STEREO VOCAL EXTRACTION USING ADRESS AND NEAREST NEIGHBOURS MEDIAN FILTERING

Derry FitzGerald, *

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

ABSTRACT

An efficient and effective stereo vocal extraction algorithm is presented, which combines two existing approaches. A Nearest Neighbours Median Filtering algorithm is used to separate the vocals and the instrumental backing track from the stereo mixture. The separated vocal track is then passed through a mask generated by the Adress algorithm and high-pass filtered to extract the vocals. The separated instrumental backing track is then improved by adding to it the residual backing track energy extracted by Adress. Also investigated is a variant on this algorithm which uses a difference spectrogram to calculate the nearest neighbours. The effectiveness of these algorithms is then demonstrated on a test dataset, and results show that the proposed algorithms give performance comparable to the state of the art, but at a low computational cost.

1. INTRODUCTION

Vocal extraction has been the subject of much interest in recent years due to its numerous applications. These include uses in melody line extraction, query by humming systems, automatically aligning lyrics to audio, singer identification, and also for obtaining "samples" to be repurposed in new musical pieces.

A wide variety of different techniques have been utilised in attempting to tackle this problem. Li and Wang made use of a vocal/non-vocal region classifier in conjunction with a predominant melody estimator to separate the vocal melody [1]. Ozerov et al made use of pre-trained Bayesian models of vocals and instrumental tracks which were then adapted to the mixture signal, using vocal/non-vocal classification to aid the algorithm to perform vocal separation [2]. Vembu et al [3] again used a vocal/non-vocal classifier followed by a factorisation technique to perform vocal separation, as did Raj et al [4]. However, a shortcoming of the previous three techniques was that they depended on there being sufficient non-vocal regions to allow successful separation. Durrieu et al proposed a source-filter model to extract the melody, combined with Non-negative Matrix Factorisation to model the backing track [5]. More recently, techniques have been proposed which use the fact that the instrumental backing track is more repetitive than the main melody to extract vocals and lead instruments from musical signals [6],[7], which were shown to give good results.

Other techniques were also proposed which specifically require the use of stereo mixtures. The Adress algorithm [8] separates sources based on their pan position in the stereo field and has been used to separate vocals from stereo signals, and a variant on this algorithm has been used commercially for vocal removal for karaoke games. A problem with Adress is that there are often multiple instruments in the same position in the stereo field as the vocals, such as bass guitar and drums. More recently, the SEMANTICS algorithm proposed by Sofianos et al [9] has been used for stereo vocal extraction. Here, Independent Component Analysis is performed on the input stereo signal, resulting in one signal containing vocals and some instruments, and another containing instruments. The non-vocal signal is then used to remove these instruments from the original stereo signal before an amplitude thresholding step is used further reduce the presence of instruments in the signal, resulting in a separated vocal signal.

The focus in this paper is on developing a computationally efficient but effective technique for the extraction of vocals from stereo signals, and to this end, a technique combining two previously proposed algorithms has been implemented. The first of these is the Adress algorithm and the second is the Nearest Neighbours Median Filtering vocal separation algorithm described in [6]. The following two sections describe these algorithms, before the combination of these two algorithms is described in Section 4. Following from that, an evaluation of the proposed algorithm is presented in Section 5, before conclusions and future work are discussed in Section 6.

2. THE ADRESS ALGORITHM

The Adress algorithm performs sound source separation on stereo audio signals[8]. It assumes a linear instantaneous mixing model, and has been demonstrated to perform well on a wide range of recorded music. It assumes that each source occupies a unique point in the stereo field, and works by estimating sets of time-frequency bins which are associated with a given pan position. However, as previously noted, a shortcoming of the algorithm lies in the fact that instruments which occur at the same point in the stereo field will be separated together. This is often the case in modern commercial releases, for example, the centre position in the stereo field typically contains vocals, bass guitar, snare and kick drums.

The linear instantaneous mixing model used in Adress is given by:

$$L(t) = \sum_{j=1}^{J} Pl_j S_j(t) \tag{1}$$

$$R(t) = \sum_{j=1}^{J} Pr_j S_j(t) \tag{2}$$

* This work was supported by the Science Foundation Ireland Stokes Lectureship programme.

with S_j indicating the jth source, Pl_j and Pr_j, the panning coefficients for the jth source, and L and R indicating the left and right channel mixtures respectively. Let the amplitude ratio of a given source be defined as:

$$I_j = \frac{Pl_j}{Pr_j} \tag{3}$$

It can be seen that, given linear instantaneous mixing, $L - I_j R$ cancels the jth source from the mixture. However, this cancellation technique does not allow recovery of the cancelled source. Instead the cancelled source is estimated using time-frequency domain techniques.

