DEPLOYING NONLINEAR IMAGE FILTERS TO SPECTROGRAM FOR HARMONIC/PERCUSSIVE SEPARATION

Aggelos Gkiokas
Institute for Language and Speech Processing / "R.C. Athena"
National Technical University of Athens
Athens, Greece
agkiokas@ilsp.gr

Vassilis Papavassiliou
Institute for Language and Speech Processing / "R.C. Athena"
Athens, Greece
vpapa@ilsp.gr

Vassilis Katsouros
Institute for Language and Speech Processing / "R.C. Athena"
Athens, Greece
vsk@ilsp.gr

George Carayannis
National Technical University of Athens
Athens, Greece
carayan@softlab.ece.ntua.gr

ABSTRACT

In this paper we present a simple yet novel technique for harmonic/percussive separation of monaural audio music signals. Under the assumption that percussive/harmonic components exhibit vertical/horizontal lines in the spectrogram, image morphological filters are applied to the spectrogram of the input signal. The structure elements of the morphological filters are chosen to accentuate regions of the spectrogram corresponding to harmonic and percussive components. The proposed method was evaluated on the SISEC 2008/2010 development data and outperformed the baseline method adopted.

1. INTRODUCTION

The design of systems that automatically separate a music signal to its percussive and harmonic components has gained a great attention in the last few years. Such systems meet many applications such as remixing, but also serve as a pre-processing step in many applications in the MIR domain. Tonal characteristics such as chords, melody, and musical key of a music piece can be easier extracted from the signal under the absence of percussive instruments. Moreover drums transcription can be more efficient if some could pre-extract the drums sound from the audio mixture. Harmonic/Percussive separation can also enhance the performance of rhythmic analysis methods as shown in [1].

Harmonic/Precussive separation was firstly found in the literature mostly in terms of drums separation and transcription. FitzGerald et al. utilized tensor factorization models to separate drums from polyphonic music [2]. Gillet et al. extracted drum sounds via harmonic and noise decomposition. The noise component was interpreted as drum sound [3]. Yoshii et al. proposed a drum sound recognition system that matches drum templates after the application of harmonic structure suppression to spectrogram [4]. Helen and Virtanen applied non-negative matrix factorization and SVM for the separation of drums from polyphonic music [5]. In [6] Rigaud et al. proposed a method for drum extraction by spectro-temporal modelling of drum events.

A major milestone in Harmonic/Percussive separation was Ono et al. work [7][8]. They adopted the assumption that percussive sounds exhibit vertical lines in the spectrogram while horizontal lines are interpreted as harmonic content. Horizontal/vertical gradients for harmonic/percussive components of the compressed spectrogram were mutually minimized to derive the respective content. Duong et al. [9] recently extended this work by considering spectral and spatial continuity of harmonic and percussive components from stereo recordings to derive maximum a posteriori estimates.

The same assumption was adopted by FitzGerald [10], who proposed a simple yet efficient system that applies a median filter to rows/columns of the compressed spectrogram in order to derive masks for the harmonic/percussive components. In [11] Thoshkahna and Ramakrishnan extend FitzGerald's work by proposing a post-processing technique on the percussive component. Harmonic leakages are suppressed by signal reconstruction using envelope properties on sub-bands. Their method was evaluated on Indian and Western music and outperformed FitzGerald's method.

In this paper we present a method for separating a monaural music signal into harmonic/percussive components, considering the image of the spectrogram and using the assumption that vertical/horizontal lines correspond to percussive/harmonic components respectively. In particular, we apply nonlinear morphological filtering on the spectrogram. Nonlinear morphological filters are widely used in image processing tasks, such as noise reduction, object detection and image segmentation. However, applications of such filters in audio processing are rarely found in the literature.

2. IMAGE MORPHOLOGICAL FILTERS

2.1. Binary Operators on Binary Images

Let $\mathbf{X},\mathbf{B}:\mathbb{Z}^2 \to \{0,1\}$ be binary images defined on $\mathbb{Z}^2$. For the

