathered in the harmonic and the percussive component flows towards the residual component (decreasing values in Figure 3a/c and increasing values in Figure 3b for increasing β). Listening to the decomposition results shows that the harmonic and the percussive component thereby become more and more extreme in their respective characteristics. For medium frame sizes, this allows us to find settings that lead to decompositions in which the harmonic component contains the violin, the percussive component contains the castanets, and the residual contains the applause. This is reflected by Figure 3, where for N=1024 and β=2 the three sound components all hold roughly one third of the overall energy. For very large or very small frame sizes it is not possible to get such a good decomposition. For example, considering β=1 and N=4096, we already observed that the harmonic component holds most of the signal's energy and also contains some of the percussive sounds. However, already for small $\beta > 1$ these percussive sounds are shifted towards the residual component (see the large amount of energy assigned to the residual in Figure 3b for β=1.5, N=4096). Furthermore, also the energy from the percussive component moves towards the residual. The large frame size therefore results in a very clear harmonic component while the residual holds both the percussive as well as all other non-harmonic sounds, leaving the percussive component virtually empty. For very small N the situation is exactly the other way around. This observation can be exploited to define a refined decomposition procedure which we discuss in the next section.

3.3 Iterative Procedure

In [13], Tachibana et al. described a method for the extraction of human singing voice from music recordings. In this algorithm, the singing voice is estimated by iteratively applying the harmonic-percussive decomposition procedure described in [9] first to the input signal and afterwards again to one of the resulting components. This yields a decomposition of the input signal into three components, one of which containing the estimate of the singing voice. The core idea of this algorithm is to perform the two harmonic-percussive separations on spectrograms with two different time-frequency resolutions. In particular, one of the spectrograms is based on a large frame size and the other on a small frame size. Using this idea, we now extend our proposed harmonic-percussive-residual separation procedure

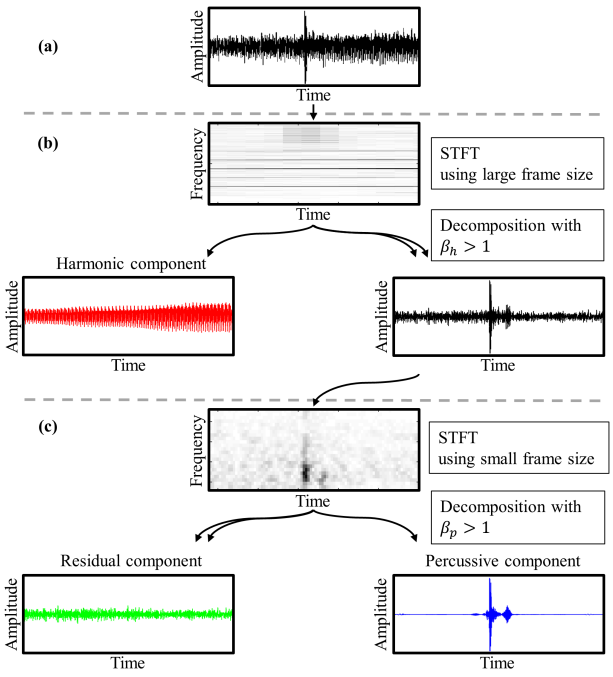

Figure 4. Overview of the refined procedure. **(a):** Input signal x. **(b):** First run of the decomposition procedure using a large frame size N_h and a separation factor β_h. **(c):** Second run of the decomposition procedure using a small frame size N_p and a separation factor β_p.

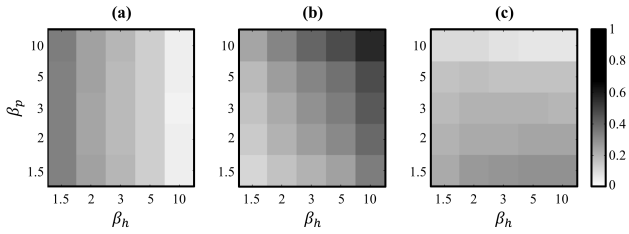

Figure 5. Energy distribution between the harmonic, residual, and percussive components for different separation factors β_h and β_p. **(a):** Harmonic components. **(b):** Residual components. **(c):** Percussive components.

presented in Section 3.1. So far, although it is possible to find a good combination of N and β such that both the harmonic as well as the percussive component represent the respective characteristics of the input signal well (see Section 3.2), the computation of the two components is still coupled. It is therefore not clear how to adjust the content of the harmonic and the percussive component independently. Having made the observation that large N lead to good harmonic but poor percussive/residual components for β>1, while small N lead to good percussive components but poor harmonic/residual components for β>1, we build on the idea from Tachibana et al. [13] and compute the decomposition in two iterations. Here, the goal is to decouple the computation of the harmonic component from the computation of the percussive component. First, the harmonic component is extracted by applying our basic procedure with a large frame size N_h and a separation factor β_h>1, yielding x_h^{first}, x_r^{first} and x_p^{first}. In a second run,

