Phase-based Harmonic/Percussive Separation

Conference Paper · October 2014

CITATIONS
14

READS
244

3 authors:

Estefanía Cano
Fraunhofer Institute for Digital Media Technology IDMT
32 PUBLICATIONS **149** CITATIONS
SEE PROFILE

Mark D. Plumbley
University of Surrey
366 PUBLICATIONS **5,671** CITATIONS
SEE PROFILE

Christian Dittmar
Friedrich-Alexander-University of Erlangen-Nürnberg
76 PUBLICATIONS **542** CITATIONS
SEE PROFILE

Phase-based Harmonic/Percussive Separation

Estefanía Cano[1], *Mark Plumbley*[2], *Christian Dittmar*[1]

[1]Fraunhofer Institute for Digital Media Technology IDMT.
[2]Centre for Digital Music, Queen Mary University of London
cano@idmt.fraunhofer.de, mark.plumbley@qmul.ac.uk, dmr@idmt.fraunhofer.de

Abstract

In this paper, a method for separation of harmonic and percussive elements in music recordings is presented. The proposed method is based on a simple spectral peak detection step followed by a phase expectation analysis that discriminates between harmonic and percussive components. The proposed method was tested on a database of 10 audio tracks and has shown superior results to the reference state-of-the-art approach.

Index Terms: harmonic/percussive separation, phase expectation, perceptual quality

1. Introduction

Separating harmonic instruments from percussive ones is a relevant task in contexts such as audio re-mixing, beat tracking, tempo analysis, and music education among others. Harmonic/percussive separation can also be used as an intermediate step to allow more complex types of analysis such as automatic music transcription and pitch tracking, where the presence of percussive elements can make the extraction of relevant harmonic information, more difficult.

Percussive sounds appear in the spectrogram as vertical (broadband) elements spanning a short interval of time. This can be observed in Figure 1 where the spectra of two percussive music signals are displayed. This clear characteristic of percussive sounds has been used as a mean to separate percussive elements from harmonic ones - which in contrast, appear in the spectrogram as horizontal events [1, 2]. An additional challenge that methods for harmonic/percussive separation need to face is the fact that some elements of musical sounds, which in a strict sense are not percussive sounds but exhibit similar spectral characteristics to percussive sounds, are very difficult to discriminate. This is the case for example of unvoiced fricative and plosive vocal sounds and musical instrument attacks.

This paper addresses the problem of harmonic/percussive separation using phase as main element in the processing chain. The remainder of this paper presents the proposed method and is organized as follows: in Section 2, the state-of-the-art in harmonic/percussive separation is presented. Section 3 presents a theoretical background related to phase processing. In Section 4, the proposed method is described and results are presented in Section 5.

2. State-of-the-art

As mentioned in Section 1, proposed methods for harmonic/percussive separation have taken advantage of the vertical structure of percussive events to perform separation. This is the case for the method proposed by Ono et al. [1] where the anisotropy—or dependency to direction—of the power spectrogram is exploited. The method works under the assumption that the power spectrogram $W(k,n)$ can be expressed as a sum of percussive $P(k,n)$ and harmonic elements $H(k,n)$, that is:

$$W(k,n) = P(k,n) + H(k,n) \tag{1}$$

where k represents the frequency bin index, and n the time index.

The authors evaluate the anisotropic smoothness of the power spectrogram gradients, $P(k,n-1) - P(k,n)$ and $H(k,n-1) - H(k,n)$, by minimizing an auxiliary function given by:

$$J(\mathbf{H},\mathbf{P}) = \frac{1}{2\sigma_H^2}\sum_{k,n}(H(k,n-1) - H(k,n))^2 + \frac{1}{2\sigma_P^2}\sum_{k,n}(P(k,n-1) - P(k,n))^2 \tag{2}$$

Here, σ_H and σ_P are parameters to control the weights of the horizontal and vertical smoothness.