Figure 1. *From left to right: a) Original image. b) Image erosion by a 45° angle line. c) Image erosion by a 135° angle line.*

binary image A we denote $a \in \mathbf{A}$ if $\mathbf{A}(a) = 1$. The following basic operations are defined:

a) Translation of $\mathbf{X}$ by $z \in \mathbb{Z}^2$:

$$\mathbf{X}_z(x) \triangleq \mathbf{X}(x - z) \tag{1}$$

b) Symmetric $\mathbf{X}^s$ and complement $\mathbf{X}^c$ of $\mathbf{X}$:

$$\mathbf{X}^s(x) \triangleq \mathbf{X}(-x), \quad \mathbf{X}^c(x) \triangleq 1 - \mathbf{X}(x) \tag{2}$$

c) Dilation of $\mathbf{X}$ by $\mathbf{B}$:

$$(\mathbf{X} \oplus \mathbf{B})(x) \triangleq \max_y \{\mathbf{X}(x - y), y \in \mathbf{B}\} \tag{3}$$

d) The dual operator of dilation is called erosion:

$$(\mathbf{X} \ominus \mathbf{B})(x) \triangleq \min_y \{\mathbf{X}(x + y), y \in \mathbf{B}\} = (\mathbf{X}^c \oplus \mathbf{B}^s)^c \tag{4}$$

e) Opening of $\mathbf{X}$ by $\mathbf{B}$:

$$\mathbf{X} \circ \mathbf{B} = (\mathbf{X} \ominus \mathbf{B}) \oplus \mathbf{B} \tag{5}$$

f) Closing of $\mathbf{X}$ by $\mathbf{B}$:

$$\mathbf{X} \bullet \mathbf{B} = (\mathbf{X} \oplus \mathbf{B}) \ominus \mathbf{B} \tag{6}$$

The geometrical interpretation of dilation and erosion is that erosion shrinks image $\mathbf{X}$ according to structure element $\mathbf{B}$ while dilation expands $\mathbf{X}$. For further reading we refer to [12].

2.2. Expansion of Binary Operators on Real Sets

In the case of real valued images, every image $\mathbf{X}: \mathbb{Z}^2 \to \mathbb{R}$ is associated with a real function $f_X: \mathbb{Z}^2 \to \mathbb{R}$ as

$$f_X(y) = \mathbf{X}, y \in \mathbb{Z}^2 \tag{7}$$

In order to expand the notions of dilation and erosion to real valued images the *level sets* of $\mathbf{X}$ are defined as:

$$\mathbf{X}_\upsilon(f_X) = \{x \in \mathbb{Z}^2 : f_X(x) \geq \upsilon\}, \quad \upsilon \in \mathbb{R} \tag{8}$$

Level sets are binary images and have the property that can uniquely represent the signal:

$$f_X(x) = \sup\{\upsilon \in \mathbb{R} : x \in \mathbf{X}_\upsilon(f_X)\} \tag{9}$$

Then the erosion and dilation of $\mathbf{X}$ by $\mathbf{B}$ are defined as:

$$(f_X \oplus \mathbf{B})(x) \triangleq \sup_{y \in B} f_X(x - y) \tag{10}$$

$$(f_X \ominus \mathbf{B})(x) \triangleq \inf_{y \in B} f_X(x + y) \tag{11}$$

Figure 1 shows the erosion of a real valued image for different structure elements. It should be noted how the different shape of the structure elements affect the distortion on the image. Regions on the image with high contrast (dots on dice) are "stretched" towards the direction of the structure element.

3. APPLICATION TO HARMONIC/PERCUSSIVE SEPARATION

3.1. Fundamental Morphological Operations on Spectrogram

Since harmonic and percussive components "appear" in the spectrogram as horizontal and vertical lines, the application of morphological filters on the spectrogram with structure elements B_h, B_p that represent horizontal and vertical lines seem practical operations. Let $\mathbf{W}$ denote the magnitude spectrogram of the input signal. From $\mathbf{W}$ we calculate the harmonic/percussive enhanced spectrograms $\tilde{\mathbf{H}}, \tilde{\mathbf{P}}$ by applying erosion / dilation / opening / closing operators with the corresponding structure elements B_h, B_p respectively. B_h corresponds to a horizontal line structure element, while B_p denotes a vertical line element. In particular,

$$\tilde{\mathbf{H}}^D = \mathbf{W} \oplus B_h, \ \tilde{\mathbf{H}}^E = \mathbf{W} \ominus B_h, \ \tilde{\mathbf{H}}^O = \mathbf{W} \circ B_h, \ \tilde{\mathbf{H}}^C = \mathbf{W} \bullet B_h \tag{12}$$

$$\tilde{\mathbf{P}}^D = \mathbf{W} \oplus B_p, \ \tilde{\mathbf{P}}^E = \mathbf{W} \ominus B_p, \ \tilde{\mathbf{P}}^O = \mathbf{W} \circ B_p, \ \tilde{\mathbf{P}}^C = \mathbf{W} \bullet B_p \tag{13}$$

where $type \in \{E, D, O, C\}$ stands for erosion, diation, opening and closing operations. From the enhanced spectrograms the elements of the corresponding masks are computed as follows [9]:

