HARMONIC-PERCUSSIVE-RESIDUAL SOUND SEPARATION USING THE STRUCTURE TENSOR ON SPECTROGRAMS

Richard Füg[1], *Andreas Niedermeier*[1], *Jonathan Driedger*[2], *Sascha Disch*[1,2], *Meinard Müller*[2]

[1]Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany
[2]International Audio Laboratories Erlangen

ABSTRACT

Harmonic-percussive-residual (HPR) sound separation is a useful preprocessing tool for applications such as pitched instrument transcription or rhythm extraction. Recent methods rely on the observation that in a spectrogram representation, harmonic sounds lead to horizontal structures and percussive sounds lead to vertical structures. Furthermore, these methods associate structures that are neither horizontal nor vertical (i.e., non-harmonic, non-percussive sounds) with a residual category. However, this assumption does not hold for signals like frequency modulated tones that show fluctuating spectral structures, while nevertheless carrying tonal information. Therefore, a strict classification into horizontal and vertical is inappropriate for these signals and might lead to leakage of tonal information into the residual component. In this work, we propose a novel method that instead uses the structure tensor—a mathematical tool known from image processing—to calculate predominant orientation angles in the magnitude spectrogram. We show how this orientation information can be used to distinguish between harmonic, percussive, and residual signal components, even in the case of frequency modulated signals. Finally, we verify the effectiveness of our method by means of both objective evaluation measures as well as audio examples.

Index Terms— Harmonic-Percussive-Residual Separation, Structure Tensor, Spectrogram, STFT

1. INTRODUCTION

Being able to separate a sound into its harmonic[1] and percussive component is an effective preprocessing step for many applications. Using the separated percussive component of a music recording for example can lead to a quality improvement for beat tracking [1], rhythm analysis [2] and transcription of rhythm instruments. The separated harmonic component is suitable for the transcription of pitched instruments and chord detection [2, 3]. Furthermore, harmonic-percussive separation can be used for remixing purposes like changing the level ratio between both signal components [4], which may lead to an either "smoother" or "punchier" overall sound perception.

Recent methods for harmonic-percussive sound separation rely on the assumption that harmonic sounds have a horizontal structure in the magnitude spectrogram of the input signal (in time direction), while percussive sounds appear as vertical structures (in frequency direction). Ono et al. presented a method that first creates harmonically/percussively enhanced spectrograms by diffusion in time/frequency direction [5]. By comparing these enhanced representations afterwards, a decision if a sound is either harmonic or percussive could be derived. A similar method was published by Fitzgerald, where the enhanced spectrograms were calculated by using median filtering in perpendicular directions in-

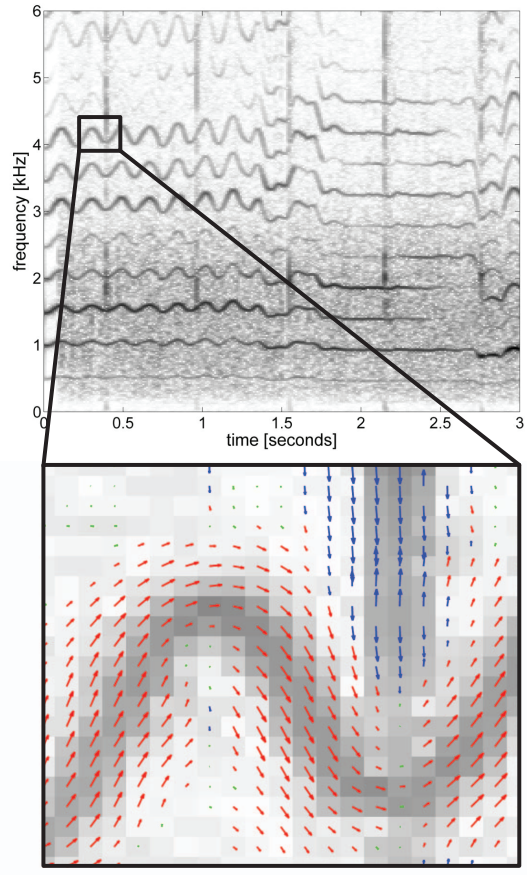

Fig. 1. Spectrogram of a mixture of a singing voice, castanets, and applause with zoomed in region additionally showing direction (orientation of arrows) and anisotropy measure (length of arrows) obtained by the structure tensor. The color of the arrows indicate whether the respective time-frequency bin is assigned to the harmonic (red), the percussive (blue), or the residual (green) component based on the orientation and anisotropy information.

