2010-01-01

Single Channel Vocal Separation Using Median Filtering and Factorisation Techniques

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

Mikel Gainza
Technological University Dublin, Mikel.Gainza@tudublin.ie

Recommended Citation

FitzGerald, D. & Gainza, M. (2010) Single Channel Vocal Separation using Median Filtering and Factorisation Techniques, *ISAST Transactions on Electronic and Signal Processing*, No. 1, Vol. 4,2010 (ISSN 1797-2329), pages: 62 - 73, 2010.

Single Channel Vocal Separation using Median Filtering and Factorisation Techniques

Derry FitzGerald, Mikel Gainza, Audio Research Group, Dublin Institute of Technology, Kevin St, Dublin 2, Ireland

***Abstract*— This paper deals with the problem of the extraction of vocals from single channel audio signals containing both vocals and other instruments, including both pitched instruments and percussion instruments. A novel median filtering-based approach for the extraction of vocal tracks is described, which is simple and efficient to implement. Further improvements in separation quality are then obtained by the application of tensor factorisation techniques to further extract residual instruments from the vocal mix. Finally, a novel use of non-negative partial matrix cofactorisation is demonstrated as a means of further improving separation quality. Here the original single channel mixture is partially cofactorised in conjunction with the separated vocal signal in order to obtain improved separation of the vocal and instrumental tracks. The effectiveness of these techniques is then demonstrated on a test set of real world signals.**

***Index Terms*— Single channel sound source separation, vocal separation and suppression, Non-negative partial cofactorisation, tensor factorisation.**

I. Introduction

The topic of singing voice (or vocal) separation/extraction, a subset of the more general sound source separation problem, has recieved attention over the past number of years. Here, in this case, the separation problem is limited to extracting the singing voice from a recording of polyphonic music, with no restrictions on the instrumentation present.

Vocal separation is a topic of interest due to its numerous applications. For example, once the vocals have been extracted, the vocal melody line can be more easily transcribed by pitch estimation algorithms, the output of which can then be used in query by humming systems. The separated vocals can also be repurposed or "sampled" for use in other pieces of music. This is commonplace in popular music, and the availabilty of high quality vocal separations would greatly increase the amount of material available for this purpose.

Further, much existing research on sound source separation has focused on pitched instruments and/or percussion instruments, and the addition of vocal separation algorithms in conjunction with these existing approaches would allow other applications, such as the upmixing of old recordings from mono to stereo or 5.1 surround sound. Other applications include automatically aligning lyrics to music and singer identification.

A. Previous Research

Much of the existing research on vocal separation has focused on stereo or two channel recordings, where the position of the vocal in the stereo field is often used to aid separation. Work in this area includes the system proposed by Sofianos et al [1], which makes use of both Independent Component Analysis [2] and the Azimuth Discrimination and Resynthesis algorithm (ADRess) [3] to extract vocals from stereo signals. A variant of ADRess has also been used commercially for vocal removal for kareoke games. However, a problem with such approaches is that there are often multiple instruments occupying the same position in the stereo field as the vocals, such as bass guitar and drums such as the kick and snare drums. Further, a large proportion of older recordings from the 1950s and before are only single channel recordings. Therefore, it can be seen that a system that is capable of separating vocals from single channel or mono recordings would be advantageous, both to handle old recordings, and deal with the source overlap problem in modern recordings. To this end, the work in this paper focuses on the problem of vocal extraction from single channel recordings of polyphonic music, and so a brief overview of previous research on this topic is now presented.

Li and Wang [4] presented a system which consisted of three stages, the first divided the input signal into regions where vocals were present and regions where they were not. The regions with vocals were then passed to a predominant pitch estimator, which attempted to identify the vocal melody, Then knowledge of this melody was used to separate out the singing voice from the recordings using an adaptation of a previous technique for single channel speech separation [5]. However, as it is based on a predominant melody estimator, it cannot deal with vocal harmonies.

