A Segmental Spectral Flatness Measure for Harmonic-Percussive Discrimination

Julian M. BECKER[1], *Christian ROHLFING*[1]

[1]Institute of Communications Engineering, RWTH Aachen University, 52056 Aachen, Germany

{becker, rohlfing}@ient.rwth-aachen.de

Abstract. *In a variety of applications of audio signal processing, for example blind source separation (BSS), harmonic signals need to be treated differently from percussive signals. Thus, recognizing if a signal is harmonic or percussive is a very helpful and frequently used preprocessing step. Different measures have been proposed to capture either the harmonicity or the percussivity of a signal, using knowledge of the spectral shape of the signal or investigating its periodic behavior.*

In this paper, we propose a new method for distinguishing harmonic and percussive signals and evaluate it in a BSS environment. We compare it to three other state-of-the-art measures. Our experiments show, that our method obtains better results than the traditionally used methods.

Keywords

Harmonicity, percussivity, spectral flatness, harmonic ratio.

1. Introduction

Knowing if an audio signal originates from a harmonic or a percussive source can be very helpful for further processing in a lot of audio signal processing applications such as blind source separation ([4]). Different methods have been developed to extract information about either harmonicity or percussivity of a signal. Measures of spectral flatness ([1, 2]) are very popular for this task. These measures use the fact, that percussive signals usually have a relatively flat spectrum. Other measures, such as the harmonic ratio (HR), which is part of the audio harmonicity descriptor in the MPEG-7 audio content description ([3]), make use of the periodicity of the signal.

In the context of blind source separation (BSS) of audio signals, these methods can be helpful to estimate the harmonicity or percussivity of separated sound events (or components). A lot of BSS algorithms, such as the one described in [4], are using Nonnegative Matrix Factorization (NMF) to separate the magnitude spectrogram of a signal into different sound events. These components are separated in an iterative process, which can be improved by treating harmonic components differently to percussive ones. However, classifying the components during the iterative process is challenging, because the components are not separated optimally at early stages of the process. Thus a method is needed which is robust to artifacts in the signal.

In this paper we present a new measure for distinguishing between harmonic and percussive signals. We evaluate the method in a BSS environment and show that it leads to better results than traditionally used methods.

The paper is structured as follows: In Section 2 we describe three traditional measures, before describing the proposed method in Section 3. In Section 4 we show our experimental results followed by our conclusions in Section 5.

2. Traditional Measures

In this section we desribe three traditionally used measures for the hormonicity, respectively the percussivity, of a signal. Two of these methods use information about the spectral shape of the signal, whereas the third measure uses information about the periodic behavior of the signal.

2.1. Spectral Flatness

The spectral flatness measure was introduced as a measure of the noisiness respectively sinusoidaliy of a signal in [2]. The idea is to measure the "whiteness" of a discrete time signal $x(t)$ with the magnitude spectrum $X(k)$ of length K, by using the fact that white noise has a perfectly flat spectrum. The measure is defined as

$$\mathcal{F}_1(X) = \frac{\left[\Pi_{k=0}^{K-1} X(k)\right]^{\frac{1}{K}}}{\frac{1}{K}\sum_{k=0}^{K-1} X(k)}, \tag{1}$$

where the numerator equals the geometric mean and the denominator the arithmetic mean of $X(k)$. We obtain $\mathcal{F}_1(X) = 0$ for a purely sinusoidal signal $x(t)$ and $\mathcal{F}_1(X) = 1$ for white noise. This flatness measure will be called spectral flatness (SF) in the following.

The fact that transient signals usually have a noisy spectrum, whereas harmonic signals have a sparse spectrum, made SF very popular to distinguish between percussive and har-

monic signals. It has also been included in the MPEG-7 framework ([3]).
One downside of this measure is its sensitivity to small values in the spectrum. The product in the numerator leads to a very small value of $\mathcal{F}_1(X)$ already, if only one coefficient in the spectrum is close to zero. Thus low values of $\mathcal{F}_1$ do not necessarily indicate a harmonic signal, but a signal with at least one very small value in the spectrum.