Similarly, Fitzgerald proposed in [2] a method for harmonic/percussive separation that exploits the vertical structure of percussive events and applies median filtering in the magnitude spectrogram to perform separation. To extract harmonic components from the audio signal, a median filter of length L_H is applied to each frequency slice (bin) in the magnitude spectrogram $M(k,n)$. That is, for each frequency bin k, a median filter in the time direction is applied to obtain a harmonic-enhanced magnitude spectrogram. What this approach effectively accomplishes is to smooth the transient-like percussive events from the temporal envelopes of each frequency bin k in the spectrogram. Similarly, percussive events are detected by applying a median filter of length L_P in the frequency direction in every time frame n of the magnitude spectrogram. The use of soft Wiener masks is proposed to obtain the final percussive and harmonic signals. The harmonic mask, for example, is defined as:

$$M_H(k,n) = \frac{H^p(k,n)}{H^p(k,n) + P^p(k,n)} \tag{3}$$

where p is the power to which the components are raised to obtain range compressed spectrograms.

Not directly applied to harmonic/percussive separation but extremely similar in nature is the *Tonalness Spectrum* presented in [3]. The tonalness spectrum $\mathcal{T}(k,n)$ shows the likelihood of a spectral bin to be a tonal or non-tonal component. To do so, a set of spectral features t_i with $i = 0,\ldots,V$ is defined and finally combined to produce the tonalness spectrum:

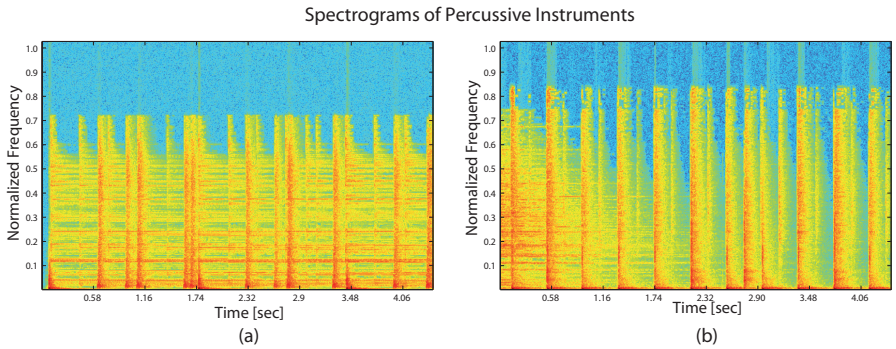

Figure 1: Spectrograms of percussive instruments. Clear verticals events in the spectrogram mark the percussive hits.

$$\mathcal{T}(k,n) = (\prod_{i=1}^{V} t_i(k,n))^{1/\eta} \quad (4)$$

The feature set is composed by an amplitude continuity feature—closely related to the spectrogram gradient in the time direction—, a frequency continuity feature—related to spectrogram gradient in the frequency direction—, a frequency deviation feature that accounts for spectral leakage, a frequency coherence feature related to the instantaneous frequency, a peakiness feature, and a time window center of gravity feature.

Other works applied *Non-negative Matrix Factorization* [4, 5] or *Independent Subspace Analysis* [6] for harmonic/percussive separation. They all first decompose the music spectrogram into components and then classify them as either percussive or harmonic.

3. Background

For the remainder of this paper, the following notation applies: let $S(k,n)$ be the complex-valued spectrogram of an audio signal $f(t)$ sampled with sampling frequency f_s. Additionally, let N be the frame length of the transform, and H its hop size. Frequency bins are indicated with the index k , and time frames with the index n. $|S(k,n)|$ represents the magnitude spectrum and the (unwrapped) phase spectrum is given by $\phi(k,n)$.

As opposed to the methods described in Section 2, the proposed method for harmonic/percussive separation is based on two concepts related to spectral phase: *Phase Expectation* and *Instantaneous Frequency Distribution IFD*. These concepts are further explained in the following sections.

3.1. Instantaneous Frequency Distribution IFD

An important aspect in the prediction of phase spectrum is the concept of *unwrapped phase*. To better handle phase information, it is a common procedure to unwrap the phase spectrum to obtain a continuous representation where the discontinuities are removed. The most common process of phase unwrapping corrects the phase values by adding multiples of $\pm 2\pi$ when absolute jumps between adjacent values are greater than or equal to a pre-defined tolerance. The tolerance is normally chosen to be π.