Ozerov et al proposed the use of Bayesian methods for the purposes of single channel vocal separation [6]. Their system required the use of training data consisting of a set of solo vocal recordings, which were then used to train a Bayesian model for singing voice. Similarly, a set of instrumental tracks were then used to create a model for instrumental parts found in music. Their algorithm consisted of a number of stages. First the input signal was segmented into regions where the vocal was present, and regions where only instruments were present. The instrument-only segments were then used to adapt the general instrumental model to better fit the actual instruments present in the input signal. This adapted model, in conjunction with the existing general vocal model, was then used to attempt separation of the vocals. Both of these were further adapted during the course of the separation to better match the characteristics of the input signal, Finally, separation was obtained by using the adapted models to create adaptive Wiener filters which were then applied to the input signal. This

Manuscript completed in Oct, 2010. The work reported in this paper is supported by Science Foundation Ireland.

system was found to be capable of giving good separation results, but did have a number of shortcomings. Firstly, the input signal must have sufficient non-vocal segments to allow the instrumental model to adapt to the characteristics of the input signal. Secondly, the music from the non-vocal parts had to be similar to that during the vocal parts, and finally, the system was designed to deal with solo singing voice, in other words, the system performed better in the absence of backing vocals.

Vembu et al [7] proposed a singing voice separation system based on non-negative matrix factorisation (NMF) [22]. The first stage of their technique was to build a classifier to discriminate automatically between sections of the music where no vocals were present and segments where they were present. They then proceeded to decompose a spectrogram of the input signal using NMF, and then to cluster the basis functions into vocal and nonvocal basis functions. Two techniques for clustering were tested, the first used the vocal/nonvocal classifier to discriminate between vocal and non-vocal basis functions, while the second was an unsupervised classifier based on features known to discriminate between vocal and non-vocal segments of music. The first method was not found to perform well, while the second was capable of giving good results in simple music, such as just voice and guitar with no other instruments present.

Raj et al also proposed a factorisation-like technique for separating singing voice [8]. Here, they manually identified regions which did not contain vocals and used these segments to train a model of the accompaniment. The vocal parts were then learned from the mixture while keeping the accompaniment model fixed. This suffered from a similar drawback to the method proposed by Ozerov et al, namely, the mixure signal needed to have sufficient non-vocal segments to accurately train the model.

Hsu et al extended work by Li et al to improve the separation of singing voice [10]. They used Hidden Markov Models to identify regions where the accompaniment was dominant, where voiced singing (ie where a discernible pitch was present in the vocals) was dominant, and where unvoiced singing was dominant. They then used the method proposed by Li et al to identify and separate the voiced parts of the singing. Further enhancements included the use of a spectral subtraction method to reduce the level of accompaniment in the separated singing, as well as using statistical models to attempt to separate the unvoiced regions of singing. However, it still suffered from the disadvantage that the system is based on predominant melody estimation.

Hsu et al. later proposed another single-channel singing voice separation algorithm in [9]. This algorithm consists of a number of steps. First, sinusiodal partials are extracted from the signal. Then parameters measuring vibrato and tremelo are estimated for each of the partials. Then the vocal partials are discriminated from the instrumental partials by thresholding on the extracted parameters. A technique called Normalised sub-harmonic summation [10] was then applied as a means of further enhancing the vocal harmonics, and improve the separations. This principal application of this paper was melody estimation, but the technique appears to give good separation results. However, a drawback of this approach is that it focuses on the extraction of a solo singing voice, without attempting to extract backing vocals.

As can be seen from the above, the principal problems with existing vocal separation algorithms is that they depend on either previous training data, or training on non-vocal segments of the music, or a predominant melody estimation stage, which can introduce problems if the incorrect pitch is determined. Further, it can be seen that these algorithms are all designed to deal with a single solo voice, as opposed to handling backing vocals and other vocal harmony parts. These are shortcomings which are addressed in the algorithms proposed in this paper.

