2012

Vocal Separation Using Nearest Neighbours and Median Filtering

Derry Fitzgerald
Technological University Dublin, derry.fitzgerald@tudublin.ie

Funder: SFI

Vocal Separation using Nearest Neighbours and Median Filtering

Derry FitzGerald[†]

Audio Research Group,
Dublin Institute of Technology,
IRELAND

E-mail: [†]derry.fitzgerald@dit.ie

Abstract — **Recently, single channel vocal separation algorithms have been proposed which exploit the fact that most popular music can be regarded as a repeating musical background over which a locally non-repeating vocal signal is superimposed. In this paper we describe a novel vocal separator inspired by these approaches which finds the k nearest neighbours to each frame of a spectrogram of the mixture signal. The median value of these frames is then used as the estimate of the background music at the current frame. This is then used to generate a mask on the original complex-valued spectrogram before inversion to the time domain. The effectiveness of the approach is demonstrated on a number of real-world signals.**

Keywords — **Sound Source Separation, Vocal Extraction, Median filtering**

I Introduction

Recently, the topic of extracting vocals from audio mixtures such as commercial pop songs has received much attention. This has numerous applications including automatically creating karaoke tracks, automatic melody transcription, singer identification, automatic alignment of lyrics to music, remixing and sampling for use in new compositions.

A wide variety of approaches have been used to try to extract vocals, including factorisation-based approaches [1, 2] where training material was required to distinguish between vocal basis functions and non-vocal basis functions, allowing the vocals to be separated. Other approaches made use of a predominant melody estimation algorithm to detect the vocal melody, which was then used to enable separation of the vocal in conjunction with techniques which identified vocal and non-vocal regions of the mixture signal [3, 4]. A median filtering-based approach to vocal separation was described in [5], which made use of the fact that vocals appear as pitched instrument-like harmonics at low frequency resolution, while appearing as broadband noise at higher frequency resolutions. This technique also made use of various matrix factorisation algorithms as post-processing steps in order to further improve the results.

More recently, approaches have been proposed based on the idea of repetition in audio signals [6, 7]. The underlying idea here is that many recordings of popular music can be viewed as having a repeating musical structure in the background accompaniment. Over this, the vocal signal occurs without any immediate repeating structure, though obviously repetition of vocal melodies and lyrics can occur, but at a much larger timescale than that of the background music. Further, it was also assumed that the vocal was sparse in the time-frequency domain, and so the number of time-frequency bins in which the vocal is active are very much less than the total number of bins.

In [7], a beat spectrogram was calculated to estimate a local periodicity for the spectrogram of the mixture signal. Once this period had been identified, frames at integer multiples of this period around the current frame being processed were collected together, as it was assumed that these frames would be similar in nature due to the local repetition of the background music. The median

value of these frames at each frequency bin was then taken as the estimate of the background music for the current frame. The motivation for using the median filter lies in the observation that when vocal energy occurs it will not follow the repeating pattern of the background music, and so bins with significant vocal energy will appear as outliers, which median filtering is particularly suited to removing [8]. The resulting background music spectrogram was then used to create a mask which was applied to the original complex-valued spectrogram before inversion to the time domain. This approach was demonstrated to give separation quality comparable to that in [5], but at a much reduced computational cost.

This paper describes a vocal separation technique inspired by the above repetition based approach. The proposed algorithm is described in section 2, while section 3 describes the use of the technique on a number of real-world examples. Section 4 then contains evaluation of the technique proposed on a test set of real world recordings. Finally, section 5 contains conclusions and areas for future research.

II Vocal separation algorithm

The vocal separation technique proposed in this paper initially takes the input mixture signal and obtains a magnitude spectrogram by performing a short-time Fourier transform on the signal. Instead of assuming local periodicity as was done in [7], we here make the assumption that the local periodicity is not necessary. Instead we focus on finding the most similar frames to the current frame being analysed using a suitable distance metric. This is because the closest frames to any given frame may not occur within a short distance of that frame, instead it may occur in another verse or chorus of the song, and so calculating distances globally in the spectrogram as opposed to locally should give an advantage when attempting to estimate the background music.