2.2. Entropy Spectral Flatness

In [1] the problem of $\mathcal{F}_1$ being very sensitive to small values was adressed and a new flatness measure was introduced. For a signal $x(t)$ with the magnitude spectrum $X(k)$ of length K the measure is defined as

$$\log_2\left(\mathcal{F}_2(X)+1\right) = -\frac{1}{\log_2(K)}\sum_k \hat{X}(k)\log_2\left(\hat{X}(k)\right), \tag{2}$$

with

$$\hat{X}(k) = \frac{X(k)}{\sum_{k=0}^{K-1} X(k)} \tag{3}$$

being the normalized spectrum X. We still obtain $\mathcal{F}_2(X) = 0$ for a purely sinusoidal signal $x(t)$ and $\mathcal{F}_2(X) = 1$ for white noise, but for a perfectly flat spectrum with only one element close to zero we obtain a value close to one.
Due to its relation to Shannon's entropy this measure will in the following be called entropy spectral flatness (ESF).

2.3. Harmonic Ratio

Spectral flatness measures such as SF and ESF can be used to distinguish between percussive and harmonic signals under the assumption, that percussive signals have a relatively flat spectrum, whereas the spectrum of a harmonic signal is sparse. Unfortunately this assumption does not always hold: Some percussive instruments, for example a bass drum, are only active in a small frequency range. Thus the spectrum is not flat over the whole frequency range and using a flatness measure would lead to these instruments being classified as harmonic. Instead of using spectral flatness measures it is also possible to estimate the harmonicity of a signal $x(t)$. A commonly used measure for harmonicity is the harmonic ratio (HR) which is part of the audio harmonicity descriptor of the MPEG-7 framework ([3]).
First the signal $x(t)$ is partitioned into frames of size N_w with a hop of N_{hop} samples between successive frames. For each frame the normalized autocorrelation function is calculated as

$$\mathrm{r}_l(m) = \frac{\sum_{t=0}^{N_w-1} x_l(t)x_l(t-m)}{\sqrt{\sum_{t=0}^{N_w-1} x_l(t)^2 \sum_{t=0}^{N_w-1} x_l(t-m)^2}} \tag{4}$$

where $x_l(t)$ is defined as $x(lN_{hop}+t)$, with $0 \le l \le L-1$ being the frame index. $1 \le m \le M$ is the lag index of the autocorrelation. The maximum lag M corresponds to the minimum fundamental frequency that can be estimated, which is usually set to 25 Hz. HR is then defined as

$$\mathcal{F}_3(x) = \max_{M_0 \le m \le M}\{r_l(m)\}, \tag{5}$$

where M_0 is usually defined as the lag corresponding to the first zero crossing of the autocorrelation. The mean HR of all frames of the signal can be used to classify it as harmonic or percussive signal.
HR has the downside of needing more complex calculations than the spectral flatness measures, due to the calculation of the autocorrelation functions. It also only uses one peak of the autocorrelation function. This means it is not assured, that the signal is really periodic, which would also lead to peaks at multiples of the estimated lag. It also does not take into account that a harmonic signal usually contains harmonics at multiples of the fundamental frequency.

3. A Segmental Entropy Spectral Flatness Measure

In this section we present a new measure to distinguish between harmonic and percussive signals, based on spectral flatness as well as characteristics of harmonic signals. Harmonic signals idealy have a sparse spectrum with peaks only at multiples of the fundamental frequency. Percussive signals on the other hand usually have a more flat spectrum, but not necessarily over the whole frequency range. However, for smaller segments of the frequency range the assumption of flatness should hold. Therefor we use a segmental spectral flatness measure, where each segment should contain exactly one peak for harmonic signals. The measure is calculated as follows:

Peak Picking In a first step, the first peak of the magnitude spectrum $X(k)$, denoted as k_p, is searched. k_p is assumed to correspond to the fundamental frequency f_0 (or a multiple of it) for a harmonic signal. The signal is only searched for peaks starting at index k_0, where k_0 corresponds to a frequency of 25 Hz. A peak is defined as a value that is larger than both neighboring values. We set the additional conditions, that a peak has to have an amplitude of at least $a_{\min}$ and that two peaks have to have a distance of at least k_0. The value for $a_{\min}$ can be chosen relative to the maximum value of $X(k)$.

Segmentation In the next step, the signal is partitioned into different segments. Each segment should contain exactly one peak, if the signal is harmonic. The segments $\sigma_i(k')$ are estimated symmetric around multiples i of k_p with a segment length of $l = 2 \cdot \lfloor \frac{k_p}{2} \rfloor + 1$.

$$\sigma_i(k') = X(i \cdot k_p - \lfloor \frac{l}{2} \rfloor) \tag{6}$$

The segmentation is illustrated in Figure 1.