Other research of interest, though on source separaration of drum sounds in particular, includes the Harmonic-Percussive Median Filtering-based algorithm developed by the authors of this paper [11]. This uses median filtering on spectrograms to separate harmonic and percussive components in audio signals. As will be seen later, the properties of this algorithm can be modified for use to separate vocals, and so section II describes this algorithm in greater detail. A further technique, also proposed by the authors, uses tensor factorisations to separate musical sources, both pitched and percussion [17]. This can be used as a means of improving vocal separations by reducing artifacts remaining after initial vocal separation. Finally, a partial cofactorisation approach to drum separation was described in [21]. This approach uses pre-existing samples of drums which are cofactorised in conjunction with the mixture spectrogram in order to constrain some of the basis functions to correspond to drum instruments. However, there is no restriction on the source to be used in the cofactorisation, and so the vocal separations obtained from one method can be used to drive the cofactorisation, resulting in a further improved separation of the vocals.

B. Paper Overview

In the following section, a simple median filtering-based approach to harmonic-percussive separation is discussed, with particular reference to it's properties when dealing with singing voice. This leads to Section III where a multipass extension of this algorithm is proposed for the purposes of vocal separation. Section IV describes how the vocal separation output from the multipass median filtering algorithm can be enhanced by the addition of another separation algorithm based on non-negative tensor factorisation. The following section then proposes a novel use for non-negative partial cofactorisation. Here, the output of the previous vocal separation algorithm is used as a guide to perform a new factorisation of the original mixture into vocal and instrumental tracks, resulting in improved separation over previous approaches. Section VI then describes the test set and the testing procedures used, as well as detailing the performance of the algorithms. Finally Section VII offers some conclusions on the methods proposed and highlights areas for future research.

II. Harmonic-Percussive Separation using Median Filtering

Recently, a median filtering-based technique for the separation of harmonic and percussive events from single channel

audio signals has been proposed by the author [11]. This is based on the idea that broadband impulsive noises such as drums and percussion form stable vertical ridges in a magnitude or power spectrogram,typically obtained from a Short-Time Fourier Transform (STFT), while harmonics from pitched instruments form stable horizontal ridges in a magnitude or power spectrogram, This is illustrated in Figure 1, which shows a spectrogram of an audio signal containing a snare drum and a piano. It can be seen that the harmonics of the piano form stable horizontal lines in the spectrogram, while the snare drum forms a vertical line in the spectrogram. Therefore, a technique that emphasises vertical lines while suppressing horizontal lines should result in a spectrogram that contains mainly percussion instruments. Similarly emphasising horizontal lines at the expense of the vertical lines should result in a spectrogram which contains mainly pitched instruments.

This principal was first used for the separation of harmonic and percussion instruments by Ono et al [12], who used an iterative diffusion-based approach to emphasise horizontal lines and vertical lines in spectrograms respectively. In effect, this process smoothed out vertical lines in the spectrogram by reducing spikes associated with harmonics, and smoothed horizontal lines by reducing spikes associated with transients due to drums or percussion. Another way of looking at this problem is to regard the spikes due to harmonics within a given time frame as outliers, and to regard spikes due to percussion onsets as outliers across a given frequency slice in time. Therefore, the problem of separating percussive and pitched events reduces to the identification and removal of outliers from each individual frame for percussive events, and for each frequency slice to recover pitched instruments.

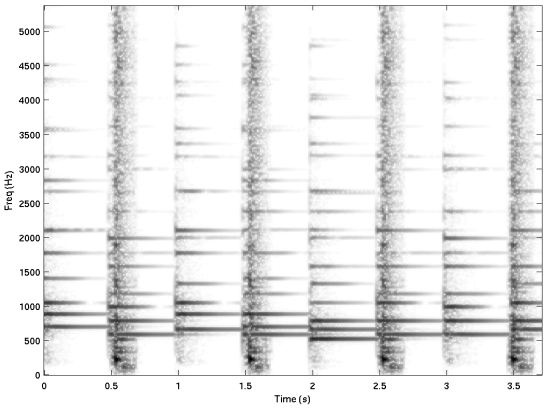

Fig. 1. *Spectrogram of a drum and piano. The drum can be seen to form stable vertical ridges, while the harmonics of the piano form stable horizontal ridges in the spectrogram*

To this end, it was proposed in [11] to use median filters to remove these outliers, as median filters have been widely used in image processing for the removal of speckle noise and salt and pepper noise from images. These forms of noise can also be regarded as outliers in an image [13]. Median filters have proved better than moving average filters in removing impulse noise because they are not dependent on values which are outliers from the typical values in the area surrounding the original sample. Median filters filter a signal by replacing a given sample 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:

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

Where l is even, the mean of the two values at the center of the sorted list is used.