A Short Time Fourier Transform (STFT) is carried out on each of the two mixture signals. Let β refer to the user-chosen azimuth resolution, where here azimuth is used to indicate position in the stereo space defined across the two channels. This determines how many equally spaced gain scaling values are used to create the frequency-azimuth plane across the stereo field. The amplitude ratio I, as previously defined is not bounded if Pr_j is 0, and so a bounded gain scale vector g is defined. The gains g for a given azimuth resolution are defined as:

$$g_i = \begin{cases} \dfrac{i}{\beta} & \text{if} \quad i \leq \beta/2 \\ \dfrac{\beta - i}{\beta} & \text{if} \quad i > \beta/2 \end{cases} \tag{4}$$

where $0 \leq i \leq \beta$ and where i and β are integers. A position index is then defined as:

$$P_i = \begin{cases} g_i - 1 & \text{if} \quad i \leq \beta/2 \\ 1 - g_i & \text{if} \quad i > \beta/2 \end{cases} \tag{5}$$

where P then ranges from -1 for sources panned hard left, to 1 for sources panned hard right, with 0 indicating a position in the centre. Then, g ranges from 0 for sources hard left, increasing to 1 for centre positioned sources, before decreasing to 0 for sources panned hard right.

The frequency-azimuth plane can then be defined by:

$$Az_{k,i} = \begin{cases} |Lf_k - g_i Rf_k| & \text{if} \quad i \leq \beta/2 \\ |Rf_k - g_i Lf_k| & \text{if} \quad i > \beta/2 \end{cases} \tag{6}$$

where Rf_k and Lf_k denote the kth frequency bin of the current right and left frames of the STFT respectively.

To allow extraction or resynthesis of a given source, a source position d is given by the user, which is a value taken from P. For a source occurring at this position, the energy in the frequency bins associated with the source will cancel out, resulting in energy minima at those frequency bins associated with the source. The residual energy associated with these minima then contains energy present due to other sources in the mixture. Due to frequency overlap between different sources in the mixture, the position of a given frequency minimum can move away from that of the actual source position, depending on the relative strength of the sources active in the frequency bin. To overcome this effect, an azimuth subspace width, H is defined, so that H spans a subset of the possible values of P. Taken with d, this defines which positions in P are to be used for resynthesis. In the original version of Adress, the source spectrogram for the current frame are estimated from

$$Y_k = \begin{cases} E & \text{if} \quad d - H/2 \leq \text{argmin}(Az_{ki}) \leq d + H/2 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where E is defined as:

$$E = \begin{cases} Lf_k - \min(Az_{ki}) & \text{if} \quad d \leq 0 \\ Rf_k - \min(Az_{ki}) & \text{if} \quad d > 0 \end{cases} \tag{8}$$

The phase information from the channel in which the source is dominant can then be applied to this spectrogram to allow resynthesis in the time domain via an inverse STFT once all the frames have been estimated. However, for the purposes of this paper, resynthesis is carried out by defining a binary mask for the source:

$$M_k = \begin{cases} 1 & \text{if} \quad d - H/2 \leq \text{argmin}(Az_{ki}) \leq d + H/2 \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

This binary mask is then multiplied with the associated left and right complex spectrogram frames to estimate complex source spectrograms for each channel. The separated stereo signal is then recovered in the time domain via inverse STFTs on each of the separated complex source spectrograms. In the case of vocal extraction, we assume in this paper that the vocal position is in the centre, i.e. with a source position d of 0, and an azimuth width of 0.8. Therefore, extracting time-frequency bins associated with this region is assumed to extract the vocals from the input stereo signal. However, if the vocals are not in the center position, the choice of source position can be modified by the user. Similarly, the energy associated with the other sources outside the given associated azimuth can be recovered from the residual mask Q_k, defined as $1 - M_k$.

3. VOCAL SEPARATION USING NEAREST NEIGHBOURS MEDIAN FILTERING

The vocal separation technique proposed assumes 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 is also assumed that the vocal is 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.

A magnitude spectrogram is then calculated via the Short-Time Fourier Transform, and then the distance between all frames is calculated to identify repeating parts in the spectrogram. 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 used is the squared Euclidean distance between the frames, given by:

$$D_{k,l} = \sum_{r=1}^{n} (\mathbf{X}_k - \mathbf{X}_l)^2 \tag{10}$$

where $\mathbf{X}$ is the mixture magnitude spectrogram of size $n \times m$, with n the number of frequency bins and m 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.

