HARMONIC/PERCUSSIVE SEPARATION USING MEDIAN FILTERING

Derry FitzGerald, *

Audio Research Group
Dublin Institute of Technology
Kevin St.,Dublin 2, Ireland
derry.fitzgerald@dit.ie

ABSTRACT

In this paper, we present a fast, simple and effective method to separate the harmonic and percussive parts of a monaural audio signal. The technique involves the use of median filtering on a spectrogram of the audio signal, with median filtering performed across successive frames to suppress percussive events and enhance harmonic components, while median filtering is also performed across frequency bins to enhance percussive events and supress harmonic components. The two resulting median filtered spectrograms are then used to generate masks which are then applied to the original spectrogram to separate the harmonic and percussive parts of the signal. We illustrate the use of the algorithm in the context of remixing audio material from commercial recordings.

1. INTRODUCTION

The separation of harmonic and percussive sources from mixed audio signals has numerous applications, both as an audio effect for the purposes of remixing and DJing, and as a preprocessing stage for other purposes. This includes the automatic transcription of pitched instruments,key signature detection and chord detection, where elimination of the effects of the percussion sources can help improve results. Similarly, the elimination of the effects of pitched instruments can help improve results for the automatic transcription of drum instruments, rhythm analysis beat tracking.

Recently, the authors proposed a tensor factorisation based algorithm capable of obtaining good quality separation of harmonic and percussive sources [1]. This algorithm incorporated an additive synthesis based source-filter model for pitched instruments, as well as constraints to encourage temporal continuity on pitched sources. A principal advantage of this approach was that it required little or no pretraining in comparison to many other approaches [2, 3, 4]. Unfortunately, a considerable shortcoming of the tensor factorisation approach is that it is both processor and memory intensive, making it impractical for use when whole songs need to be processed, for example such as when remixing a song.

In an effort to overcome this, it was decided to investigate other approaches capable of separating harmonic and percussive components without pretraining, but which were also computationally less intensive. Of particular interest was the approach developed by Ono et al [5]. This technique was based on the intuitive idea that stable harmonic or stationary components form horizontal ridges on the spectrogram, while percussive components form vertical ridges with a broadband frequency response. This can be seen in Figure 1, where the harmonic components are visible as horizontal lines, while the percussive events can be seen as vertical lines. Therefore, a process which emphasises the horizontal lines in the spectrogram while suppressing vertical lines should result in a spectrogram which contains mainly pitched sources, and vice-versa for the vertical lines to recover the percussion sources. To this end, a cost function which minimised the L_2 norm of the power spectrogram gradients was proposed.

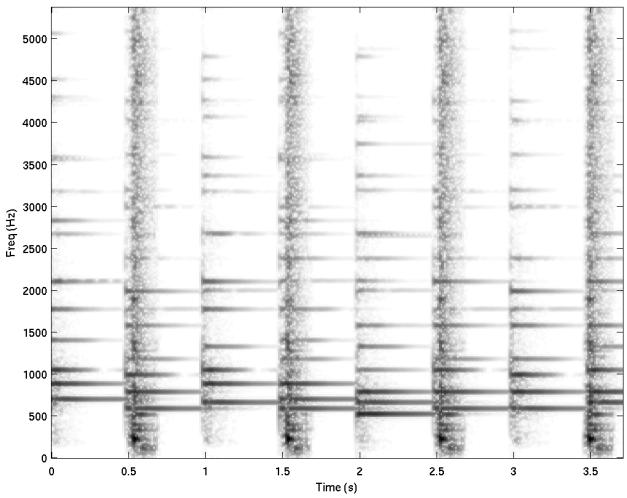

Figure 1: *Spectrogram of pitched and percussive mixture*

Letting $W_{h,i}$ denote the element of the power spectrogram $\mathbf{W}$ of a given signal at frequency bin h and the ith time frame, and similarly defining $H_{h,i}$ as an element of $\mathbf{H}$ the harmonic power spectrogram and $P_{h,i}$ as an element of $\mathbf{P}$ the percussive power spectrogram, the cost function can then be defined as:

$$J(\mathbf{H},\mathbf{P}) = \frac{1}{2\sigma_H^2}\sum_{h,i}(H_{h,i-1}-H_{h,i})^2 + \frac{1}{2\sigma_P^2}\sum_{h,i}(P_{h-1,i}-P_{h,i})^2 \quad (1)$$