The Instantaneous Frequency Distribution (IFD) is derived from the first-order time derivative of the unwrapped phase spectrum. It is defined as follows:

$$\Phi(k,n) = \frac{1}{2\pi}\frac{d\phi(k,n)}{dn} \quad (5)$$

where $\phi(k,n)$ is the unwrapped phase spectrum. In practice, the differentiation in (5) is approximated by taking the difference between two consecutive values of the phase spectrum. The division by 2π is used to normalize the instantaneous frequency (IF). The normalized IF can be used to obtain the IF in Hertz simply by multiplying it by the sampling frequency fs [7]. In Figure 2a, the IFD of the the first 10 partials of a trumpet tone is displayed.

3.2. Phase Expectation

The main idea behind phase expectation is that the frame-wise change in phase of a harmonic source can be predicted given the pitch of the source, and the hop size H in samples of the time-frequency transform. For a given harmonic source evident in frequency bin k of $S(k,n)$, the phase change in radians from time frame n to time frame $n+1$ is given by:

$$\Delta\phi_k(n) = \frac{2\pi f_k \cdot H}{fs} \quad (6)$$

with f_k the center frequency of bin k in Hz. The concept of phase expectation has been applied for estimation of overlapped harmonics in [8], for monaural source separation in [9], and time-stretching via phase vocoder [10].

Each frequency bin k in $S(k,n)$ covers a band of frequencies defined by the center frequency f_k and the width of each band fs/N. Thus, the band of frequencies covered by bin k is given by $[f_{k_{Low}}, f_{k_{High}}]$ with $f_{k_{Low}} = f_k - fs/(2N)$ and $f_{k_{High}} = f_k + fs/(2N)$. Equation (6) can be used to calculate the range of expected phase changes $[\Delta_{k_{Low}}, \Delta_{k_{High}}]$ for the frequency band covered by each bin k. These values of expected phase changes can be used to predict whether the energy falling in a given time-frequency bin belongs to a harmonic source or not. If the change in phase between two consecutive time frames falls in the expected radian ranges, the source can be assumed harmonic. If on the contrary, the phase change falls outside the expected ranges, the source exhibits transient or noise-like characteristics.

The concept of phase expectation is depicted in Figure 2b. Due to spectral leakage in the time-frequency transform, the presence of a harmonic source in bin k is likely to affect its adjacent frequency bins, $k-1$ and $k+1$. Figure 2b shows the phase expectation plot of two consecutive A4 saxophone tones. The vertical axis displays three frequency bins for each of the ten first partials p of the tones: the bin where the peak is observed k_p, its lower adjacent bin $k_p - 1$, and its higher adjacent bin $k_p + 1$. The natural order of the frequency bins was preserved in the plot, being $k_p - 1$ the lower bin shown, k_p the

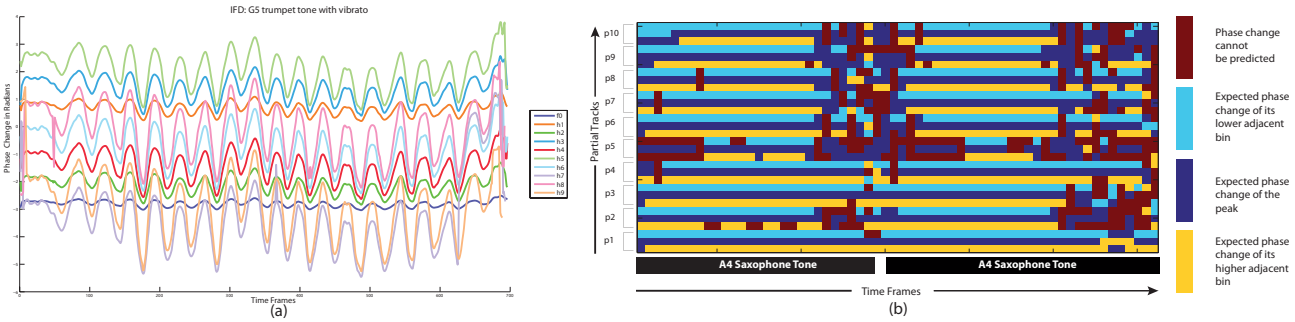

Figure 2: (a) Instantaneous Frequency Distribution (IFD) $\Phi(k,n)$ of the first 10 partials of a G5 trumpet tone with vibrato. $f0$ stands for fundamental frequency and h_i are the first nine harmonics with $i = 1,\ldots,9$. (b) Phase expectation plot for the first 10 partials of two saxophone tones.