The resulting distance matrix is a symmetric matrix $\mathbf{D}$ of size $m \times m$. Each row of $\mathbf{D}$ is sorted in ascending order, and the frame indices obtained of the p nearest neighbours to the current (kth) frame. These nearest neighbours are then extracted from the magnitude spectrogram and stored in a $n \times p$ matrix $\mathbf{P}$. An estimate of the background music for the kth frame is then obtained from:

$$\mathbf{Y}_k = \mathcal{M}(\mathbf{P}) \tag{11}$$

where $\mathbf{Y}_k$ is the kth frame of the estimated background music spectrogram $\mathbf{Y}$, and where $\mathcal{M}$ denotes the median operator. Median filtering is used as the bins with vocal energy will be outliers amongst the different repetitions, and median filtering has been shown to be good at removing outliers [10].

We further assume that the background music cannot have a greater energy at a given time-frequency bin than that of the mixture signal and so values in $\mathbf{Y}$ which are greater than those of the original mixture are replaced with the original energy at those bins:

$$\mathbf{Y}_{f,k} = \min(\mathbf{X}_{f,k}, \mathbf{Y}_{f,k}) \tag{12}$$

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

A mask is then generated based on a Gaussian radial basis function approach, as described in the vocal separation algorithm proposed in [11]:

$$\mathbf{W}_{f,k} = \exp\left(-\frac{(\log \mathbf{X}_{f,k} - \log \mathbf{Y}_{f,k})^2}{2\lambda^2}\right) \tag{13}$$

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:

$$\mathbf{B} = \mathbf{W} \otimes \mathbf{R} \tag{14}$$

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 be estimated from:

$$\mathbf{V} = (1 - \mathbf{W}) \otimes \mathbf{R} \tag{15}$$

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 a cutoff frequency from the vocal signal and to add these frequencies back into the background track as was done in [12]. This method was shown to give good results in vocal separation from single channel mixtures.

4. VOCAL SEPARATION USING ADRESS AND NEAREST NEIGHBOURS MEDIAN FILTERING

The underlying idea of the stereo vocal separation algorithm proposed here is to combine the two previously described techniques to give improved vocal separation by taking advantage of the strengths of both techniques, while using the different nature of the techniques to overcome their individual weaknesses. For example, the vocal extracted using Nearest Neighbours Median Filtering will often contain traces of other instruments. If these come from a pan position away from the vocal position then Adress can be used to remove these parts without compromising the vocal separation.

The algorithm proceeds by performing an STFT on each of the channels individually. These are then passed through the Adress algorithm to obtain $\mathbf{M}$, which gathers the frame masks M_k into a full spectrogram mask which identifies which spectrogram bins belong to the vocal region. Next, the nearest-neighbours median filtering algorithm is then performed on the magnitude spectrograms of each channel individually to yield $\mathbf{V}_L$ and $\mathbf{V}_R$, which denotes the separated complex vocal spectrograms for the left and right channels respectively. Improved estimates of the vocal in each channel are then obtained from:

$$\hat{\mathbf{V}}_L = \mathbf{V}_L \otimes \mathbf{M} \tag{16}$$

$$\hat{\mathbf{V}}_R = \mathbf{V}_R \otimes \mathbf{M} \tag{17}$$

This assumes that any bins which fall outside the region of the vocal position in stereo space belong to the backing track and should be removed. Again, all frequencies below a chosen cutoff frequency are removed from the vocal signal and added to the backing track signal. This can be done by setting all bins in $\mathbf{M}$ below the cutoff frequency to zero. The time domain vocal signal is then obtained via inverse STFT.

Improved estimates of the backing track can then be obtained from:

$$\hat{\mathbf{B}}_L = \mathbf{B}_L + \mathbf{V}_L \otimes \mathbf{Q} \tag{18}$$

$$\hat{\mathbf{B}}_R = \mathbf{B}_R + \mathbf{V}_R \otimes \mathbf{Q} \tag{19}$$

where $\mathbf{B}_L$ and $\mathbf{B}_R$ denote the complex backing track left and right spectrograms respectively, with $\mathbf{Q} = 1 - \mathbf{M}$, except for bins in $\mathbf{Q}$ below the chosen cutoff frequency, which are set to 1. Again the time domain backing track is recovered using the inverse STFT.

A version of the algorithm was also implemented where Adress was performed and then the Nearest-Neighbour Median Filtering algorithm performed on the extracted vocal region, which still contained drums and bass. In this case it was observed that more of the vocal was extracted with the repeating drum and bass backing track, giving poorer results for vocal separation.

4.1. Difference Nearest Neighbours Median Filtering

A problem with the Nearest Neighbours Median Filtering algorithm is that traces of the vocals still show up in the separated backing track. These traces will not be reduced by the use of Adress as previously described, and so it was decided to investigate alternate means of trying to separate the vocals. To this end, it was decided to investigate calculating the Nearest Neighbours using a different method, namely calculating the distance matrix based on the difference between successive frames of the magnitude spectrogram. This will have the effect of reducing the effect of the pitched instruments on the distance measure as they will often be relatively stationary between successive overlapped frames.