	SDR						SIR						SAR					
	BL	HP	HP-I	HPR	HPR-I	HPR-IO	BL	HP	HP-I	HPR	HPR-I	HPR-IO	BL	HP	HP-I	HPR	HPR-I	HPR-IO
Violin	-3.10	-5.85	0.08	8.23	7.65	8.85	-3.10	-5.09	1.08	17.69	14.58	21.65	274.25	8.33	9.44	8.82	8.78	9.11
Castanets	-2.93	3.58	2.86	8.29	9.14	9.28	-2.93	6.06	10.45	22.34	20.66	24.41	274.25	8.14	4.07	8.49	9.50	9.44
Applause	-3.04	–	-7.03	4.25	4.93	5.00	-3.04	–	14.69	8.41	12.80	9.04	274.25	–	-6.85	6.95	5.93	7.69

Table 1. Objective evaluation measures. All values are given in dB.

the procedure is applied again to the sum $x_r^{\text{first}} + x_p^{\text{first}}$, this time using a small frame size N_p and a second separation factor $\beta_p > 1$. This yields the components x_h^{second}, x_r^{second} and x_p^{second}. Finally, we define the output components of the procedure to be

$$x_h := x_h^{\text{first}},\ x_r := x_h^{\text{second}} + x_r^{\text{second}},\ x_p := x_p^{\text{second}}.$$

For an overview of the procedure see Figure 4. While fixing the values of N_h and N_p to a small and a large frame size, respectively (in our experiments we chose N_h=4096 and N_p=256), the separation factors β_h and β_p yield handles that give simple and independent control over the harmonic and percussive component. Figure 5, which is based on the same audio example as Figure 3, shows the energy distribution among the three components for different combinations of β_h and β_p, see also [1]. For the harmonic components (Figure 5a) we see that the portion of the signals energy contained in this component is independent of β_p and can be controlled purely by β_h. This is a natural consequence from the fact that in our proposed procedure the harmonic component is always computed directly from the input signal x and β_p does not influence its computation at all. However, we can also observe that the energy contained in the percussive component (Figure 5c) is fairly independent of β_h and can be controlled almost solely by β_p. Listening to the decomposition results confirms these observations. Our proposed iterative procedure therefore allows to adjust the harmonic and the percussive component almost independently what significantly simplifies the process of finding an appropriate parameter setting for a given input signal. Note that in principle it would also be possible to choose β_h=β_p=1, resulting in an iterative application of Fitzgerald's method [5]. However, as discussed in Section 3.2, Fitzgerald's method suffers from component leakage when using very large or small frame sizes. Therefore, most of the input signal's energy will be assigned to the harmonic component in the first iteration of the algorithm, while most of the remaining portion of the signal is assigned to the percussive component in the second iteration. This leads to a very weak, although not empty, residual component.

4. EVALUATION

In a first experiment, we applied objective evaluation measures to our running example. Assuming that the violin, the castanets, and the applause signal represent the characteristics that we would like to capture in the harmonic, the percussive, and the residual components, respectively, we treated the decomposition task of this mixture as a source separation problem. In an optimal decomposition the harmonic component would contain the original violin signal, the percussive component the castanets signal, and the residual component the applause. To evaluate the decomposition quality, we computed the *source to distortion ratios* (SDR), the *source to interference ratios* (SIR), and the *source to artifacts ratios* (SAR) [15] for the decomposition results of the following procedures.

As a baseline (BL), we simply considered the original mixture as an estimate for all three sources. Furthermore, we applied the standard harmonic-percussive separation procedure by Fitzgerald [5] (HP) with the frame size set to N=1024, the HP method applied iteratively (HP-I) with N_h=4096 and N_p=256, the proposed basic harmonic-percussive-residual separation procedure (HPR) as described in Section 3.1 with N=1024 and β=2, and the proposed iterative harmonic-percussive-residual separation procedure (HPR-I) as described in Section 3.3 with N_h=4096, N_p=256, and β_h=β_p=2. As a final method, we also considered HPR-I with separation factor β_h=3 and β_p=2.5, which were optimized manually for the task at hand (HPR-IO). The filter lengths in all procedures were always fixed to be equivalent to 200 milliseconds in time direction and 500 Hertz in frequency direction. Decomposition results for all procedures can be found at [1].

The results are listed in Table 1. All values are given in dB and higher values indicate better results. As expected, BL yields rather low SDR and SIR values for all components, while the SAR values are excellent since there are no artifacts present in the original mixture. The method HP yields low evaluation measures as well. However, these values are to be taken with care since HP decomposes the input mixture in just a harmonic and a percussive component. The applause is therefore not estimated explicitly and, as also discussed in Section 2, randomly distributed among the harmonic and percussive component. It is therefore clear that especially the SIR values are low in comparison to the other procedures since the applause heavily interferes with the remaining two sources in the computed components. When looking at HP-I, the benefit of having a third component becomes clear. Although here the residual component does not capture the applause very well (SDR of -7.03 dB) this already suf-

Item name	Description
CastanetsViolinApplause	Synthetic mixture of a violin, castanets and applause.
Heavy	Recording of heavily distorted guitars, a bass and drums.
Stepdad	Excerpt from *My Leather, My Fur, My Nails* by the band *Stepdad*.
Bongo	Regular beat played on bongos.
Glockenspiel	Monophonic melody played on a glockenspiel.
Winterreise	Excerpt from "Gute Nacht" by Franz Schubert which is part of the *Winterreise* song cycle. It is a duet of a male singer and piano.