where σ_H and σ_P are parameters used to control the weights of the harmonic and percussive smoothness respectively. The cost function is further subject to the additional constraints that

$$H_{h,i} + P_{h,i} = W_{h,i} \quad (2)$$

* This work was supported by Science Foundation Ireland's Stokes Lecturer Program

$$H_{h,i} \geq 0, \quad P_{h,i} \geq 0 \tag{3}$$

In effect, this is equivalent to assuming that the spectrogram gradients $(H_{h,i-1} - H_{h,i})$ and $(P_{h-1,i} - P_{h,i})$ follow gaussian distributions. This is not the case, and so a compressed version of the power spectrogram, $\tilde{\mathbf{W}} = \mathbf{W}^{\gamma}$ where $0 < \gamma \leq 1$, is used instead to partially compensate for this. Iterative update equations to minimise $J(\mathbf{H}, \mathbf{P})$ for $\mathbf{H}$ and $\mathbf{P}$ were then derived, and the recovered harmonic and percussive spectrograms used to generate masks which were then applied to the original spectrogram before inversion to the time domain.

In [6] an alternative cost function based on the generalised Kullback-Liebler divergence was proposed:

$$\begin{aligned} &J_{KL}(\mathbf{H}, \mathbf{P}) \\ &= \sum_{h,i} \{W_{h,i} \log \frac{W_{h,i}}{H_{h,i} + P_{h,i}} - W_{h,i} + H_{h,i} + P_{h,i}\} \\ &+ \frac{1}{2\sigma_H^2} \sum_{h,i} (\sqrt{H_{h,i-1}} - \sqrt{H_{h,i}})^2 \\ &+ \frac{1}{2\sigma_P^2} \sum_{h,i} (\sqrt{P_{h-1,i}} - \sqrt{P_{h,i}})^2 \end{aligned} \tag{4}$$

and new update equations for $\mathbf{H}$ and $\mathbf{P}$ derived from this cost function. A real-time implementation of the algorithm using a sliding block analysis, rather than processing the whole signal, was also implemented and described in [6].

The system was shown to give good separation performance at low computational cost, thereby making it suitable as a preprocessor for other applications. Further, the underlying principle of the algorithm represents a simple intuitive idea that can be used to derive alternate means of separating harmonic and percussive components, as will be seen in the next section.

2. MEDIAN FILTERING BASED SEPARATION

As was shown previously, regarding percussive events as vertical lines and harmonic events as horizontal lines in a spectrogram is a useful approximation when attempting to separate harmonic and percussive source. Taking the percussive events as an example, the algorithms described above in effect smooth out the frequency spectrum in a given time frame by removing large "spikes" in the spectrum which correspond to harmonic events. Similarly, harmonic events in a given frequency bin are smoothed out by removing "spikes" related to percussive events. Another way of looking at this is to regard harmonic events as outliers in the frequency spectrum at a given time frame, and to regard percussive events as outliers across time in a given frequency bin. This brings us to the concept of using median filters individually in the horizontal and vertical directions to separate harmonic and percussive events.

Median filters have been used extensively in image processing for removing speckle noise and salt and pepper noise from images [7]. Median filters operate by replacing a given sample in a signal by the median of the signal values in a window around the sample. Given an input vector $x(n)$ then $y(n)$ is the output of a median filter of length l where l defines the number of samples over which median filtering takes place. Where l is odd, the median filter can be defined as:

$$y(n) = \text{ median } \{x(n-k : n+k), k = (l-1)/2\} \tag{5}$$

In effect, the original sample is replaced with the middle value obtained from a sorted list of the samples in the neighbourhood of the original sample. In cases where l is even, the median is obtained as the mean of the two values in the middle of in the sorted list. As opposed to moving average filters, median filters are effective in removing impulse noise because they do not depend on values which are outliers from the typical values in the region around the original sample.