The effectiveness of median filtering in the suppression of harmonic and percussive events in audio spectrograms is demonstrated in the following figures. Figure 2(a) shows the energy in a frequency bin across time (henceforth refered to as a frequency slice) from the same mixture of instruments, namely piano and snare drum, as shown in figure 1. The onset of the snare drum can be seen as a large jump in energy in the frequency slice, while the energy due to the piano note harmonic is more constant across the slice. In comparison, figure 2(b) shows the energy in the frequency slice subsequent to median filtering. It can be seen that most of the energy due to the drum onset has been eliminated by median filtering, resulting in the suppression of the drum sound. Denoting the input magnitude spectrogram $\mathbf{S}$, the ith time frame as S_i, and the hth frequency slice as S_h, then a harmonic-enhanced spectrogram frame H_h can be obtained from:

$$H_h = \mathcal{M}\{S_h, l_{harm}\} \quad (2)$$

These individual frequency slices can then be combined to yield $\mathbf{H}$.

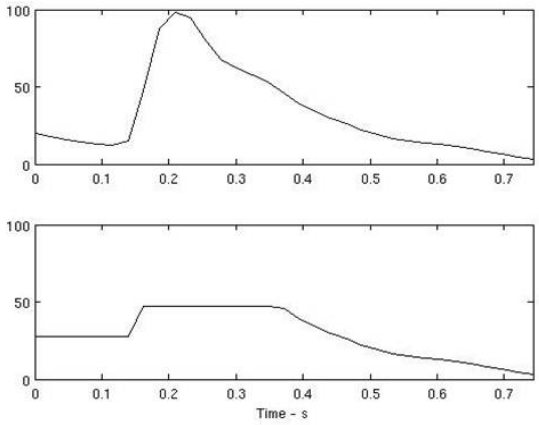

Fig. 2. *Spectrogram frequency slice from a spectrogram containing a mixture of snare drum and piano a) The original slice, b) the slice after median filtering. It can be seen that a large amount of the energy of the snare has been removed*

Figure 3(a) then shows a spectrogram frame from the same mixture as above. The harmonics due to the presence of the piano are evident as large spikes in energy in the frame. Figure 3(b) then shows the same frame after median filtering. The harmonics have been removed by the median filtering, leaving a frame in which percussive energy predominates. A

percussion-enhanced spectrogram frame P_i can be generated by performing median filtering on S_i:

$$P_i = \mathcal{M}\{S_i, l_{perc}\} \tag{3}$$

where $\mathcal{M}$ indicates the median filtering operation, and l_{perc} is the length of the percussion-enhancing median filter. Repeating this for each spectrogram frame will result in a percussion-enhanced spectrogram $\mathbf{P}$.

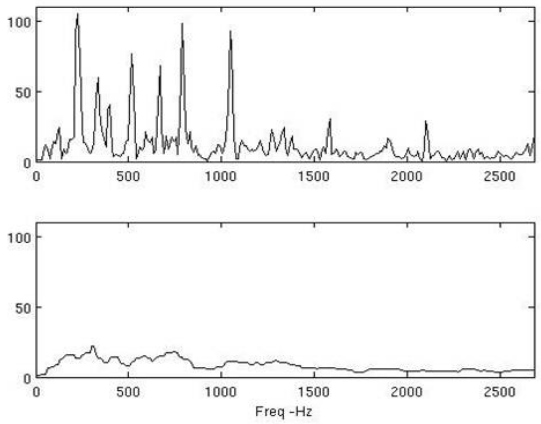

Fig. 3. *Spectrogram time frame from a spectrogram containing a mixture of snare drum and piano a) The original frame, b) the frame after median filtering. It can be seen that a large amount of the energy of the harmonics of the piano have been removed*