The kth frame of the difference matrix $\mathbf{T}$ is then given as:

$$T_k = \mathbf{X}_k - \mathbf{X}_{k-1} \tag{20}$$

When k is 1, it is assumed that $\mathbf{X}_0$ is a vector of all zeros. The distance matrix is then calculated as per eqn. (10), with $\mathbf{T}$ substituted for $\mathbf{X}$, and the algorithm then proceeds as before.

It was found in this case that the resulting repeating backing track consisted predominantly of snare and kick drum and parts

of the pitched instruments with only small traces of the vocals. The downside of this is that most of the pitched instruments now come out with the vocal. Nevertheless, it should be noted that the repetitive backing track contains the drum sources which typically occur in the same point in stereo space as the lead vocal in modern recordings, and that there is very little vocal in this separated signal. Further, carrying out subsequent processing as described above using the mask generated from Adress then removes the pitched instruments which are not positioned in the center, allowing recovery of the separated lead vocal and backing track.

4.2. Example separations

Figure 1 shows the spectrogram of the right channel of an excerpt from "Knowing me, Knowing you" by Abba. Figures 2 and 3 respectively show the separated vocals and track obtained via the Adress algorithm. It can be seen that there is still considerable drum energy present in the separated vocals and that the drum energy is reduced in the separated track. This is also noticable on listening. Figures 4 and 5 show the vocal and track separations obtained using the Nearest Neighbour algorithm. Significantly more pitched instrument energy is visible in the separated vocal, while it can be seen that some of the drums have been reduced in volume in comparison to Adress, while on listening, due to the Nearest Neighbours method, the track is significantly distorted and there is a greater presence of vocals than with Adress.

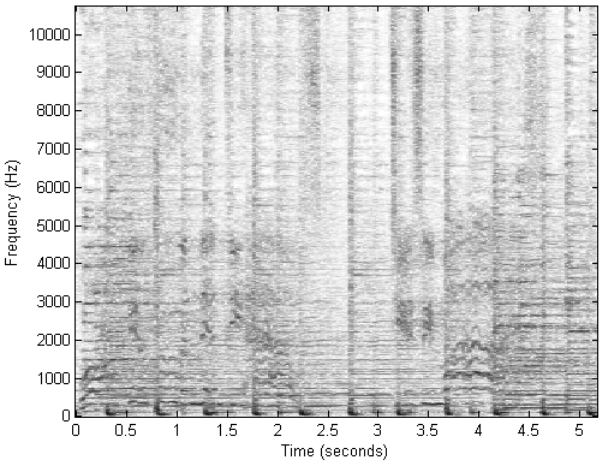

Figure 1: *Spectrogram of right channel (Knowing me, knowing you, by Abba)*

Figures 6 and 7 then show the separated vocals and track using the combined Adress and Nearest Neighbours Median Filtering approach proposed in this paper. It can be seen that the vocal is clearer, with less interference from both drums and pitched instruments. This is borne out on listening to the example. The separated track now contains more drum energy than the original Adress separation, but at the expense of an audible increase in vocal levels in comparison to Adress on its own.

Figures 8 and 9 then show the separations obtained using the difference Nearest Neighbours Adress approach. The vocal is still well separated, but there is more noise visible than with the combined Adress and Nearest Neighbours Median Filtering approach.