A number of examples are now presented to illustrate the effects of median filtering in suppressing harmonic and percussive events in audio spectrograms. Figure 2(a) shows the plot of a frequency spectrum containing a mixture of noise from a snare drum and notes played by a piano. The harmonics from the piano are clearly visible as large spikes in the spectrum. Figure 2(b) shows the same spectrum after median filtering with a filter length of 17. It can be seen that the spikes associated with the harmonics have been suppressed, leaving a spectrum where the drum noise now predominates. Similarly, Figure 3(a) shows the output of a frequency bin across time, again taken from a mixture of snare drum and piano. The onset of the snare is clearly visible as a large spike in energy in the frequency bin, while the harmonic energy is more constant over time. Figure 3(b) shows the output of the frequency bin after median filtering, and it can be appreciated that the spike associated with the onset is removed by median filtering, thereby suppressing the energy due to the percussion event.

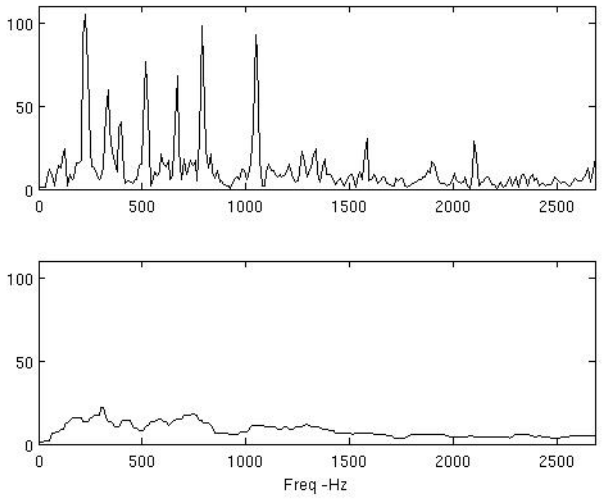

Figure 2: *Spectrogram frame containing mixture of piano and snare a) Original spectrum b) Spectrum after median filtering*

Given an input magnitude spectrogram $\mathbf{S}$, and denoting the ith time frame as S_i, and the hth frequency slice as S_h a percussion-enhanced spectrogram frame P_i can be generated by performing median filtering on S_i:

$$P_i = \mathcal{M}\{S_i, l_{perc}\} \tag{6}$$

where $\mathcal{M}$ denotes median filtering and l_{perc} is the filter length of the percussion-enhancing median filter. The individual percussion-enhanced frames P_i are then combined to yield a percussion-enhanced spectrogram $\mathbf{P}$. Similarly, a harmonic-enhanced spectrogram frequency slice H_h can be obtained by median filtering frequency

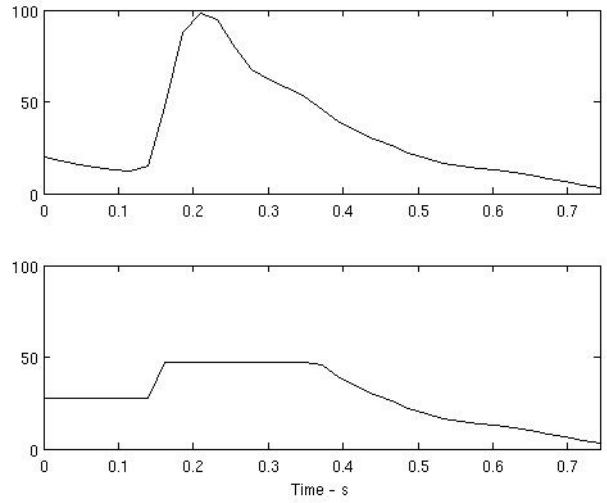

Figure 3: *Spectrogram frequency slice containing mixture of piano and snare a) Original slice b) Slice after median filtering*

slice S_h.

$$H_i = \mathcal{M}\{S_h, l_{harm}\} \tag{7}$$

where l_{harm} is the length of the harmonic median filter. The slices are then combined to give a harmonic enhanced spectrogram $\mathbf{H}$.

The resulting harmonic and percussion suppressed spectrograms can then be used to generate masks which can then be applied to the original spectrogram. Two families of masks were then investigated for the separation of the sources. The first of these is a hard or binary mask, where it is assumed that each frequency bin in the spectrogram belongs either to the percussion or to the harmonic source. In this case, the masks are defined as:

$$\mathbf{M}_{\mathbf{H}h,i} = \begin{cases} 1, & \text{if } \mathbf{H}_{h,i} > \mathbf{P}_{h,i} \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

$$\mathbf{M}_{\mathbf{P}h,i} = \begin{cases} 1, & \text{if } \mathbf{P}_{h,i} > \mathbf{H}_{h,i} \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

The second family of masks are soft masks based on Wiener Filtering and are defined as:

$$\mathbf{M}_{\mathbf{H}h,i} = \frac{\mathbf{H}_{h,i}^p}{(\mathbf{H}_{h,i}^p + \mathbf{P}_{h,i}^p)} \tag{10}$$

$$\mathbf{M}_{\mathbf{P}h,i} = \frac{\mathbf{P}_{h,i}^p}{(\mathbf{H}_{h,i}^p + \mathbf{P}_{h,i}^p)} \tag{11}$$

where p denotes the power to which each individual element of the spectrograms are raised. Typically p is given a value of 1 or 2.