[1]While "Harmonic-Percussive(-Residual) Separation" is a common term, it is misleading as it implies a harmonic structure with sinusoidals having a frequency of an integer multiple of the fundamental frequency. Even though the correct term should be "Tonal-Percussive-(Residual) Separation", we will refer to the common term and "harmonic" instead of "tonal" throughout this paper for easier understanding and coherence with existing work.

stead of diffusion [6], which led to similar results while reducing the computational complexity. Inspired by the sines+transients+noise (S+T+N) signal model [7, 8, 9]—a framework that aims to describe the respective signal components by means of a small set of parameters—Fitzgerald's method was then extended to harmonic-percussive-residual (HPR) separation in [10]. As audio signals often consist of sounds that are neither clearly harmonic nor percussive, this procedure captures these sounds in a third, *residual* component. While some of these residual signals clearly have an isotropic, neither horizontal nor vertical, structure (as for example noise), there exist sounds that do not have a clear horizontal structure but nevertheless carry tonal information and may be perceived as harmonic part of a sound. An example are frequency modulated tones like they can occur in recordings of violin playing or vocals, where they are said to have "vibrato". Due to the strategy of recognizing either horizontal or vertical structures, the aforementioned methods are not always able to capture such sounds in their harmonic component. It is worth noting that a harmonic-percussive separation procedure based on non-negative matrix factorization that is capable of capturing harmonic sounds with non-horizontal spectral structures in the harmonic component was recently proposed in [11]. However it did not include a third residual component.

In this work, we propose a novel approach for HPR separation. Instead of searching only for strictly horizontal and vertical structures, our method calculates predominant orientation angles as well as the local anisotropy in the spectrogram by using the *structure tensor* known from image processing. The provided information about the orientation of spectral structures can then be used to distinguish between harmonic, percussive, and residual signal components by setting appropriate thresholds, see figure 1.

2. THE STRUCTURE TENSOR ON SPECTROGRAMS

The structure tensor [12, 13] is a well known tool in image processing where it is applied to grey scale images for edge and corner detection [14] or to estimate the orientation of an object. While the structure tensor has already been used for preprocessing and feature extraction in audio processing [15, 16], its robust orientation information has not been exploited so far in this scope.

We now revisit the mathematical basics of the structure tensor and interpret it in the context of audio processing. In the following, matrices and vectors are written as bold letters for notational convenience. Furthermore, the $(\cdot)$ operator is used to index a specific element. In this case the matrix or vector is written as a non-bold letter to show its scalar use.

2.1. Calculation of the spectrogram

In this work, we apply the structure tensor to the spectrogram representation of a discrete input audio signal $\mathbf{x} \in \mathbb{R}^M$ with a sampling frequency of f_s. For the spectral analysis of $\mathbf{x}$, the short-time Fourier-transform (STFT)

$$X(b,k) := \sum_{n=0}^{N-1} w(n)x(n+Hb)\exp\left(-i2\pi nk/N\right) \qquad (1)$$

is used, where $X(b,k) \in \mathbb{C}$, b denotes the frame index, k the frequency index and $\mathbf{w} \in \mathbb{R}^N$ is a window function of length N. $H \in \mathbb{N}$, $H \leq N$ represents the analysis hop size of the window. Note that since the STFT spectrum has a certain symmetry around the Nyquist point at $N/2$, we restrict our processing to $0 \leq k \leq N/2$, as the symmetry can be reconstructed during the inverse STFT.