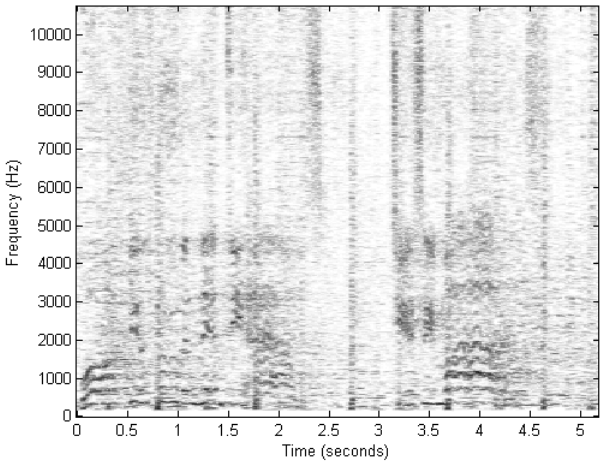

Figure 2: *Spectrogram of separated vocals (Adress)*

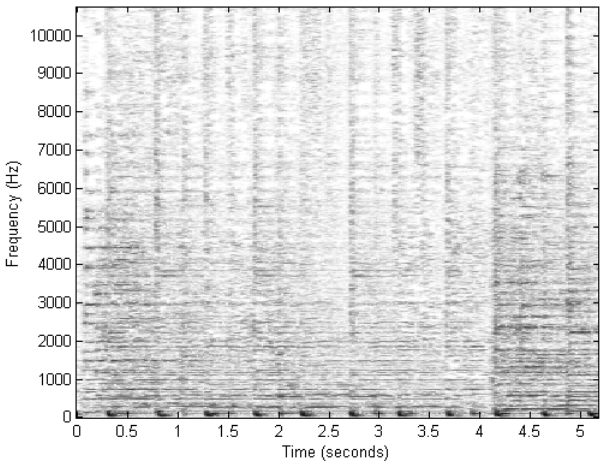

Figure 3: *Spectrogram of separated track (Adress)*

This extra noise is evident on listening. However, the separated track has vocal levels comparable to that achieved with Adress, while still capturing a lot of energy from the drums, making it the best sounding track separation. This suggests that what yields the best vocal separation will not always result in an optimal separation of the backing track, and that the separation method should be chosen based on what is more important to the user, with the difference Nearest Neighbours preferable for backing track separation. This is further borne out in informal listening tests. Example separations for this track are available for listening at [13].

5. PERFORMANCE EVALUATION

The proposed algorithms (both Nearest Neighbours and difference Nearest Neighbours) were evaluated using the development signals used in Sisec 2011 for the Professionally produced music recordings separation task [14]. Also tested as baselines were the original Adress algorithm and the original Nearest Neighbours

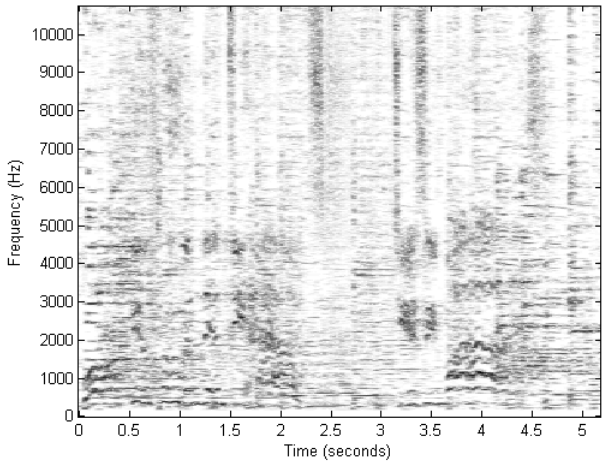

Figure 4: *Spectrogram of separated vocals (Nearest Neighbours)*

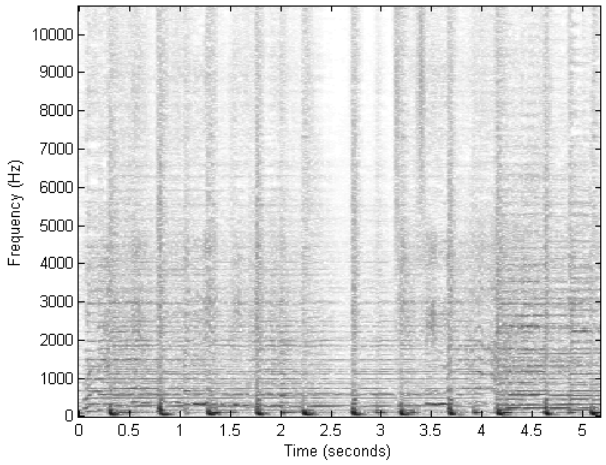

Figure 5: *Spectrogram of separated track (Nearest Neighbours)*

Median Filtering algorithm. This allows the performance of the proposed algorithms to be compared with other algorithms which were evaluated on the development set. The set consists of 5 snippets from songs recorded in a variety of styles, all at a samplerate of 44.1 kHz. Vocal extraction performance was evaluated using the PEASS toolbox, which calculates a set of objective measures for the perceptual evaluation of audio source separation [15]. The metrics used were the overall perceptual score (OPS), which attempts to measure the perceived overall quality of the separation, the target-related perceptual score (TPS) a measure of how the separated source matches the spatial positioning of the original source, the artifacts-related perceptual score (APS) which relates to the perceived amount of artefacts present in the separated source, and the interference-related perceptual score (IPS) which determines how much interference due to other sources is perceived in the separated source. Also computed were the Signal to Distortion Ratio (SDR), Signal to Interference Ratio (SIR), Signal to Artefacts Ratio, and the source Image to Spatial Distortion

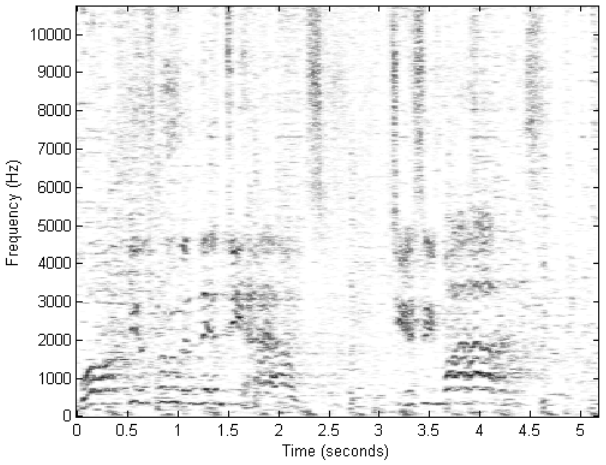

Figure 6: *Spectrogram of separated vocals (Nearest Neighbours, Adress)*

ratio(ISR). The values of these metrics are on a scale from 0-100, with 100 indicating perfect separation performance with respect to the metric in question.