Flatness Estimation For each segment $\sigma_i(k')$, the spectral flatness $F_i = \mathcal{F}_2(\sigma_i(k'))$ is calculated using the ESF measure described in Section 2.2. Also the energy $E_i = \sum_{k'} \sigma_i(k')^2$ of each segment is estimated.

Averaging Finally, the average segmental flatness of the signal is calculated by summing up the spectral flatness of each segment after weighting with its relative energy

$$\mathcal{F}_4(X) = \frac{\sum_i F_i \cdot E_i}{\sum_i E_i}. \quad (7)$$

The energy-weighting is necessery to prevent low-energy segments, which usually only contain noise, from distorting the result.

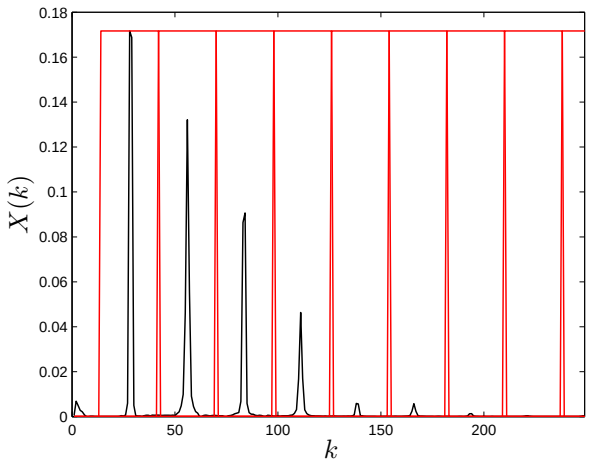

Fig. 1. Example for the segmentation of the spectrum of a harmonic signal. Each segment contains exactly one peak.

With this estimation, the measure lies in the range $0 \leq \mathcal{F}_4 \leq 1$ with $\mathcal{F}_4 = 0$ for purely harmonic signals and $\mathcal{F}_4 = 1$ for white noise.
We will call this measures segmental entropy spectral flatness (SESF) in the following.

4. Experimental Results

In this section, we compare the three traditional methods for distinguishing between harmonic and percussive signals described in Section 2 with the method proposed in Section 3 by using a blind source separation (BSS) environment. We describe the method that we used for evaluation and give a short overview over the test database and the setup of the algorithms. Finally we present the results of the experiments.

4.1. Evaluation

We use a BSS framework similar to the one described in [4] for evaluation of the different methods. We separate two-source mixtures $x(t)$, consisting of one harmonic source $s_h(t)$ and one percussive source $s_p(t)$ into different components by Nonnegative Matrix Factorization (NMF). Each component is then assigned either to the harmonic or the percussive source depending on the result of the different measures. The evaluation algorithm works as follows:

STFT First, the short-time Fourier transform (STFT) of the mixture $x(t)$ is taken. The resulting complex-valued spectrogram $\underline{\mathbf{X}}$ is of size $K \times N$ with frequency-bins $1 \leq k \leq K$ and time-bins $1 \leq n \leq N$. The magnitude spectrogram is denoted $\mathbf{X} = |\underline{\mathbf{X}}|$.

NMF In the next step, $\mathbf{X}$ is factorized by the NMF. This results in a separation of I sound events. These sound events are desribed by spectral basis vectors $b_i(k)$ and temporal activation vectors $g_i(n)$ with $1 < i < I$.

ISTFT The I sound events are transformed back into time domain by inverse short-time fourier transform, resulting in I signals $s_i(t)$.

Clustering The signals are assigned to either the harmonic or the percussive source by using the four described measures. SF, ESF and the proposed measure use the spectral basis vectors $b_i(k)$, whereas HR uses the time signals $s_i(t)$. The results of the four measures are each clustered into two clusters using the k-means algorithm [5]. The clusters correspond to the harmonic, respectively the percussive source. This step results in the estimated harmonic and percussive sources $\tilde{s}_h(t)$ and $\tilde{s}_p(t)$.