Furthermore we calculate the real valued logarithmic spectrogram

$$S(b,k) = 20\log_{10}\left|X(b,k)\right| . \qquad (2)$$

2.2. Calculation of the structure tensor

For the calculation of the structure tensor the partial derivatives of $\mathbf{S}$ are needed. The partial derivative with respect to time index b is given by

$$\mathbf{S}_b = \mathbf{S} * \mathbf{d} \qquad (3)$$

while the partial derivative with respect to frequency index k is defined as

$$\mathbf{S}_k = \mathbf{S} * \mathbf{d}^T \qquad (4)$$

where $\mathbf{d}$ is a discrete differentiation operator (for example, for central differences one could choose $\mathbf{d} = [-1, 0, 1]/2$) and $*$ denotes the 2-dimensional convolution. Furthermore we define

$$\mathbf{T}_{11} = (\mathbf{S}_b \odot \mathbf{S}_b) * \mathbf{G} \qquad (5)$$

$$\mathbf{T}_{21} = \mathbf{T}_{12} = (\mathbf{S}_k \odot \mathbf{S}_b) * \mathbf{G} \qquad (6)$$

$$\mathbf{T}_{22} = (\mathbf{S}_k \odot \mathbf{S}_k) * \mathbf{G} \qquad (7)$$

where $\odot$ is the point wise matrix multiplication, also known as the Hadamard product and $\mathbf{G}$ is a 2-D Gaussian smoothing filter having the standard deviation σ_b in time index direction and σ_k in frequency index direction. The structure tensor $\mathbf{T}(b,k)$ is then given by a 2×2 symmetric and positive semidefinite matrix

$$\mathbf{T}(b,k) = \begin{bmatrix} T_{11}(b,k) & T_{12}(b,k) \\ T_{21}(b,k) & T_{22}(b,k) \end{bmatrix} . \qquad (8)$$

The structure tensor contains information about the dominant orientation of the spectrogram at position (b,k). Note that in the special case where $\mathbf{G}$ is a scalar, $\mathbf{T}(b,k)$ does not contain more information than the gradient at this position in the spectrogram. However in contrast to the gradient, the structure tensor can be smoothed by $\mathbf{G}$ without cancellation effects, which makes it more robust against noise.

2.3. Calculation of angles and anisotropy measure

The information about the orientation for each bin in the spectrogram is obtained by calculating the eigenvalues $\lambda(b,k)$, $\mu(b,k)$ with $\lambda(b,k) \leq \mu(b,k)$ and the corresponding eigenvectors $\mathbf{v}(b,k) = [v_1(b,k), v_2(b,k)]^T$ and $\mathbf{w}(b,k) = [w_1(b,k), w_2(b,k)]^T$ of the structure tensor $\mathbf{T}(b,k)$. Note that $\mathbf{v}(b,k)$, the eigenvector corresponding to the smaller eigenvalue $\lambda(b,k)$, is pointing into the direction of lowest change in the spectrogram at index (b,k), while $\mathbf{w}(b,k)$ is pointing in to the direction of highest change. Thus, the angle of the orientation at a specific bin can be obtained by

$$\alpha(b,k) = \operatorname{atan}\left(\frac{v_2(b,k)}{v_1(b,k)}\right) \in [-\pi/2; \pi/2] . \qquad (9)$$

In addition, a measure of anisotropy

$$C(b,k) = \begin{cases} \left(\frac{\mu(b,k)-\lambda(b,k)}{\mu(b,k)+\lambda(b,k)}\right)^2 & , \ \mu(b,k)+\lambda(b,k) \geq e \\ 0 & , \ \text{else} \end{cases} \qquad (10)$$

with $e \in \mathbb{R}^{>0}$ can be determined for each bin. Note that $C(b,k) \in [0;1]$. Values of $C(b,k)$ close to 1 indicate a high anisotropy of the spectrogram at index (b,k), while a constant neighborhood leads to values close to 0. The threshold e, that defines a limit on what should be considered anisotropic, can be chosen to further increase the robustness against noise.