The average results for the separated vocals obtained from the 5 snippets are shown in Table 1 for the Adress algorithm, denoted by Ad, the original nearest-neighbours median filtering separation algorithm, denoted by NNM, the Nearest Neighbours/Adress algorithm proposed in this paper,denoted by NNAd, and the difference Nearest Neighbours Adress algorithm, also proposed here which is denoted by dNNAd. The parameters used for the Adress algorithm both standalone and in the proposed algorithms were a source position of 0 and an azimuth width of 0.8. This assumes that the vocals are positioned in the centre, but for situations where this is not the case, the user can choose a different source position and azimuth width for the vocals. For the Nearest neighbour Median filtering algorithm, 100 nearest neighbours were used, again in both the standalone version and the proposed Nearest Neighbours Adress algorithm. The same number of nearest neighbours was used for the difference Nearest Neighbours median filtering version of the proposed algorithm. All separated vocal sources were high pass filtered at 130 Hz, and any energy below this frequency was reallocated to the separated instrumental track. It should be pointed out that knowledge of the vocal melody would allow further optimisation of this cutoff, which could then be adjusted on a frame by frame basis to further improve the separations. In all cases an FFT size of 4096 samples, a Hann window of 4096 samples and a hopsize of 1024 samples were used when calculating the STFT. Total processing time for the 5 excerpts was under 5 minutes in unoptimised Matlab code on a Core 2 Q9550 processor at 2.83 GHz. This run time compares favourably to the majority of the techniques evaluated in Sisec 2011, which are available on the Sisec 2011 website [14].

The results for the individual tracks are contained in tables 2 to 6, to allow comparison with the individual results available on the Sisec 2011 website [14]. Similarly audio for the separated vocals and backing tracks can be found at [13], to again allow the reader to compare the results with the separations available on the Sisec 2011 website. Of particular interest are the results

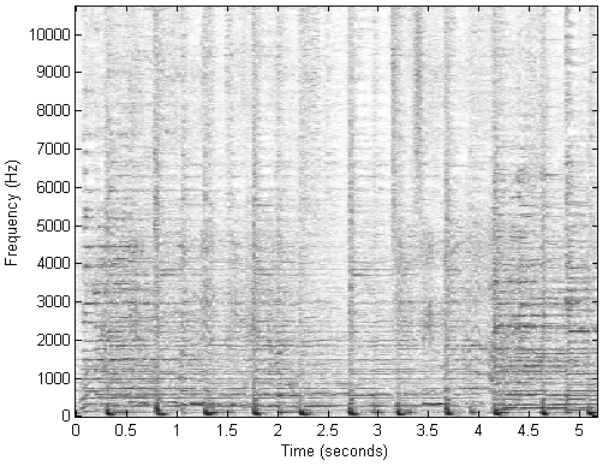

Figure 7: *Spectrogram of separated track (Nearest Neighbours, Adress)*

	NNM	Ad	NNad	dNNAd
OPS	22.5	21.0	34.2	33.9
TPS	37.8	61.5	44.0	46.1
APS	42.0	71.3	41.8	45.8
IPS	26.9	27.8	49.4	47.3
SDR	3.2	-2.3	3.3	4.1
SAR	18.8	18.1	16.8	17.2
SIR	5.4	1.7	6.8	5.3
ISR	5.9	6.7	5.3	9.37

Table 1: *Overall Performance Evaluation. Ad denotes the Adress algorithm, NNM denotes the Nearest-Neighbour Median Filtering vocal separation algorithm, while NNad and dNNAd respectively denote the Nearest Neighbour and difference Nearest Neighbour Median Filtering Adress algorithms proposed in this paper.*

obtained for “Ultimate NZ tour", which are considerably lower than those of the other tracks. The reason for this is due to the presence of a synthesiser in the same region in stereo space as the vocal, which serves to highlight a shortcoming of the proposed approaches, namely that it assumes that the only instruments in the vocal region are typically snare, kick drum and bass guitar, which are all removed to some degree by the proposed approach, while other pitched instruments in that region will not be separated. This is a potential area for future work.