Assuming that the vocal signal is sparse and is non-repeating, then the effects of the background music will predominate when calculating the distance between frames, as there will only be a small number of bins in which vocal energy is present, in comparison to the total number of bins at any given frame. This means that the distance measure should chiefly calculate the distance between the background music occurring in any pair of frames in the mixture magnitude spectrogram.

The distance metric chosen here is the squared Euclidean distance between the frames, given by:

$$D_{k,l} = \sum (\mathbf{X}_k - \mathbf{X}_l)^2 \tag{1}$$

Here $\mathbf{X}$ is the mixture magnitude spectrogram of size $n \times m$ where n is the number of frequency bins and m is the number of time frames. $\mathbf{X}_k$ denotes the kth spectrogram frame of the magnitude spectrogram, $D_{k,l}$ denotes the squared euclidean distance between frames k and l, and summation occurs over all n frequency bins.

Calculating the distance between all frames results in a symmetric matrix $\mathbf{D}$ of size $n \times n$. This matrix is then sorted in ascending order, and the frame indices obtained of the p nearest neighbours to the current (kth) frame. These frames are then extracted from the magnitude spectrogram and stored in a $n \times p$ matrix $\mathbf{P}$. The background music estimate for the kth frame is then estimated as:

$$\mathbf{Y}_k = \mathcal{M}(\mathbf{P}) \tag{2}$$

where $\mathbf{Y}_k$ is the kth frame of the estimated background music spectrogram $\mathbf{Y}$, and where $\mathcal{M}$ denotes the median operator.

In line with [7] we assume that the background music cannot have a greater energy at a given time-frequency bin than that of the mixture signal and so further processing is carried out on $\mathbf{Y}$ to eliminate any values which are greater than those of the original mixture:

$$\mathbf{Y}_{f,k} = \min(\mathbf{X}_{f,k}, \mathbf{Y}_{f,k}) \tag{3}$$

where f denotes the fth frequency bin and k the kth time frame.

A binary mask could then be obtained for separating the background music from the vocal signal by comparing the values in $\mathbf{X}$ with those in $\mathbf{Y}$. For time-frequency bins where there is no vocal energy present, if the model of the background music is correct then $\mathbf{X}_{f,k} \approx \mathbf{Y}_{f,k}$ and for bins with significant vocal energy $\mathbf{X}_{f,k} \gg \mathbf{Y}_{f,k}$. Therefore, a suitably chosen threshold should be able to discriminate between bins containing vocal energy from those containing background music. However, the use of binary masking can often introduce audible artifacts in the resynthesis, and so we chose to follow the soft masking approach used by Liutkus et al in [7], which forms a mask based on a Gaussian radial basis function approach:

$$\mathbf{W}_{f,k} = \exp\left(-\frac{(\log \mathbf{X}_{f,k} - \log \mathbf{Y}_{f,k})^2}{2\lambda^2}\right) \tag{4}$$

where $\mathbf{W}$ is a soft mask to be applied to the original complex-valued spectrogram and λ is a tolerance parameter which can be used to control the weights obtained in the mask.

The complex-valued background music spectrogram $\mathbf{B}$ can then be estimated as:

$$\mathbf{B} = \mathbf{W} \otimes \mathbf{R} \tag{5}$$

where $\mathbf{R}$ denotes the original complex-valued mixture spectrogram and $\otimes$ denotes elementwise multiplication. The background music signal can then

be recovered via an inverse short-time Fourier transform.

Similarly, the complex-valued vocal spectrogram $\mathbf{V}$ can then be estimated from:

$$\mathbf{V} = (1 - \mathbf{W}) \otimes \mathbf{R} \tag{6}$$

where all operations are carried out elementwise. Again, the vocal signal can then be recovered using the inverse short-time Fourier transform.

A simple post-processing step which typically improves the results further is a low pass filtering approach which removes all frequencies below 100 Hz from the vocal signal and to add these frequencies back into the background track as was done in [5]. The effects of doing this will be evaluated later in section 4.

The technique described is a very simple approach which can be implemented in a computationally efficient manner. This means that is is capable of running considerably faster than other approaches such as those described in [5].