The resulting harmonic and percussion suppressed spectrograms could then be inverted to the time domain by applying the phase information from the original spectrogram before performing an inverse short time fourier transform. However, the use of median filtering introduces many artifacts into these spectrogram, and a better strategy to ensure a high quality resynthesis is to use $\mathbf{H}$ and $\mathbf{P}$ to generate masks which can then be applied to the original spectrogram before inversion to the time domain. Of particular interest in this case are masks based on Wiener Filtering. These masks are defined as:

$$\mathbf{M}_{\mathbf{H}h,i} = \frac{\mathbf{H}_{h,i}^2}{(\mathbf{H}_{h,i}^2 + \mathbf{P}_{h,i}^2)} \tag{4}$$

$$\mathbf{M}_{\mathbf{P}h,i} = \frac{\mathbf{P}_{h,i}^2}{(\mathbf{H}_{h,i}^2 + \mathbf{P}_{h,i}^2)} \tag{5}$$

Complex spectrograms are then recovered for inversion from:

$$\hat{\mathbf{H}} = \hat{\mathbf{S}} \otimes \mathbf{M}_{\mathbf{H}} \tag{6}$$

and

$$\hat{\mathbf{P}} = \hat{\mathbf{S}} \otimes \mathbf{M}_{\mathbf{P}} \tag{7}$$

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. A further advantage of using this technique for resynthesis is that the separated signals will sum together to give a perfect reconstruction of the original signal. This is useful for the purposes of remixing the audio, and for upmixing recordings from mono to stereo. Good separations were typically obtained from an STFT with FFT size of 4096 samples, and a hopsize of 1024 samples, when CD quality audio with a sampling rate of 44.1 kHz was used. In this case, l_{perc} and l_{harm} were set to 17.

The above technique has been shown to be effective in separating single channel mixtures of percussion and pitched instruments. It should also be noted that for separation to take place, the percussion instruments do not have to be broadband in the sense that they have energy across the entire spectrogram. Instead, if a percussion instrument is locally broadband in a given portion of the spectrum, determined by the length of the median filter, then the percussion source can be recovered. Further, as will be described below, an adaptation of the technique can also be used for the separation of vocals from single channel mixtures of vocals with both pitched and percussion instruments.

III. Vocal Separation using Multipass Median Filtering-based Separation

In contrast to pitched instruments where the harmonics are typically stable over the course of the entire note or notes played by an instrument, the singing voice constantly varies between voiced regions with a discernible pitch such as when vowels are being sung, and unvoiced regions where consonants and plosives occur. The singing voice moves smoothly back and forth between such regions depending on the words being sung, the duration of the individual voiced and unvoiced parts of the words, and the characteristics of the vocalist. Even in regions where a pitch is discernible, the voice is at best pseudoharmonic, and has often been modeled as a broadband excitation being filtered by formant filters.

When using the harmonic-percussive separation algorithm described above to separate pitched and percussive instruments in cases where singing voice was present, using the parameters used above, it was noted that the voiced parts of the singing tended to be separated with the pitched instruments, while the unvoiced regions tended to be separated with the percussion instruments. Further investigation of this revealed that the proportion of voice which was separated with the pitched instruments varied according to the frequency resolution of the STFT used. At low frequency resolution, around an FFT size of 512 samples, the majority of the voice tended to be separated with the pitched instruments, while at high frequency resolution, such as an FFT size of 16384 samples, the majority of the voice tended to be separated with the percussion instruments.

The reason for this phenomenon is that at low frequency resolution, more and more of the voice energy is collected within a single frequency bin, leading to the singing voice appearing as a harmonic instrument at low frequency resolution. At high frequency resolution, the pseudoharmonic nature of singing voice begins to dominate, and instead of the energy of the various partials of the voice being concentrated within a single frequency bin, the energy is spread out across a range of frequency bins. This is in contrast to pitched instruments,

where, regardless of the frequency resolution used, the energy of the harmonics of a source will occur in a very narrow number of bins around the frequency of the harmonic.