The results show that the performance of the algorithms is comparable to the best of the algorithms entered in Sisec 2011, with the average overall perceptual score for both versions of the proposed algorithm being higher than the majority of the algorithms tested on the development set. The Nearest Neighbours version of the algorithm has a slightly higher OPS than the difference Nearest Neighbours algorithm, but informal listening tests suggest that the difference Nearest Neighbours algorithm gives better sounding results when removing the vocals to create instrumental backing tracks, such as required for karaoke applications. It should be noted that evaluation of the backing track separation performance was not possible as only the individual tracks are given,

	NNM	Ad	NNad	dNNAd
OPS	18.1	31.9	34.2	31.9
TPS	35.8	57.3	40.0	40.3
APS	44.3	58.0	39.5	43.2
IPS	19.7	40.6	55.2	58.5
SDR	4.7	-1.39	4.2	0.9
SAR	19.4	16.7	16.6	14.9
SIR	6.0	2.2	5.9	5.2
ISR	7.0	6.5	6.0	6.1

Table 2: *Performance Evaluation for “The ones we love" by Another Dreamer, Legend as per Table 1*

	NNM	Ad	NNad	dNNAd
OPS	25.4	31.6	39.3	40.2
TPS	35.7	61.2	46.0	55.2
APS	25.2	64.7	36.3	48.0
IPS	34.4	42.2	56.8	57.3
SDR	3.3	1.1	3.4	3.8
SAR	22.2	18.5	22.2	19.6
SIR	12.1	6.5	17.1	11.1
ISR	3.6	6.2	3.6	6.1

Table 3: *Performance Evaluation for “Que Pena / Tanto Faz" by Tamy, Legend as per Table 1*

	NNM	Ad	NNad	dNNAd
OPS	20.5	18.1	32.6	32.3
TPS	35.7	59.7	40.7	32.9
APS	35.5	71.8	39.9	39.5
IPS	24.3	21.3	45.5	53.1
SDR	3.4	0.0	2.9	3.8
SAR	18.9	18.2	16.6	18.3
SIR	5.1	4.1	3.8	9.8
ISR	4.9	7.2	4.6	6.8

Table 4: *Performance Evaluation for “Roads" by Bearlin, Legend as per Table 1*

	NNM	Ad	NNad	dNNAd
OPS	33.8	12.2	40.5	38.0
TPS	46.5	68.5	52.8	51.9
APS	49.1	82.9	48.1	47.7
IPS	44.2	26.1	59.6	61.8
SDR	3.5	-4.2	3.69	1.4
SAR	15.0	18.5	12.3	13.4
SIR	4.2	-0.5	6.0	5.1
ISR	5.4	6.8	5.1	8.3

Table 5: *Performance Evaluation for “Remember the Name" by Fort Minor, Legend as per Table 1*

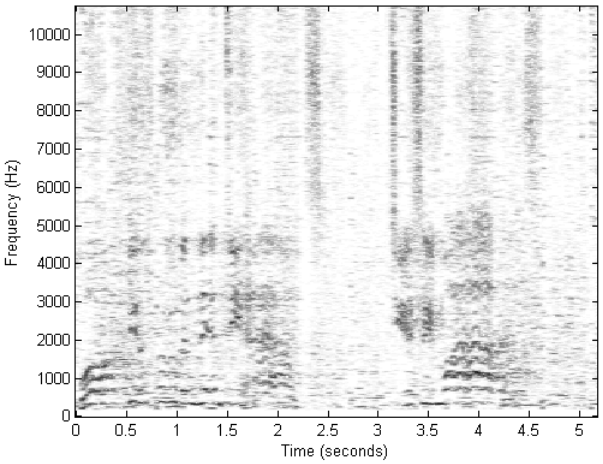

Figure 8: *Spectrogram of separated vocals (difference Nearest Neighbours, Adress)*

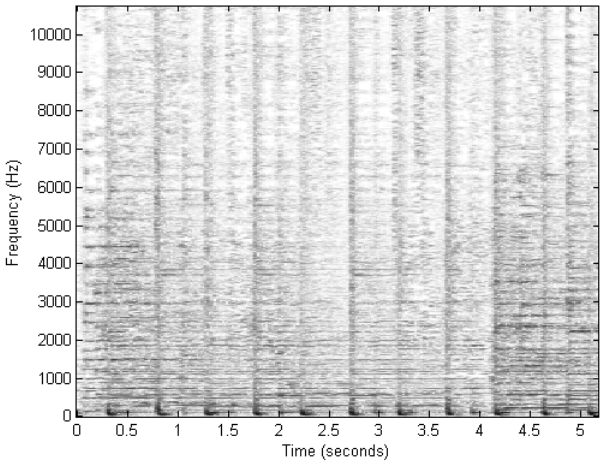

Figure 9: *Spectrogram of separated vocals (difference Nearest Neighbours, Adress)*