III Separation Examples

We now present the use of the example on a real-world example, from "Wouldn't it be nice" by the Beach Boys. Figure 1 shows the mixture spectrogram from a piece of music containing vocals and multiple instruments, both pitched and percussive. Figure 2 shows the original unmixed vocal spectrogram, while figure 3 then shows the original unmixed instrumental spectrogram. Figure 4 then shows the separated vocal as obtained using the nearest neighbour median filtering approach, while figure 5 shows the separated backing track spectrogram. It should be noted that low pass filtering was not used in this example. The test signal had a sampling rate of 44.1 kHz, and an fft/window size of 4096 samples and a hopsize of 1024 samples was used. The number of nearest neighbours was set to $p = 80$, with $\lambda = 1$.

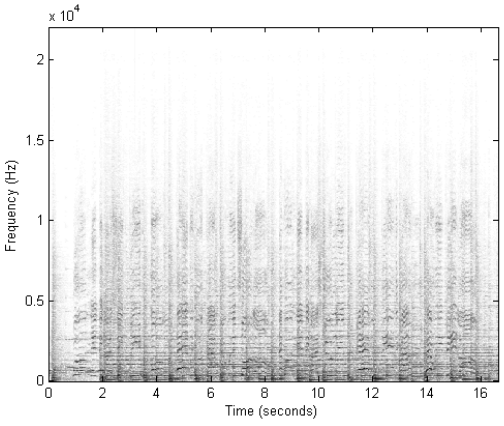

Fig. 1: Original Mixture spectrogram.

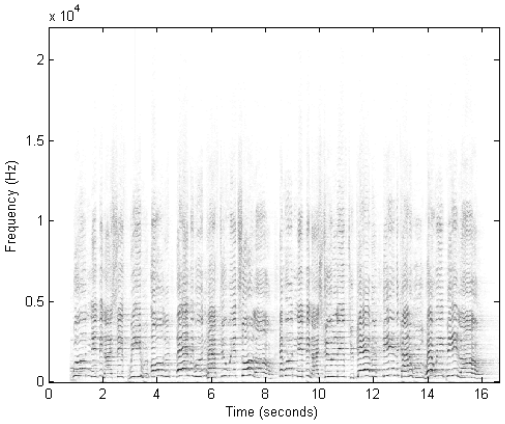

Fig. 2: Original vocal spectrogram.

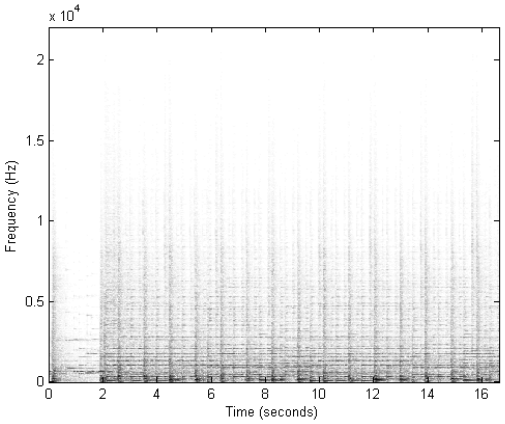

Fig. 3: Original instrumental spectrogram.

It can be seen that the separated vocal spectrogram obtained using the nearest neighbour median filtering algorithm has successfully captured the main characteristics of the original vocal spectrogram, though there are still some traces of the other instruments in the separated spectrogram. Similarly, the separated instrumental spectrogram has captured the main characteristics of the original instrumental spectrogram though close inspection will reveal the presence of traces of the vocals. On listening to the separated sources, the vocals clearly predominate in the vocal signal, though there is some audible interference from the other instruments. Similarly in the separated instrumental signal, the vocals can still be heard, though at a greatly reduced amplitude. This demonstrates that the technique is capable of working on real-world signals. Audio examples can be found at `http://eleceng.dit.ie/derryfitzgerald/index.php?uid=489&menu_id=65`.