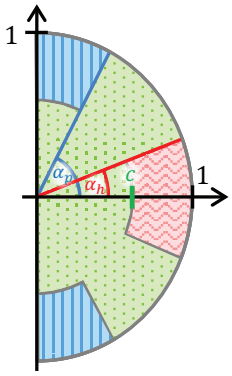

Fig. 2. Range of orientation/anisotropy values computed by the structure tensor. Values in areas marked in red wavy lines lead to an assignment to the harmonic component, in blue vertical lines to the percussive component, and in dotted green to the residual component.

2.4. Interpreting the angles and smoothing parameters

The physical meaning of angle $\alpha(b,k)$ can be understood by considering a continuous signal with a change of instantaneous frequency Δf during a time interval Δt. Thus the instantaneous frequency change rate R is denoted by

$$R = \frac{\Delta f}{\Delta t}. \tag{11}$$

Considering sample rate, length and hop-size of the applied STFT analysis, a relation between the angles in the spectrogram and the instantaneous frequency change rate $R(b,k)$ for each bin can be derived by

$$R(b,k) = \frac{{f_s}^2}{HN} \cdot \tan(\alpha(b,k)). \tag{12}$$

Also the standard deviations of the smoothing filter $\mathbf{G}$ in the discrete domain σ_b and σ_k can be converted to the continuous physical parameters σ_t and σ_f by

$$\sigma_t = \frac{H}{f_s}\sigma_b, \qquad \sigma_f = \frac{f_s}{N}\sigma_k. \tag{13}$$

3. HARMONIC-PERCUSSIVE-RESIDUAL SEPARATION USING THE STRUCTURE TENSOR

Now, the information obtained via the structure tensor can be applied to the problem of HPR separation. Our goal is to classify each bin in the spectrogram as being part of either the harmonic, the percussive or the residual component of the input signal. As discussed in section 1, bins assigned to the harmonic components should belong to rather horizontal structures, while bins belonging to rather vertical structures should be assigned to the percussive component. Bins that do not belong to any kind of oriented structure should be assigned to the residual component. Intuitively, for a bin (b,k) to be assigned to the harmonic component, it should satisfy two constraints. First, the absolute value of the angle $\alpha(b,k)$ should be smaller than some threshold $\alpha_h \in [0;\pi/2]$. This means, that the bin should be part of some spectral structure that does not have a slope bigger or smaller than α_h. This way also frequency modulated sounds can be considered to be part of the harmonic component, depending on the parameter α_h. Secondly, the measure of anisotropy $C(b,k)$ should support that the bin (b,k) is part of some directed, anisotropic structure, and therefore exceeds a second threshold c. Note that for a given bin (b,k), the angle $\alpha(b,k)$ together with the measure of anisotropy $C(b,k)$ define a point in $\mathbb{R}^2$ given in polar coordinates. Figure 2 depicts the subset of all points that lead to an assignment to the harmonic component (red regions). Similarly, we introduce another angle threshold α_p to define when a bin should be assigned to the percussive component (blue regions in figure 2). Finally, all bins that are assigned to neither the harmonic nor the percussive component are assigned to the residual component (green regions in figure 2).

This assignment process can be expressed by defining a mask for the harmonic component $\mathbf{M}_h$, a mask for the percussive component $\mathbf{M}_p$ and a mask for the residual component $\mathbf{M}_r$. Note, that instead of using the thresholds α_h and α_p we define thresholds on the maximum absolute frequency change rate $r_h, r_p \in \mathbb{R}^{>0}$ with $r_p \geq r_h$ to give the choice of parameters a better physical interpretation. The masks are then given by:

$$M_h(b,k) = \begin{cases} 1 & ,|R(b,k)| \leq r_h \ \wedge \ C(b,k) > c \\ 0 & ,\text{else} \end{cases} \tag{14}$$

$$M_p(b,k) = \begin{cases} 1 & ,|R(b,k)| > r_p \ \wedge \ C(b,k) > c \\ 0 & ,\text{else} \end{cases} \tag{15}$$

$$M_r(b,k) = 1 - M_h(b,k) - M_p(b,k). \tag{16}$$

Finally, the STFT of the harmonic component $\mathbf{X}_h$, the percussive component $\mathbf{X}_p$ and the residual component $\mathbf{X}_r$ are obtained by

$$\mathbf{X}_h = \mathbf{M}_h \odot \mathbf{X} \tag{17}$$

$$\mathbf{X}_p = \mathbf{M}_p \odot \mathbf{X} \tag{18}$$

$$\mathbf{X}_r = \mathbf{M}_r \odot \mathbf{X}. \tag{19}$$

The corresponding time signals can then be calculated via the inverse STFT.