	NNM	Ad	NNad	dNNAd
OPS	14.5	11.3	24.4	23.5
TPS	35.5	60.7	40.6	41.2
APS	55.7	79.2	45.2	52.1
IPS	11.7	9.0	29.7	28.6
SDR	1.2	-7.0	2.2	-4.1
SAR	18.6	18.3	16.4	16.7
SIR	-0.4	-3.9	1.2	-1.6
ISR	8.6	6.7	7.3	7.2

Table 6: *Performance Evaluation for "Ultimate NZ Tour", Legend as per Table 1*

and so the actual mixes of the backing tracks were unavailable. Future work will be done to quantify this informal result. It can also be seen that the proposed algorithms significantly outperform the results obtained when the Adress algorithm and the Nearest Neighbours Median Filtering algorithm are used individually. This shows that the combination of the two algorithms helps overcome the weaknesses inherent in the original techniques, while emphasising the strengths of the individual algorithms.

The above results demonstrate that the proposed algorithms give good vocal extraction results from stereo signals, comparable with state of the art algorithms, but at a relatively low computational cost. It further demonstrates the utility of combining different types of separation techniques to help overcome the respective weaknesses of the individual methods.

6. CONCLUSIONS

In this paper, a new algorithm for extracting vocals from stereo recordings was proposed. The algorithm combines two existing approaches used for vocal separation, the Adress algorithm, and the Nearest Neighbours Median filtering vocal separation algorithm to give improved vocal separation results. The Nearest Neighbours Median Filtering algorithm is performed on each channel individually, and then a mask generated by the Adress algorithm is used to remove bins which do not belong to the vocal region of the stereo space. A variant of this algorithm, where the nearest neighbours were calculated using a first-order difference spectrogram was also investigated.

The effectiveness of these algorithms were evaluated on a test-set of real world recordings and were found to yield separations comparable to the state of the art, but at a relatively low computational cost. Future work will concentrate on evaluation of the performance of the algorithms for generating instrumental backing tracks, as well as the incorporation of other source separation techniques to further improve the results obtained, such as improving the removal of the bass guitar from the separated vocals, and developing techniques to remove any other pitched instruments which occupy the same point in stereo space as the vocals.

7. REFERENCES

[1] Y. Li and D. Wang, "Separation of singing voice from music accompaniment for monaural recordings," *IEEE Transactions on Audio Speech and Language Processing*, 2006.

[2] 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.

[3] S. Vembu and S. Baumann, "Separation of vocals from polyphonic audio recordings," in *Proc. International Symposium on Music Information Retrieval (ISMIR05)*, 2005.

[4] B. Raj, P. Smaragdis, M. Shashanka, and R. Singh, "Separating a foreground singer from background music," in *Proc. Int Symp. Frontiers Research Speech Music (FRSM)*, India, 2007.

[5] J.L. Durrieu, B. David, and G. Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation," *IEEE Journal of Selected Topics on Signal Processing*, October 2011.

[6] D. FitzGerald, "Vocal separation using nearest neighbours and median filtering," in *23nd IET Irish Signals and Systems Conference*, 2012.

[7] Z. Rafii and B. Pardo, "Music/voice separation using the similarity matrix," in *13th International Society for Music Information Retrieval*, 2012.

[8] D. Barry, B. Lawlor, and E. Coyle, "Sound source separation: Azimuth discrimination and resynthesis," in *Proc.of 7th Int. Conference on Digital Audio Effects, (DAFX 04)*, 2004.

[9] S. Sofianos, A. Ariyaeeinia, and R. Polfreman, "Singing voice separation based on non-vocal independent component subtraction and amplitude discrimination," in *Proc.of 13th Int. Conference on Digital Audio Effects, (DAFX10)*, 2010.

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

[11] A. Liutkus, Z. Rafii, R. Badeau, B. Pardo, and G. Richard, "Adaptive filtering for music/voice separation exploiting the repeating musical structure," 2012.

[12] D. FitzGerald and M. Gainza, "Single channel vocal separation using median filtering and factorisation techniques," *ISAST Transactions on Electronic and Signal Processing*, vol. No. 1 Vol. 4, pp. 62–73, 2010.

[13] Audio Examples, "Stereo vocal extraction using adress and nearest neighbours median filtering," Webpage, `http://eleceng.dit.ie/derryfitzgerald/index.php?uid=489&menu_id=72`.

[14] Sisec 2011, "2011 signal separation evaluation campaign, professionally produced music recordings," Webpage, `http://www.irisa.fr/metiss/SiSEC11/professional/dev_eval2011.htm/`.

[15] 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*, 2011.