IV Separation Results

Having demonstrated the use of the nearest neighbour median filtering approach on a real world sig-

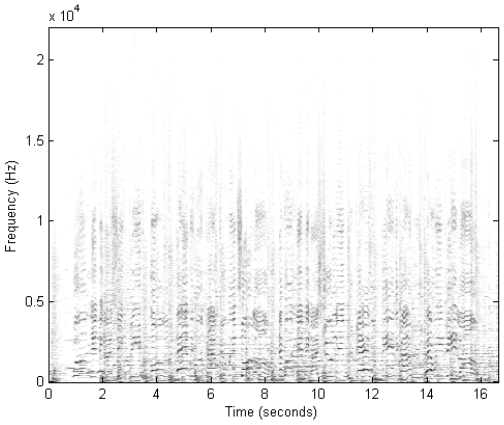

Fig. 4: Separated vocal spectrogram.

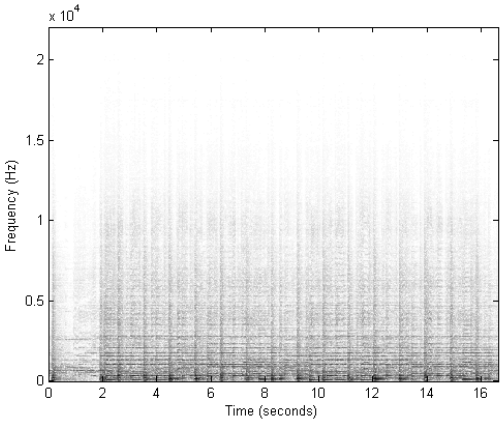

Fig. 5: Separated instrumental spectrogram.

nal, we now evaluate its performance on a set of 10 real-world audio examples. The test set used was that used to evaluate the algorithms in [5], and consisted of mixtures made from separated backing tracks and vocal tracks available in [10, 11]. The mixes were made at 3 different levels to see how the algorithm would perform under a range of circumstances. The first set dropped the level of the vocal tracks by 6dB to allow the intrumental track to predominate, the second set were mixed with no level changes to the signal, while the third increased the level of the vocals by 6dB to model cases where the vocal signal predominates.

Quantitative evaluation of the separation performance of these mixtures was carried out using the metrics defined in [12], and were evaluated using the toolbox available at [13]. The metrics are the Signal to Distortion ratio (SDR), which measures the overall separation quality, Signal to Interference ratio (SIR), which measures the amount of interference due to the presence of other sources in the separated signal, and Signal to Artifacts ratio (SAR), which measures the artifacts present in the signal due to the separation algorithm and the

	SIR	SAR	ISR
-6dB V	-1.00	13.25	-0.06
0dB V	1.83	19.81	2.12
6dB V	2.83	17.17	3.31
-6dB T	7.42	11.50	9.99
0dB T	3.19	6.59	6.89
6dB T	-2.57	0.16	3.88

Table 1: Performance Evaluation of Multipass Median filtering approach. These are the best results obtained from all the techniques described in [5]. V denotes separated vocal signals, T denotes separated backing track signal. -6dB indicates mixes made by reducing the level of the vocals by 6bB, 0dB indicates the signals were mixed with no gain changes, and 6dB indicates mixes made by boosting the level of the vocals by 6dB

resynthesis technique used. Evaluation was carried out on both the separated vocal and instrumental tracks, and the parameters of the algorithm were as described for the example given in section 3.

Table 1 shows the best results obtained from the various techniques described in [5], to provide a baseline against which to measure the performance of the algorithm. Table 2 then shows the results obtained using the nearest neighbour median filtering approach described in the paper, while table 3 then shows the results obtained when using low pass filtering as a post-processing stage.

It can be seen that the techniques proposed here give better vocal separation than those in [5] in all cases, with the low pass filtering post-processing improving the results obtained using nearest neighbour median filtering algorithm. In the case of the separations for the instrumental tracks, the results are still good, but are worse for both the 0dB and 6dB cases. Nevertheless, it should be pointed out that these results were obtained at a much lower computational load, and that applying the matrix factorisation-based post-processing approaches detailed in [5] would further improve the results obtained, though at the expense of considerably increased computational load.

Overall, the results show that the proposed algorithm is more than competitive when separating vocals and at a much reduced computational load. In fact the algorithm runs approximately 8 times faster than the reference implementation described in [5]. This demonstrates the efficacy of the simple approach to separating vocals detailed in this paper.