$$\begin{aligned} \mathbf{M}_{f,i}^{H,type} &= \frac{\left(\tilde{\mathbf{H}}_{f,i}^{type}\right)^2}{\left(\tilde{\mathbf{H}}_{f,i}^{type}\right)^2 + \left(\tilde{\mathbf{P}}_{f,i}^{type}\right)^2}, \ type = \{E, D, O, C\} \\ \mathbf{M}_{f,i}^{P,type} &= \frac{\left(\tilde{\mathbf{P}}_{f,i}^{type}\right)^2}{\left(\tilde{\mathbf{H}}_{f,i}^{type}\right)^2 + \left(\tilde{\mathbf{P}}_{f,i}^{type}\right)^2}, \ type = \{E, D, O, C\} \end{aligned} \tag{14}$$

where f, i denote the frequency/time indexes of the spectrogram respectively. Spectrograms of the separated sources are derived from the computed masks as:

$$\mathbf{H}_{f,i} = \mathbf{M}_{f,i}^H \mathbf{W}_{f,i}, \quad \mathbf{P}_{f,i} = \mathbf{M}_{f,i}^P \mathbf{W}_{f,i} \tag{15}$$

The separated sources' signals are finally reconstructed by the inverse STFT as:

$$\tilde{h} = \text{ISTFT}(\mathbf{H} e^{i\angle \mathbf{W}}), \quad \tilde{p} = \text{ISTFT}(\mathbf{P} e^{i\angle \mathbf{W}}) \tag{16}$$

Figure 2 illustrates examples of the application of the various operators on the spectrogram of a music excerpt. As expected, erosion of the spectrogram with a vertical SE results in an image in which all bright and thin horizontal lines have been suppressed as shown in Fig.2b At the same time, many parts of the image have become darker since erosion replace intensity of a point with the minimum value of its neighbors (i.e. defined by the SE). This undesirable effect could be partially restored by following the erosion by dilation with the same SE. As a result (Fig. 2e), opening produces an image that: i) does not include horizontal lines and ii) has similar brightness with the original spectrogram.. The use of dilation as the initial processing step results in an image in which the horizontal lines have been "transformed" to bright almost rectangular areas with height similar to the size of the adopted vertical SE. Consequently, harmonic/percussive separation based on this processed spectrogram

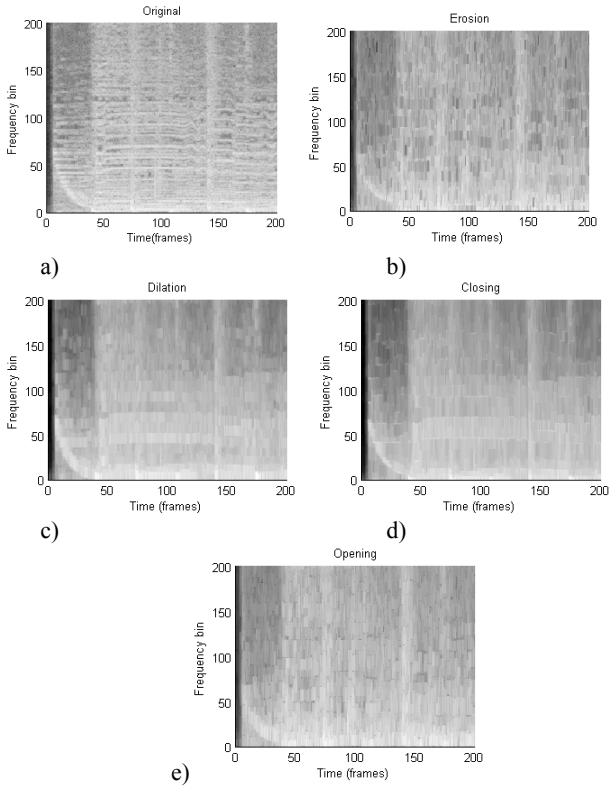

Figure 2. *a) original spectrogram b) after erosion, c) dilation, d) closing and e) opening with the percussive structure element* B_p

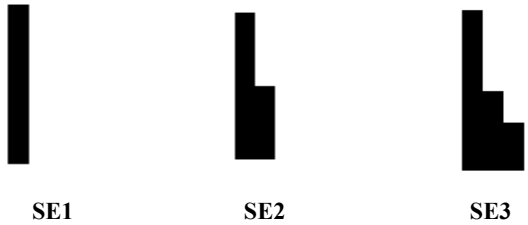

Figure 3. *Different structure elements for percussive masking*

will be insufficient. However applying erosion after dilation (i.e. closing) blurs these bright areas (Fig. 2d).