4. EVALUATION

To show the effectiveness of our proposed procedure in capturing frequency modulated sounds in the harmonic component, we compared our HPR method based on the structure tensor (HPR-ST) with the non-iterative method based on median filtering presented in [10] (HPR-M). Additionally, we also computed the metrics for the separation results with ideal binary masks (IBM) that served as a reference for the maximal achievable separation quality.

4.1. System-under-test parameters

For both HPR-ST as well as HPR-M, the STFT parameters were chosen to be f_s=22050Hz, N=1024 and H=256, using a sine window for $\mathbf{w}$. The separation parameters for HPR-M were chosen as in the experiments performed in [10]. For our method, the structure tensor was calculated using the Scharr-Operator [17] as discrete differentiation operator $\mathbf{d}$. The smoothing was performed using a 9×9 isotropic Gaussian filter with the standard deviations σ_b=σ_k=1.4 which leads to $\sigma_t \approx$16ms and $\sigma_f \approx$30Hz. Finally, the thresholds for the separation were set to e=20, c=0.2 and r_h=r_p=10000Hz/s.

Note that by our choice of r_h and r_p, even very steep structures in the spectrogram are assigned to the harmonic component. Our choice is motivated by observations about real world vibrato sounds as for example shown in figure 1. Here, you can see that at some instances the vibrato in the singing voice has a very high instantaneous frequency change rate. Furthermore, note that by choosing r_h=r_p,

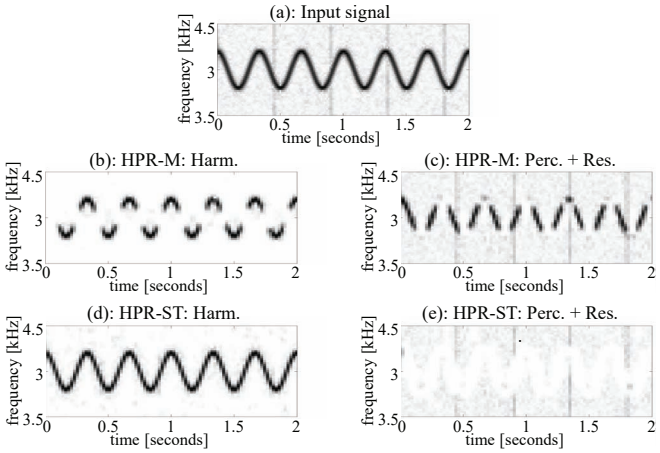

Fig. 3. Comparison between the HPR-M and HPR-ST method for an excerpt of the synthetic input signal (item 1). For enhanced visibility the spectrograms were calculated with different STFT parameters than used for the separation algorithms.

an assignment of a bin in the spectrogram to the residual component is purely dependent on its anisotropy measure.

4.2. Objective results

To compare the behavior of HPR-ST and HPR-M when applied to signals containing frequency modulated sounds, we generated two test items. Test item 1 consists of the superposition of purely synthetic sounds. The harmonic source was chosen to be a vibrato tone with a fundamental frequency of 1000Hz, a vibrato frequency of 3Hz, vibrato extent of 50Hz and 4 overtones. For the percussive source several impulses are used, while white noise represents the neither harmonic nor percussive residual source. Item 2 was generated by superimposing real world signals of singing voice with vibrato (harmonic), castanets (percussive), and applause (neither harmonic nor percussive). Interpreting the HPR separation of these items as a source separation problem, we computed standard source separation evaluation metrics (*Source to distortion ratio* SDR, *source to interference ratio* SIR, and *source to artifacts ratios* SAR, as introduced in [18]) for the separation results of both procedures. The results are shown in table 1.