Complex spectrograms are then recovered for inversion from:

$$\hat{\mathbf{H}} = \hat{\mathbf{S}} \otimes \mathbf{M}_{\mathbf{H}} \tag{12}$$

and

$$\hat{\mathbf{P}} = \hat{\mathbf{S}} \otimes \mathbf{M}_{\mathbf{P}} \tag{13}$$

where $\otimes$ denotes elementwise mulitiplication and where $\hat{S}$ denotes the original complex valued spectrogram. These complex spectrograms are then inverted to the time domain to yield the separated harmonic and percussive waveforms respectively.

In comparison to the iterative approach developed by Ono et al., which typically requires 30-50 iterations to converge, only two passes are required through the input spectrogram, one each for $\mathbf{H}$ and $\mathbf{P}$. This means that the median filter based algorithm is faster, which is of considerable benefit when used as preprocessing for other tasks. In tests, the proposed algorithm performs approximately twice as fast as that of Ono et al with the number of iterations set to 30. This raises the possibility of performing real-time harmonic/percussive separation on stereo files, as the proposed algorithm can easily be extended to handle stereo signals.

3. SEPARATION AND REMIXING EXAMPLES

We now present examples of the use of the median filtering harmonic/percussive separation algorithm. Figure 4 shows an excerpt from "Animal" by Def Leppard, as well as the separated harmonic and percussive waveforms respectively, obtained using a median filter of length 17 for both harmonic and percussive filters, as well as using a soft mask with $p = 2$. It can be seen that the recovered harmonic waveform contains little or no evidence of percussion events, while the percussive waveform contains little or no evidence of the harmonic instruments. On listening to the waveforms, some traces of the drums can be heard in the harmonic waveform, though at a very reduced level, while the attack portion of some of the instruments such as guitar has been captured by the percussive waveform, as well as traces of some guitar parts where the pitch is changing constantly. This is to be expected as the attacks of many instruments such as guitar and piano can be considered as percussive in nature, and as the algorithm assumes that the pitched instruments are stationary in pitch. This also occurs in other algorithms for separating harmonic and percussive components.

Also shown in Figure 4 are remixed versions of the original signal, the first has the percussion components reduced by 6dB, while the second shows the harmonic components reduced by 6dB. On listening to these waveforms, there are no noticeable artifacts in the resynthesis, while the reduction in amplitude of the respective sources can clearly be heard. This demonstrates that the algorithm is capable of generating audio which can be used for high-quality remixing of the separated harmonic and percussive sources.

Figure 5 shows an excerpt from "Billie Jean" by Michael Jackson, the separated harmonic and percussive waveforms, and remixed versions with the percussion reduced by 6dB, and the harmonics reduced by 6dB respectively. Again, the algorithm can be seen to have separated the harmonic and percussive parts well. On listening, the attack of the bass has been captured by the percussive part, and a small amount of drum noise can be heard in the harmonic waveform. In the remixed versions, no artifacts can be heard.

Both of the above examples were carried out using an FFT size of 4096, with a hopsize of 1024, with a sampling frequency of 44.1 kHz. Testing showed that better separation quality was achieved at larger FFT lengths. The median filter length was set to 17 for both the harmonic and percussive filters, and testing showed that once the median filter lengths were above 15 and below 30,the separation quality did not vary dramatically, with good separation achieved in most cases. Further, informal listening tests suggest that the quality of separation is comparable to that achieved by the algorithms proposed by Ono et al. The use of the soft masking was found to result in less artifacts in the resynthesis, though at the expense of a slight increase in the amount of interference between the percussive and harmonic sources. In general, it was observed that setting $p = 2$ gave considerably better separation results than us-