middle bin shown, and $k_p + 1$ the higher bin shown for each partial. The legend of Figure 2b describes the color convention used. The time-frequency bins whose phase change falls in the expected radian ranges as calculated with (6), are shown in dark blue. Those frequency bins whose expected phase change falls in the radian ranges of their higher adjacent bins are shown in yellow. Those frequency bins whose expected phase change falls in the radian ranges of their lower adjacent bins are shown in light blue. Those frequency bins whose phase cannot be explained or predicted based on radian ranges calculated with (6), are shown in dark red.

The plots clearly show a structured behavior for the phase expectation of each partial and its adjacent bins. It can be observed that the frequency bin k_p, where each partial p is observed, exhibits phase change values mostly in the predicted ranges, showing very clear dark blue horizontal trajectories. Phase expectation of k_p+1 and k_p-1 clearly follow the behavior of k_p, $k_p - 1$ being mostly pulled to higher phase changes (closer to k_p) than the expected ones. This can be observed by the clear yellow horizontal trajectories in the plots. Phase expectation of $k_p + 1$ is also clearly affected by the main peak k_p, showing phase change values lower to the ones predicted and mainly being pulled down to be closer to k_p. This can be observed by the horizontal light blue trajectories in the plots. A very important observation to be made is the behavior of phase expectation when the higher partials of the tones in Figure 2b decay. It can be seen that as the partials decay at the end of each of the two tones, their phase changes cannot longer be accurately predicted, mainly showing dark red color in the plot. This clear difference in the behavior of phase expectation between harmonic sources and non-harmonic ones will be exploited in the separation context.

4. Proposed Method

The concept of phase expectation is exploited to perform the separation task. The fact that for a certain frequency bin k, phase values of tonal components will fall within a radian range determined by the frequency band covered by k, and the hop size H of the time-frequency transform, is exploited. Phase values outside the calculated range are assumed non-harmonic and classified as percussive components. The method works as follows:

1. Calculate $|S(k,n)|$ and $\phi(k,n)$ by means of the *Short Time Fourier Transform (STFT)*.
2. For each frequency bin k, use Eq. (6) to calculate $[\Delta_{k_{Low}}, \Delta_{k_{High}}]$
3. For each time frame n, find the p_{max} peaks with the largest amplitude in the power spectrogram $|S(k,n)|^2$. The peak detection algorithm proposed in [11] was used. The detected peaks in frame n form a set defined as $Q(n)$.
4. Create a spectral mask $M(k,n)$ such that:

$$M(k,n) = \begin{cases} 1 & \text{for all peaks in } Q(n) \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

5. Use the Hadamard product $\bigodot$ to calculate the masked phase spectrogram: $\hat{\phi}(k,n) = \phi(k,n) \bigodot M(k,n)$.
6. Use Eq. 5 to calculate $\Phi(k,n)$ for each bin in $\hat{\phi}(k,n)$.
7. Calculate binary spectral masks $H(k,n)$ and $P(k,n)$ for the harmonic and percussive components, respectively. For every time frame n and bin k in $\hat{\phi}(k,n)$:

$$H(k,n) = \begin{cases} 1 & \text{if } \Delta_{k_{Low}} < \Phi(k,n) < \Delta_{k_{High}} \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

$$P(k,n) = 1 - H(k,n) \quad (9)$$

It has been observed that for those time-frequency bins where both harmonic and percussive components overlap, phase characteristics of percussive instruments tend to prevail. These time-frequency bins are classified as percussive and no attempt is made to estimate the underlying harmonic component.