For item 1 HPR-ST yields a SDR of 21.25dB for the vibrato tone, and is therefore closer to the optimal separation result of IBM (29.43dB) than to the the separation result of HPR-M (11.51dB). This indicates that HPR-ST improves on capturing this frequency modulated sound in the harmonic component in comparison to HPR-M. This is also shown in figure 3, where the spectrograms of the harmonic components and the sum of the percussive and residual component computed for both procedures are plotted. It can be seen that for HPR-M the steep slopes of the vibrato tone leaked into the residual component (figure 3b+c), while HPR-ST (figure 3d+e) yields a good separation. This also explains the very low SIR values of HPR-M for the residual component compared to HPR-ST (-11.99dB vs. 14.12dB). Note that the high SIR value of HPR-M for the harmonic component only reflects that there are little interfering sounds from the other components, not that the sound of the vibrato is well captured as a whole. In general most of the observations for item 1 are less pronounced, but also valid for the mixture of real world sounds in item 2. For this item, the SIR value of HPR-M for the vocals even exceeds the SIR value of HPR-ST (20.83dB vs. 15.61dB). Again, the low SIR value for the applause supports that portions of the vibrato in the vocals leaked into the residual component for HPR-M (1.11dB) while the residual component of HPR-ST contains less interfering sounds (6.34dB). This indicates that our procedure was capable of capturing the frequency modulated structures of the vocals much better than HPR-M.

		SDR			SIR			SAR		
		IBM	HPR-M	HPR-ST	IBM	HPR-M	HPR-ST	IBM	HPR-M	HPR-ST
Item 1	Vibrato	29.43	11.51	21.25	34.26	27.94	30.01	31.16	11.61	21.88
	Impulses	8.56	-10.33	-1.47	20.31	-7.96	12.03	8.90	2.02	-1.00
	Noise	8.49	-13.53	2.58	24.70	-11.99	14.12	8.61	3.97	3.06
Item 2	Vocals	14.82	6.48	9.18	22.75	20.83	15.61	15.60	6.68	10.42
	Castanets	8.48	3.79	2.37	21.59	16.09	17.96	8.73	4.16	2.56
	Applause	7.39	-2.03	-0.37	20.31	1.11	6.34	7.66	3.33	1.58

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

Additionally to this objective evaluation, we also set up an accompanying website for this paper at [19] where one can find the audio signals that were used in our experiments along with all separation results.

5. CONCLUSION AND OUTLOOK

In this paper, we have proposed a novel approach for the problem of HPR separation based on the structure tensor. We have shown how frequency modulated sounds that hold tonal information can be captured in the harmonic component of our procedure by exploiting the information about the orientation of spectral structures provided by the structure tensor. Finally, we evaluated the effectiveness of our procedure HPR-ST by comparing it to the state-of-art median filtering based method HPR-M presented in [10] by means of both objective evaluation measures as well as audio examples. For signals that contain frequency modulated tones, the novel HPR-ST method was shown to provide much better separation results compared to HPR-M. Note that even though research in image processing and computer vision has already brought several additions and enhancements for the structure tensor that could be applicable to the problem discussed in this work, we have restricted ourselves to the basic version to demonstrate the general functionality and usefulness of this approach for HPR separation. Consequently, we see a high potential in a further transfer of existing image processing research to improve the method presented in this paper in future work. Finally, we think that the structure tensor might not only be beneficial for the task of HPR separation, but may also find applications in other audio processing tasks such as *singing voice detection* [20].

Acknowledgments:
The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits IIS. Jonathan Driedger is supported by the German Research Foundation (DFG MU 2686/6-1). Finally, we would like to thank Christian Uhle for his valuable feedback on the paper.

6. REFERENCES

[1] Aggelos Gkiokas, Vassilios Katsouros, George Carayannis, and Themos Stafylakis, "Music tempo estimation and beat tracking by applying source separation and metrical relations," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012, pp. 421–424.