Further, the high frequency resolution used means that correspondingly, the time resolution is lower, and so there is a much greater chance that unvoiced regions of singing will be captured in the same time frame as voiced regions of singing, resulting in further smearing of the singing voice energy across several frequency bins, resulting in the singing voice appearing as a percussion-like instrument from the point of view of the median filtering algorithm. This can be leveraged as a means of performing singing voice separation.

Having described above how the separation of singing voice varies with frequency resolution when performing harmonic-percussion separation, it is proposed to take advantage of this to separate singing voice from mixtures of pitched and percussive instruments by performing a multipass analysis of the signal. There are two potential routes for separation of the vocals from the other instruments. The first is to perform harmonic-percussive separation at a high frequency resolution to yield one signal containing percussion and vocals, and another containing pitched instruments. Harmonic-percussive separation can then be performed at a low frequency resolution to separate the vocals from the percussion instruments. The second route is to perform separation at low frequency resolution intially to yield a pitched instrument and vocals signal, which can then be processed at high frequency resolution to separate the vocals from the pitched instruments. In both these cases, the separated percussion and pitched instruments can be recombined to yield the backing track with the vocals removed.

Apart from the use of STFT-based spectrograms, it is also proposed to investigate the use of a Constant Q spectrogram as a substitute for the low frequency STFT in both of the routes described above.

The Constant Q transform (CQT) is a log-frequency resolution spectrogram [14] and has advantages for the analysis of musical signals, as the frequency resolution can be set to match that of the equal tempered scale used in western music, where the frequencies are geometrically spaced, as opposed to the linear spacing of the STFT. The frequency components of the CQT have a constant ratio of center frequency to resolution, as opposed to the constant frequency difference and constant resolution of the DFT. This constant ratio results in a constant pattern for the spectral components making up notes played on a given instrument, and this has been used to attempt sound source separation of pitched instruments from both single channel and multi-channel mixtures of instruments [15].

Given an inital minimum frequency f_0 for the CQT, the center frequencies for each band can be obtained from:

$$f_k = f_0 2^{\frac{k}{b}} \; (k = 0, 1, ...) \tag{8}$$

where b is the number of bins per octave, and k indexes over the frequency bins. The fixed ratio of center frequency to bandwidth is then given by

$$Q = \left(2^{\frac{1}{b}} - 1\right)^{-1} \tag{9}$$

The desired bandwidth of each frequency band is then obtained by choosing a window of length

$$N_k = Q \frac{f_s}{f_k} \tag{10}$$

where f_s is the sampling frequency. The CQT is defined as

$$X(k) = \frac{1}{N_k} \sum_{n=0}^{N_k - 1} W_{N_k}(n)\, x(n)\, exp^{-j2\pi Q n / N_k} \tag{11}$$

where $x(n)$ is the time domain signal and W_{N_k} is a window function, such as the hanning window, of length N_k.

Until recently, the principal disadvantage of the CQT was that there was no inverse transform. However, recent work by Schoerkhuber and Klapuri has resulted in the developement of an approximate inverse which enables a reasonable quality reconstruction of the original signal, with around 55dB signal-to-noise ratio, thereby allowing the more widespread use of the CQT for the purposes of signal analysis and modification [16].

The use of a CQT results in a low frequency resolution spectrogram, though with logarithmic frequency resolution and so it can be substituted for the low frequency resolution pass in either of the proposed algorithms above. This results in a total of four ways of attempting to separate vocals from mixtures of pitched and percussive instruments. These are outlined in Figure 4, which shows flowcharts of the proposed algorithms, where HP Median denotes Harmonic-Percussive Separation using median filtering.