3.2. Deploying Different Structure Elements

When structure elements B_h, B_p are single lines of dimensions $1\times n$ or $n\times 1$, the operations of dilation, erosion, opening and closing are equivalent to deploying max, min, maxmin, minmax filters respectively to the rows and columns of spectrogram. It could be claimed that the proposed system is very similar to FitzGerald's method [10], since there is a close relation between morphological and median filters [13]. However, the proposed method allows the expansion of structure elements to more directions.

To highlight this, we deployed more complex structure elements for enhancing the percussive component. We observed that while percussive sounds exhibit vertical lines in spectrogram, these lines tend to be "thicker" in low frequencies. Thus we adopt two additional triangular structure elements as shown in

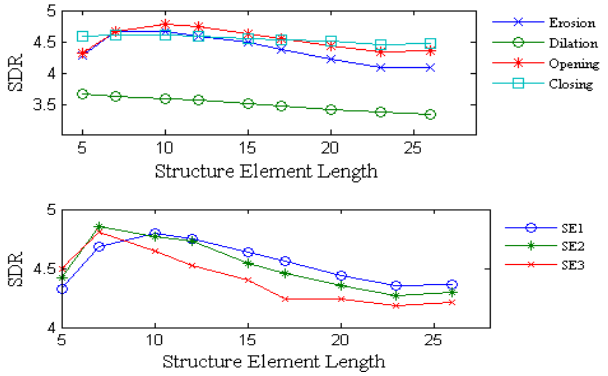

Figure 4. (*a) SDR of the proposed method for all filter types (simple structure element used). (b) SDR for the proposed method for the different structure elements using the opening filter.*

Fig. 3. These elements tend to "capture" better the percussive components of the spectrogram in some cases, while the performance is degraded in other cases, as will be demonstrated in next section. Considering the harmonic components, no such observation was made, and the utilization of more complicated elements than the single line slightly degraded the separation performance.

4. EVALUATION AND EXPERIMENTAL SETUP

The proposed method was evaluated on SISEC 2008/2010 [14] development data for the excerpts that contained percussive instruments. The dataset consists of four music excerpts of professionally recorded songs. For each excerpt the stereo recordings of the separate sources are provided.

The groundtruth data where extracted by mixing the stereo recordings to one channel for each source. Afterwards percussive and harmonic sources were separately mixed to a single channel. The full excerpt mix was obtained by adding the harmonic and percussive signals. For the evaluation we adopted the Signal-to-Distortion Ratio (SDR) evaluation measure as described in [15].

The spectrogram for each excerpt was extracted by applying a shifting Hanning window of 1024 samples length and 512 overlap at sampling frequency 44.1 kHz. Various lengths for the structure elements where used in the range of 5 to 26. Experiments were run applying all image filtering methods mentioned in Section 2 (Figure 2).

Figure 4a shows the SDR for various values of the length of structure elements for the filter operators described in Section 2. The best performance is achieved by the opening filter for a structure size between 7 and 17. When applying closing to the spectrogram, SDR is less sensitive to variations of the element size and approximately 0.2 dB lower for length in the range [7,12]. Erosion operation performs slightly better than closing in some cases, but SDR degrades faster as the length increases. The dilation operation seems impractical since SDR is at least 1dB lower than in all other cases. This can be explained by the fact that dilation introduces noise to the signal as it imposes the maximum value in the mask of the structural element. On the other hand erosion/opening/closing seem more efficient for separating the percussive/harmonic components.

To demonstrate the effect of the structure elements a comparative plot is shown in Figure 4b. When applying structure elements SE2 and SE3, the maximum SDR value is achieved for relatively low length values compared to the single line SE1. In

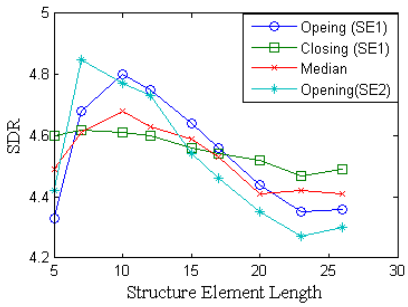

Figure 5. *SDR of the proposed method for opening/erosion filter (simple structure element used) compared to median filter proposed in [10] for various filter sizes.*

other words, it can be stated that the more complex elements require a smaller size to capture better to the harmonic/percussive components. However, when the size of these elements increases, the morphological operations start to introduce a distortion into the spectrogram.