8. Spectral leakage is considered by including in the harmonic mask, the adjacent frequency bins $k_p + 1$ and $k_p - 1$ whose phase changes follow the ones of the main peak k_p. In Figure 2b, that means that all the yellow and light blue time-frequency bins are also included in the harmonic mask.
9. Percussive and harmonic signals $p(n)$ and $h(n)$ are obtained by means of the *Inverse Short Time Fourier Transform (ISTFT)* of the masked spectrograms: $p(n) = ISFT\{S(k,n) \bigodot P(k,n)\}$, $h(n) = ISFT\{S(k,n) \bigodot H(k,n)\}$

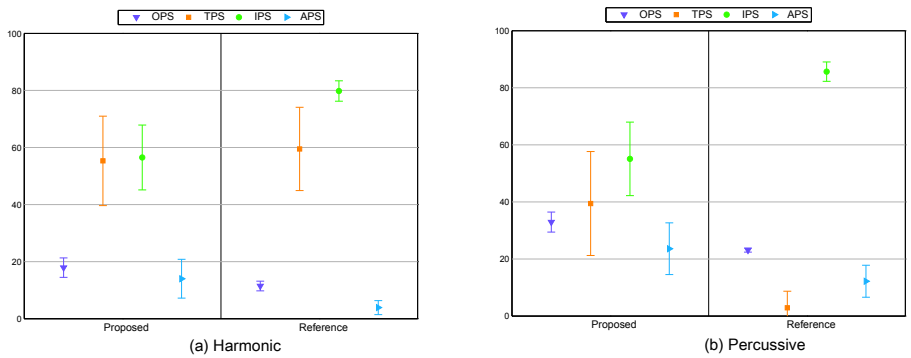

Figure 3: Results obtained with the proposed harmonic/percussive separation algorithm and the algorithm proposed by Fitzgerald [2] (Reference). Overall Perceptual Score (OPS), Target-Related Perceptual Score (TPS), Interference-Related Perceptual Score (IPS), Artifact-Related Perceptual Score (APS). Mean values with 95% confidence intervals are presented. (a) Harmonic. (b) Percussive.

5. Evaluation

The proposed algorithm for harmonic/percussive separation and the algorithm proposed by Fitzgerald in [2] were evaluated and compared under a common dataset and test conditions.

5.1. Dataset

In order to evaluate the proposed method, a dataset composed of 10 audio tracks was collected and mixed into harmonic and percussive components from the available multi-track recordings. The dataset has been made publicly available in the following [12].

5.2. Quality Metrics

To evaluate quality of the resulting separated tracks, the PEASS Toolkit [13] was used. The toolkit presents a family of four perceptually-motivated quality measures: Overall Perceptual Score (OPS) which is a measure of general separation quality, Target-Related Perceptual Score (TPS), a measure of how much the target source is preserved; Interference- Related Perceptual Score (IPS), a measure of interference from other sources, and Artifact-Related Perceptual Score (APS) which measures the amount of artifacts created during separation.

5.3. Implementation Details

The following algorithm parameters were used: a window length $N = 2048$ with a hop size $H = 128$. All tracks in the dataset had a sampling frequency $fs = 44100$ Hz. For the peak detection algorithm the frequency adaptive magnitude threshold was calculated using a frequency delta $\Delta_f = 50$ Hz.

For the reference algorithm the processing parameters recommended by the author for best performance were used: window length $N = 4096$, hop size $H = 1024$, filter length $L = 17$, and spectral compression parameter $p = 2$.

5.4. Results

Perceptual quality measures for both algorithms were obtained using the PEASS Toolkit. Mean values and 95% confidence intervals are presented in Figure 3. The proposed algorithm outperforms the reference method in terms of the OPS both for the percussive and harmonic components. Particularly noticeable is the performance improvement obtained with the proposed method for the percussive components, the proposed method obtaining a mean OPS score of 32.93, and the reference method a mean OPS score of 23.11 over the entire dataset. For the harmonic components (Figure 3a), the proposed method obtains slightly higher APS scores than the reference algorithm, but slightly lower TPS scores. For the percussive components (Figure 3b), the proposed method outperforms the reference algorithm in three of the perceptual scores, that is, OPS, TPS, and APS. It is to be noted that particularly high IPS scores are obtained by the reference algorithm for the both harmonic and percussive components, outperforming the proposed method in both cases. Informal listening tests showed that for vocal tracks, the proposed method assigns more of the fricative and plosive sounds to the percussive signal than the reference algorithm. For the instrumental tracks, more information from the attacks of the instruments is assigned by the proposed algorithm to the percussive signal than the reference algorithm. These two observations explain the clear difference in IPS scores between the two methods. Due to the transient-like characteristics of fricatives, plosives, and attacks, their corresponding phase exhibits non-harmonic characteristics and consequently, are assigned to the percussive components in the separation. Additionally, the reference algorithm shows in general smaller confidence intervals than the proposed method which suggests it can be more reliable under different signal characteristics. The resulting signals can be accessed in [12].