Table 2. List of audio excerpts.

fices to yield SDR and SIR values clearly above the baseline for the estimates of the violin and the castanets. The separation quality further improves when considering the results of our proposed method `HPR`. Here the evaluation yields high values for all measures and components. The very high SIR values are particularly noticeable since they indicate that the three sources are separated very clearly with very little leakage between the components. This confirms our claim that our proposed concept of a separation factor allows for *tightening* decomposition results as described in Section 2. The results of `HPR-I` are very similar to the results for the basic procedure `HPR`. However, listening to the decomposition reveals that the harmonic and the percussive component still contain some slight residue sounds of the applause. Slightly increasing the separation factors to β_h=3 and β_p=2.5 (`HPR-IO`) eliminates these residues and further increases the evaluation measures. This straight-forward adjustment is possible since the two separation factors β_h and β_p constitute independent handles to adjust the content of the harmonic and percussive component, what demonstrates the flexibility of our proposed procedure.

The above described experiment constitutes a first case study for the objective evaluation of our proposed decomposition procedures, based on an artificially mixed example. To also evaluate these procedures on real-world audio data, we additionally performed an informal subjective listening tests with several test participants. To this end, we applied our procedures to the set of audio excerpts listed in Table 2. Among the excerpts are complex sound mixtures as well as purely percussive and harmonic signals, see also [1]. Raising the question whether the computed harmonic and percussive components meet the expectation of representing the clearly harmonic or percussive portions of the audio excerpts, respectively, the performed listening test confirmed our hypothesis. It furthermore turned out that β_h=β_p=2, N_h=4096 and N_p=256 seems to be a setting for our iterative procedure which robustly yields good decomposition results, rather independent of the input signal. Regarding the residual component, it was often described to sound like a *sound texture* by the test participants, which is a very interesting observation. Although there is no clear definition of what a sound texture exactly is, literature states "sound texture is like wallpaper: it can have local structure and randomness, but the characteristics of the fine structure must remain constant on the large scale" [12]. In our opinion this is not a bad description of what one can hear in residual components.

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

5. REFERENCES

[1] J. Driedger, M. Müller, and S. Disch. Accompanying website: Extending harmonic-percussive separation of audio signals. `http://www.audiolabs-erlangen.de/resources/2014-ISMIR-ExtHPSep/`.

[2] J. Driedger, M. Müller, and S. Ewert. Improving time-scale modification of music signals using harmonic-percussive separation. *Signal Processing Letters, IEEE*, 21(1):105–109, 2014.

[3] C. Duxbury, M. Davies, and M. Sandler. Separation of transient information in audio using multiresolution analysis techniques. In *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-01)*, Limerick, Ireland, 12 2001.

[4] C. Duxbury, M. Davies, and M. Sandler. Improved time-scaling of musical audio using phase locking at transients. In *Audio Engineering Society Convention 112*, 4 2002.

[5] D. Fitzgerald. Harmonic/percussive separation using medianfiltering. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pages 246–253, Graz, Austria, 2010.

[6] A. Gkiokas, V. Katsouros, G. Carayannis, and T. Stafylakis. Music tempo estimation and beat tracking by applying source separation and metrical relations. In *ICASSP*, pages 421–424, 2012.

[7] D. W. Griffin and J. S. Lim. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 32(2):236–243, 1984.

[8] S. N. Levine and J. O. Smith III. A sines+transients+noise audio representation for data compression and time/pitch scale modications. In *Proceedings of the 105th Audio Engineering Society Convention*, 1998.

[9] N. Ono, K. Miyamoto, H. Kameoka, and S. Sagayama. A real-time equalizer of harmonic and percussive components in music signals. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 139–144, Philadelphia, Pennsylvania, USA, 2008.

[10] N. Ono, K. Miyamoto, J. LeRoux, H. Kameoka, and S. Sagayama. Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram. In *European Signal Processing Conference*, pages 240–244, Lausanne, Switzerland, 2008.

[11] A. Petrovsky, E. Azarov, and A. Petrovsky. Hybrid signal decomposition based on instantaneous harmonic parameters and perceptually motivated wavelet packets for scalable audio coding. *Signal Processing*, 91(6):1489–1504, 2011.

[12] N. Saint-Arnaud and K. Popat. Computational auditory scene analysis. chapter Analysis and synthesis of sound textures, pages 293–308. L. Erlbaum Associates Inc., Hillsdale, NJ, USA, 1998.

[13] H. Tachibana, N. Ono, and S. Sagayama. Singing voice enhancement in monaural music signals based on two-stage harmonic/percussive sound separation on multiple resolution spectrograms. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(1):228–237, January 2013.

[14] Y. Ueda, Y. Uchiyama, T. Nishimoto, N. Ono, and S. Sagayama. HMM-based approach for automatic chord detection using refined acoustic features. In *ICASSP*, pages 5518–5521, 2010.

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