[2] Nobutaka Ono, Kenichi Miyamoto, Hirokazu Kameoka, Jonathan Le Roux, Yuuki Uchiyama, Emiru Tsunoo, Takuya Nishimoto, and Shigeki Sagayama, "Harmonic and percussive sound separation and its application to MIR-related tasks," in *Advances in Music Information Retrieval*, vol. 274 of *Studies in Computational Intelligence*, pp. 213–236. Springer Berlin Heidelberg, 2010.

[3] Yushi Ueda, Yuuki Uchiyama, Takuya Nishimoto, Nobutaka Ono, and Shigeki Sagayama, "HMM-based approach for automatic chord detection using refined acoustic features," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, Texas, USA, 2010, pp. 5518–5521.

[4] Nobutaka Ono, Kenichi Miyamoto, Hirokazu Kameoka, and Shigeki Sagayama, "A real-time equalizer of harmonic and percussive components in music signals," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Philadelphia, Pennsylvania, USA, 2008, pp. 139–144.

[5] Nobutaka Ono, Kenichi Miyamoto, Jonathan LeRoux, Hirokazu Kameoka, and Shigeki Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in *European Signal Processing Conference*, Lausanne, Switzerland, 2008, pp. 240–244.

[6] Derry Fitzgerald, "Harmonic/percussive separation using median filtering," in *Proceedings of the International Conference on Digital Audio Effects (DAFX)*, Graz, Austria, 2010, pp. 246–253.

[7] Scott N. Levine and Julius O. Smith III, "A sines+transients+noise audio representation for data compression and time/pitch scale modications," in *Proceedings of the AES Convention*, 1998.

[8] Tony S. Verma and Teresa H.Y. Meng, "An analysis/synthesis tool for transient signals that allows a flexible sines+transients+noise model for audio," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Seattle, Washington, USA, May 1998, pp. 3573–3576.

[9] Laurent Daudet, "Sparse and structured decompositions of signals with the molecular matching pursuit," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1808–1816, September 2006.

[10] Jonathan Driedger, Meinard Müller, and Sascha Disch, "Extending harmonic-percussive separation of audio signals," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Taipei, Taiwan, 2014, pp. 611–616.

[11] Jeongsoo Park and Kyogu Lee, "Harmonic-percussive source separation using harmonicity and sparsity constraints," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Málaga, Spain, 2015, pp. 148–154.

[12] Josef Bigun and Gösta H. Granlund, "Optimal orientation detection of linear symmetry," in *Proceedings of the IEEE First International Conference on Computer Vision*, London, UK, 1987, pp. 433–438.

[13] Hans Knutsson, "Representing local structure using tensors," in *6th Scandinavian Conference on Image Analysis*, Oulu, Finland, 1989, pp. 244–251.

[14] Chris Harris and Mike Stephens, "A combined corner and edge detector," in *Proceedings of the 4th Alvey Vision Conference*, Manchester, UK, 1988, pp. 147–151.

[15] Rolf Bardeli, "Similarity search in animal sound databases," *IEEE Transactions on Multimedia*, vol. 11, no. 1, pp. 68–76, January 2009.

[16] Matthias Zeppelzauer, Angela S. Stöger, and Christian Breiteneder, "Acoustic detection of elephant presence in noisy environments," in *Proceedings of the 2nd ACM International Workshop on Multimedia Analysis for Ecological Data*, Barcelona, Spain, 2013, pp. 3–8.

[17] Hanno Scharr, *Optimale Operatoren in der digitalen Bildverarbeitung*, Dissertation, IWR, Fakultät für Physik und Astronomie, Universität Heidelberg, Heidelberg, Germany, 2000.

[18] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[19] Website, "Accompanying website: Harmonic-percussive-residual sound separation using the structure tensor on spectrograms," `www.audiolabs-erlangen.de/resources/MIR/2016-ICASSP-HPRST/`.

[20] Bernhard Lehner, Gerhard Widmer, and Reinhard Sonnleitner, "On the reduction of false positives in singing voice detection," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 7480–7484.