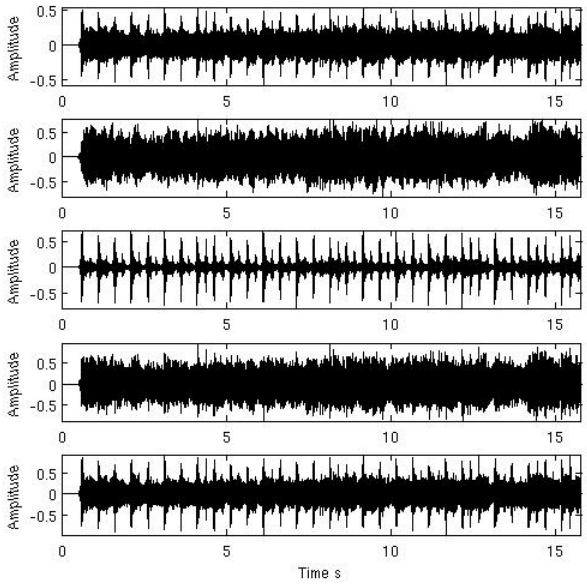

Figure 4: *Excerpt from "Animal" by Def Leppard a) Original waveform, b) Separated harmonic waveform, c) Separated percussive waveform, d) Remix, percussion reduced by 6dB, e) Remix, harmonic components reduced by 6dB*

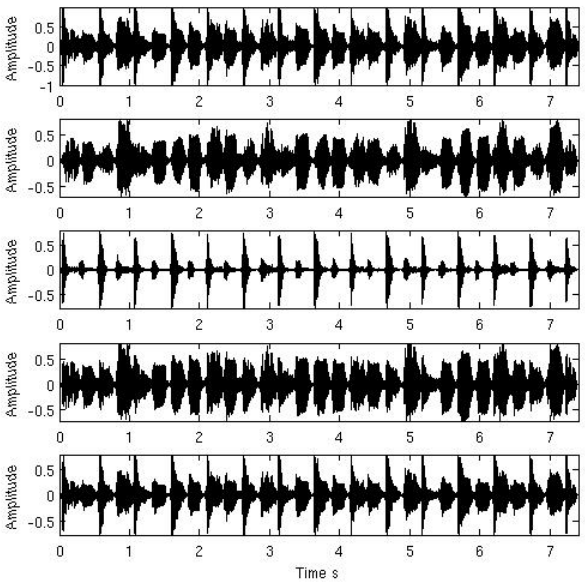

Figure 5: *Excerpt from "Billie Jean" by Michael Jackson a) Original waveform, b) Separated harmonic waveform, c) Separated percussive waveform, d) Remix, percussion reduced by 6dB, e) Remix, harmonic components reduced by 6dB*

ing $p = 1$. Audio examples are available for download at `http://eleceng.dit.ie/derryfitzgerald/index.php?uid=489&menu_id=42`

4. CONCLUSIONS

Having described an fast effective method of harmonic-percussive separation developed by Ono et al [6], which is based on the idea that percussion events can be regarded as vertical lines, and harmonic or stationary events as horizontal lines in a spectrogram, we then took advantage of this idea to develop a simpler, faster and more effective harmonic/percussive separation algorithm. This was based on the idea that harmonics could be regarded as outliers in a spectrum containing a mixture of percussion and pitched instruments, while percussive onsets can be regarded as outliers in a frequency slice containing a stable harmonic or stationary event.

To remove these outliers, we then used median filtering, as median filtering is effective at removing outliers for the purposes of image denoising. The resulting harmonic-enhanced and percussion-enhanced spectrograms were then used to generate masks which were then applied to the original spectrogram to separate the harmonic and percussive components. Real-world separation and remixing examples using the algorithm were then discussed.

Future work will concentrate on developing a real-time implementation of the algorithm and on investigating the use of the algorithm as a preprocessor for other tasks such as key signature detection and chord detection, where suppression of percussion events is helpful in improving results. Further, the use of rank-order filters, where a different percentile other than 50, which is used in median filtering will be investigated as a means of potentially improving the separation performance of the algorithm.

5. REFERENCES

[1] 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.

[2] 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.

[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] 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.

[5] 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.

[6] 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.

[7] R. Jain, R. Kasturi, and B. Schunck, *Machine Vision*, McGraw-Hill, 1995.