For a more detailed description of the BSS algorithm see [4]. The quality of the measures is evaluated in two different ways:

1. The source-to-distortion ratio (SDR) between $s_h(t)$ and $\tilde{s}_h(t)$ respectively $s_p(t)$ and $\tilde{s}_p(t)$ is calculated as

$$\mathrm{SDR}_{h/p} = 10 \log \frac{\sum_t s_{h/p}^2(t)}{\sum_t (s_{h/p}(t) - \tilde{s}_{h/p}(t))^2}. \quad (8)$$

   The average over both values is used as measure of the quality of the distinction between harmonic and percussive sources.

2. The assignments of each component to the harmonic or the percussive source are checked and the number of wrong assignments is counted.

4.2. Database & Setup

We used a database of eight different harmonic and eight different percussive sources, which were combined in all possible harmonic-percussive two-source mixtures, which made a total of 64 signals. The signals have a length of 4 to 16 seconds and are all sampled with a sampling rate of 44.1 kHz. For the BSS algorithm the windowsize of the STFT is set to 4096, the hopsize to 2048. The number of components I is set to 15. For the HR measure the framesize N_w is set to 4096 and the hopsize N_{hop} to 2048. For the peak picking step in the proposed method, $a_{\min}$ is set to 12% of the maximum value of $X(k)$.

4.3. Results

The results are displayed in Table 1. They show the following:

- By far, SF leads to the worst results in SDR as well as in the number of errors. This was to be expected, as it is very sensitive to zeros or values close to zero in the spectrum. Real signals, however, usually have at least some spectral values, that are close to zero. EF, which also measures spectral flatness, but is less sensitiv to zeros, obtains much better results. This shows, that the flatness of the spectrum can in fact be used to distinguish between harmonic and percussive sources.
- HR, which measures the harmonicity of a signal instead of the spectral flatness, obtains better results than both flatness measures.
- Our proposed method obtains the best results in SDR. Regarding the number of correctly assigned components, it performs better than both spectral flatness measures, but HR performs slightly better than the proposed method.

	SF	ESF	HR	SESF
SDR	4.54	6.98	7.93	8.27
errors	26.35%	22.08%	19.06%	19.48%

Tab. 1. Results for all four measures for SDR and relative number of incorrectly assigned components.

Besides the analysis of the classification performance we analyzed the computing time of each algorithm qualitatively. The spectral flatness measures SD and ESF are the fastest algorithms, SESF needs about five times the computational time of these two measures. HR is by far the slowest algorithm, needing around 60 times the computational time of SESF. In summary we can conclude, that the proposed method is well suited for distinguishing between harmonic and percussive signals. It obtains better results in SDR quality than all other measures and in computational time compared to HR. The ESF seems to be a good alternative for applications that depend on a low computational time.

5. Conclusions

In this paper we presented a new method to distinguish between harmonic and percussive signals. We compared our method to three well known and commonly used methods and showed, that it obtained the best results regarding the quality of the classification. Regarding computational time we showed that our method is faster than HR, but slower than the spectral flatness measures.
Further improvements in quality might be possible by combining different methods, which could be matter of future work.

References

[1] MADHU, N. Note on measures for spectral flatness. In *Electronics Letters*, 2009, vol.45, no. 23, pp. 1195 - 1196.

[2] GRAY, A.H., and MARKEL, J.D. A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis. *IEEE Trans. Acoust. Speech Signal Process.*, 1974, 22, pp. 207 - 217

[3] KIM, H.-G., MOREAU, N., Sikora, T. *MPEG-7 Audio and Beyond*. John Wiley & Sons, Ltd., 2005.

[4] SPIERTZ, M., GNANN, V. Source-filter based clustering for monaural blind source separation. In *Proceedings of International Conference on Digital Audio Effects DAFx '09*. Como (Italy), 2009.

[5] HARTIGAN, *Clustering Algorithms*. 99th ed. John Wiley & Sons, Inc., 1975.

About Authors...

Julian M. BECKER

was born in Bad Kreuznach, Germany in 1985. He received the Dipl.-Ing. degree in Electrical Engineering and Information Technology from RWTH Aachen University, Aachen, Germany, in 2010. Currently, he is pursuing the Ph.D. degree at the Institute of Communications Engineering, RWTH Aachen University, focusing on blind source separation of audio signals.

Christian ROHLFING

was born in Krefeld, Germany in 1985. He received the Dipl.-Ing. degree in Computer Engineering from RWTH Aachen University, Aachen, Germany, in 2012. Currently, he is pursuing the Ph.D. degree at the Institute of Communications Engineering, RWTH Aachen University, focusing on blind source separation of audio signals.