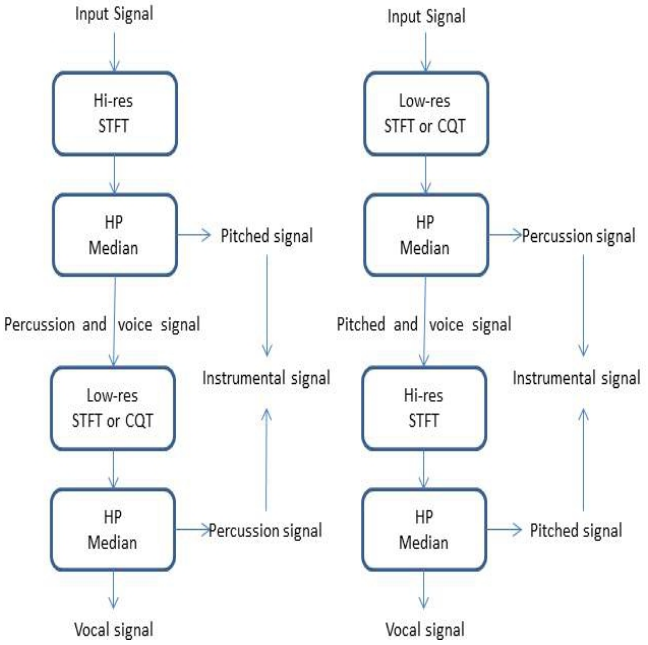

Fig. 4. *Flowcharts showing the algorithms proposed for vocal separation, where HP Median denotes Harmonic-Percussive Separation using median filtering*

In all four versions of the algorithm proposed, good separation of the vocals from the other instruments is possible, though the performance does vary from version to version, as well be seen later in the section on separation performance evaluation. The proposed multipass technique has several

advantages over other algorithms previously proposed for vocal separation from single channel mixtures. Firstly, the algorithm is completely blind, it does not depend on any predominant melody extraction techniques, or on having a score of the melody line available. Secondly, it does not require any pre-trained models of singing voice to function, or models of the instrumental part to function. Thirdly, in contrast to many of the previous algorithms, it is capable of extracting all vocal parts. including harmony vocals, whereas the majority of algorithms focus on solo singing voice. Finally, the proposed algorithm is computationally efficient, and is capable of separating the vocals in near real-time.

Despite this, the algorithm does have its disadvantages. In particular, traces of the other instruments can be heard in the separations, though at much reduced loudness. In particular, traces of some of the percussion instruments can still be heard, with elements of the kick drum often heard with the vocals. This is because at low frequency resolution, the main energy of the kick drum can sometimes be concentrated within a single frequency bin which results in the algorithm perceiving the kick drum as a pitched instrument.

This can be ameliorated to a certain extent through the use of filtering during the masking stage when resynthesising the separated sources. Setting all bins in the vocal mask which have centre frequencies below a cutoff frequency to zero will result in all the energy in those bins being removed from the vocal signal, and restored to either the percussion or pitched signal, as the case may be. In the majority of cases in popular music, setting the threshold to 100 Hz is sufficient to preserve the vocals, while removing some of the effects of the low frequency percussion. This cutoff frequency can easily be adjusted to give better results, if information about the vocal range of the music is known.

Figure 5 shows an example of the separations obtained using the multipass median filtering approach on an excerpt taken from "Sloop John B" by the Beach Boys. In this case, both the vocals and instrumental tracks were available separately and then mixed to form the mixture signal. Figure 5(a) shows the original mixture signal, while figures 5(b) and (c) show the original vocal before mixing and the separated vocal obtained from the algorithm respectively. This was obtained using a CQT spectrogram for the low resolution separation pass, with the low resolution pass performed first. The high-pass filtering approach described above was also used during the masking stage. Similarly 5(d) and (e) show the original instrumental track and the separated instrumental track respectively, obtained from the same method as for the vocals. It can be seen that the vocals have been separated quite well, with the vocal energy predominating in the separated vocal spectrogram, though some artifacts are still present. Similarly, it can be seen that the majority of the vocal energy has been removed from the instrumental track, thereby demonstrating the effectiveness of the proposed vocal separation technique.