V Conclusions

Having outlined previous approaches to vocal separation, we described a novel approach to vocal separation inspired by recent repetition based ap-

	SIR	SAR	ISR
-6dB V	-0.15	6.54	2.53
0dB V	1.92	10.08	3.51
6dB V	2.67	12.60	3.60
-6dB T	7.92	10.55	12.05
0dB T	2.23	3.86	9.19
6dB T	-4.40	-3.15	6.89

Table 2: Performance Evaluation of Nearest Neighbour-Median Filtering algorithm. Legend is as per Table 1

	SIR	SAR	ISR
-6dB V	0.17	8.08	2.31
0dB V	2.18	12.05	3.35
6dB V	2.89	15.39	3.48
-6dB T	8.10	10.74	12.25
0dB T	2.56	4.19	9.41
6dB T	-3.69	-2.44	7.11

Table 3: Performance Evaluation of Nearest Neighbour-Median Filtering algorithm plus low pass filtering. Legend is as per Table 1

proaches to vocal separation. The approach finds the nearest neighbours to any given frame in the mixture spectrogram, and uses the median of these frames to identify the background musical accompaniment at the current frame. This estimate of the background music spectrogram was then used to generate masks to apply to the original complex-valued spectrogram to allow resynthesis of the separated sources. The described approach is simple and can be implemented in a computationally efficient manner.

The effectiveness of the approach was then demonstrated on a set of real world examples, and the algorithm was found to outperform another recent vocal separation algorithm for the purposes of vocal separation. Future work will concentrate on improving the separation performance by incorporating various matrix factorisation-based techniques as post-processing to improve on the results obtained.

VI Acknowledgements

Derry FitzGerald was supported in this research by Science Foundation Ireland's Stokes Lectureship Programme.

References

[1] A. Ozerov, P. Phillipe, F. Bimbot, and R. Gribonval, *Adaption of Bayesian models for single channel source separation and its application to voice/music separation in popular songs*, IEEE Transactions on Audio Speech and Language Processing, 2007.

[2] S. Vembu and S. Baumann, *Separation of vocals from polyphonic audio recordings*, in Proc. Int. Symp. Music Inf. Retrieval (ISMIR05), 2005, pp. 337344.

[3] Y. Li and D. Wang, *Separation of Singing Voice from music accompaniment for Monaural Recordings* IEEE Transactions on Audio Speech and Language Processing, 2006.

[4] C. Hsu and J. Jang, *On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset*, IEEE Transactions on Audio Speech and Language Processing, 2010.

[5] D. FitzGerald and M. Gainza, *Single Channel Vocal Separation using Median Filtering and Factorisation Techniques*,ISAST Transactions on Electronic and Signal Processing , No. 1, Vol. 4, pages: 62 - 73, 2010

[6] Z. Rafii and B. Pardo, *A simple music/voice separation method based on the extraction of the repeating musical structure*, In IEEE International Conference on Acoustics, Speech and Signal Processing, 2011.

[7] A. Liutkus and Z. Rafii and R. Badeau and B. Pardo and G. Richard, *Adaptive filtering for music/voice separation exploiting the repeating musical structure*, In IEEE International Conference on Acoustics, Speech and Signal Processing, 2012.

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

[9] D. FitzGerald, *Harmonic/Percussive Separation using Median Filtering*, Proc. of the 13th International Conference on Digital Audio Effects (DAFx-10), Graz, Austria 2010.

[10] The Beach Boys, *Good Vibrations: Thirty Years Of The Beach Boys*, Capitol Records, Capitol C2 0777 7 81294 2 4, 1993.

[11] The Beach Boys, *The Pet Sounds Sessions*,Capitol Records, Capitol 7243 8 37662 2 2, 1997

[12] E. Vincent, R. Gribonval and C. Fvotte. *Performance measurement in Blind Audio Source Separation*, IEEE Trans. Audio, Speech and Audio Processing, vol. 14, no. 4, pp. 1462-1469, Jul. 2006.

[13] BSS_Eval toolbox available at http://bass-db.gforge.inria.fr/bss_eval/