6. Conclusions

In this paper, a method for harmonic/percussive separation based on phase processing has been presented. The system was evaluated using perceptual objective quality scores and showed superior performance in particular for the OPS, in comparison to the reference algorithm. These results support the hypothesis that the phase spectrum carries valuable information relevant to sound separation tasks, and can help improve the quality of audio source separation.

7. Acknowledgments

Parts of this work were conducted under the supervision of Dr. Mark Plumbley as part of a research stay at the Centre for Digital Music C4DM, at Queen Mary University of London. Mark Plumbley was supported by a grant EP/H043101/1 and a Leadership Fellowship EP/G007144/1 from the Engineering and Physical Sciences Research Council (EPSRC).

8. References

[1] N. Ono, K. Miyamoto, and J. L. Roux, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in *16th European Signal Processing Conference (EUSIPCO 2008)*, Lausanne, Switzerland, 2008, p. 4 pages.

[2] D. Fitzgerald, "Harmonic/percussive separation using median filtering," in *13th International Conference on Digital Audio Effects (DAFx-10)*, no. 1, Graz, Austria, 2010, pp. 10–13.

[3] S. Kraft, A. Lerch, and U. Zölzer, "The tonalness spectrum: feature-based estimation of tonal components," in *16th International Conference on Digital Audio Effects (DAFx-13)*, Maynooth, Ireland, 2013, p. 8 pages.

[4] M. Helén and T. Virtanen, "Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine," in *Proc. of the 13th European Signal Processing Conference (EUSIPCO)*, 2005.

[5] B. Schuller, A. Lehmann, F. Weninger, F. Eyben, and G. Rigoll, "Blind enhancement of the rhythmic and harmonic sections by NMF: Does it help?" in *Proceedings International Conference on Acoustics including the 35th German Annual Conference on Acoustics, NAG/DAGA 2009*, Acoustical Society of the Netherlands, DEGA. Rotterdam, The Netherlands: DEGA, March 2009, pp. 361–364, invited contribution.

[6] C. Uhle, C. Dittmar, and T. Sporer, "Extraction of drum tracks from polyphonic music using independent subspace analysis," in *Proc. of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation*, 2003.

[7] L. D. Alsteris and K. K. Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," *Digital Signal Processing*, vol. 17, no. 3, pp. 578–616, May 2007.

[8] J. Woodruff, Y. Li, and D. Wang, "Resolving overlapping harmonics for monaural musical sound separation using pitch and common amplitude modulation," in *9th International Society for Music Information Retrieval Conference (ISMIR 2008)*, Philadelphia, USA, 2008, pp. 538–543.

[9] Y. Li, J. Woodruff, S. Member, and D. Wang, "Monaural musical sound separation based on pitch and common amplitude modulation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1361–1371, 2009.

[10] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *Speech and Audio Processing, IEEE Transactions on*, vol. 7, no. 3, pp. 323–332, May 1999.

[11] E. Cano and C. Cheng, "Melody line detection and source separation in classical saxophone recordings," in *12th International Conference on Digital Audio Effects (DAFx-09)*, Como, Italy, 2009, p. 6 pages.

[12] E. Cano and Fraunhofer IDMT, "Phase-based Harmonic Percussive Separation," 2013. [Online]. Available: http://www.idmt.fraunhofer.de/en/Departments_and_Groups/smt/phase_based_harmonic_percussive_separation.html

[13] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, Sep. 2011.