Figure 5 shows the performance of the proposed system compared to the baseline method suggested in [10]. The proposed method performs slightly better when the opening filter is deployed, with the size of the filter being less than 20. In addition, the experiments demonstrate that *minmax, maxmin* filters perform better than *median*, which is the case of using the simple percussive structure element. Moreover the expansion of structure elements to more than one dimension, achieved better SDR values. Furthermore, the computational complexity of the proposed min/max filter is $O(n)$ while for the median filter the complexity is $O(nlogn)$ since it requires implicitly a sorting step. Thus, the proposed method achieves better separation accuracy while requires less computation cost.

5. CONCLUSION

The adoption of nonlinear image filters has not sufficiently been explored in the music processing domain. In this paper, we deployed common image processing techniques to separate the percussive and harmonic components of a music signal. The proposed method achieved better results than the work presented in [9], which can be considered as baseline method. Moreover, it is straightforward, easy to implement and computationally moderate.

The usage of more complex elements seems beneficial, since a) complex elements require a smaller size to better capture the desired signal properties and b) achieve higher performance rates. The exploration of more perceptually valid structure elements and additional filter types should be investigated in the future. Moreover, prior information could be integrated as knowledge to the system, e.g. by creating or generating structure masks from sound templates.

6. REFERENCES

[1] A. Gkiokas, V. Katsouros, G. Carayannis and T. Stafylakis, "Music Tempo Estimation and Beat Tracking by Applying Source Separation and Metrical Relations," in *Proc. Of the 37th IEEE International Conference on Acoustic, Speech and Signal Processing*., Kyoto, Japan, March 25-30, 2012.

[2] D. FitzGerald, E. Coyle, and M. Cranitch, "Using tensor factorisation models to separate drums from polyphonic music," in *Proc. Digital Audio Effects (DAFx-09)*, Como, Italy, 2009.

[3] O. Gillet and G. Richard, "Transcription and separation of drum signals from polyphonic music," *IEEE Transactions on Audio, Speech and Language Processing,* vol. 16, no. 3, pp. 529–540, 2008.

[4] K. Yoshii, M. Goto, and H. Okuno, "Drum sound recognition for polyphonic audio signals by adaptation and matching of spectrogram templates with harmonic structure suppression," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 333–345, 2007.

[5] M. Helen and T. Virtanen, "Separation of drums from polyphonic music using non-negative matrix factorisation and support vector machine," in *Proc. European Signal Processing Conference*, Anatalya, Turkey, 2005.

[6] F. Rigaud, M. Lagrange, A. Robel and G. Peeters, "Drum Extraction from Polyphonic Music Based on a Spectro-Temporal Model of Percussive Sounds," *in Proc. Of the 37th IEEE International Conference on Acoustic, Speech and Signal Processing*, Prague, Czech Republic, May 2011.

[7] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in *Proceedings of the EUSIPCO 2008 European Signal Processing Conference*, Aug. 2008.

[8] N. Ono, K. Miyamoto, H. Kameoka, and S Sagayama, "A real-time equalizer of harmonic and percussive components in music signals," in *Proc. Ninth International Conference on Music Information Retrieval (ISMIR08)*, 2008, pp. 139–144.

[9] N.Q.K Duong, H. Tachibana, E. Vincent, N. Ono, R. Gribonval and S. Sagayama, "Multichannel harmonic and percussive component separation by joint modeling of spatial and spectral continuity," *Proc. Of the 37th IEEE International Conference on Acoustic, Speech and Signal Processing*, Prague, Czech Republic, May 2011.

[10] D. Fitzgerald: "Harmonic/Percussive Separation using Median Filtering," *Intl Conference on Digital Audio Effects (DAFx)*,2010.

[11] B. Thoshkahna and K.R. Ramakrishnan, "A postprocessing technique for improved harmonic/percussion separation for polyphonic music," *Proc. 12th International Conference on Music Information Retrieval (ISMIR11),* Miami, USA, October 2011.

[12] P. Soille, *Morphological Image Analysis Principles and Applications*, (2nd ed.), Springer 2004.

[13] P. Maragos and R. W. Schafer, "Morphological Filters - Part II: Their Relations to Median, Order-Statistic, and Stack Filters," *IEEE Trans. on Acoustics, Speech, and Signal Processing,* vol.35, no.8, pp.1170-1184, Aug. 1987.

[14] E. Vincent, S. Araki, F.J. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, B.V. Gowreesunker, D. Lutter and N.Q.K. Duong, The Signal Separation Evaluation Campaign (2007-2010): Achievements and remaining challenges," *Signal Processing*, 92, pp. 1928-1936, 2012.

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