While the use of frequency thresholding can remove some of the low frequency percussion or noise from the separation, it was decided to explore the use of alternative approaches to source separation in order to attempt to further improve the vocal separation quality of the algorithm. This is described in the following section.

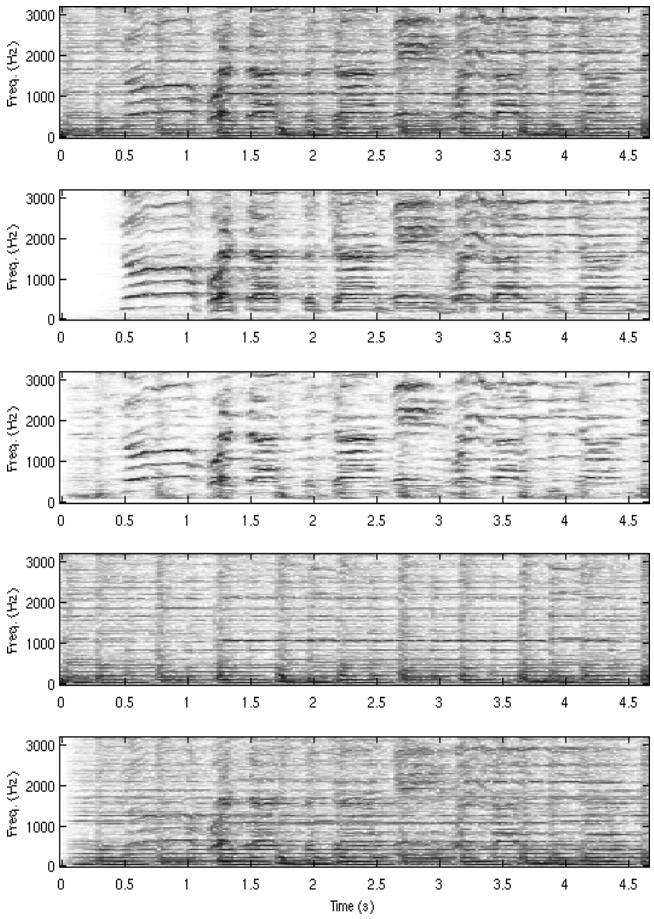

Fig. 5. *Spectrograms obtained from a) the original mixture signal, b) the unmixed vocal track, c) the separated vocal track, d) the unmixed instrumental track, e) the separated instrumental track.*

IV. Post-processing using tensor factorisation techniques

Tensor factorisation models have been used to attempt the separation of percussion instruments from pitched or voiced instruments [17], and as the principal artifacts in the vocal separation are from percussion instruments, it was decided to use this algorithm as a post processing stage. The tensor factorisation algorithm was designed to work on multichannel audio, but functions equally well on single channel mixtures, and the signal model used is described below:

Given an r-channel mixture, magnitude spectrograms are obtained for each channel, resulting in $\mathcal{X}$, an $r \times n \times m$ tensor where n is the number of frequency bins and m is the number of time frames. The tensor is then modelled as:

$$\mathcal{X} \approx \hat{\mathcal{X}} = \sum_{k=1}^{K} \mathcal{G} \circ \langle\langle\langle \mathcal{F}\mathcal{H} \rangle_{\{2,1\}} \mathcal{W} \rangle_{\{3,1\}} \mathcal{S} \rangle_{\{2,1\}} + \sum_{l=1}^{L} \mathcal{M} \circ \mathcal{B} \circ \mathcal{C} \quad (12)$$

where $\hat{\mathcal{X}}$ is an approximation to $\mathcal{X}$. The first right-hand side term models pitched sources, and the second unpitched

or percussion sources. K denotes the number of pitched sources and L denotes the number of unpitched sources. Here, all tensors, regardless of the number of dimensions, are signified by the use of caligraphic letters such as $\mathcal{A}$. $\langle \mathcal{AB} \rangle_{\{a,b\}}$ denotes contracted tensor multiplication of $\mathcal{A}$ and $\mathcal{B}$ along the dimensions a and b of $\mathcal{A}$ and $\mathcal{B}$ respectively. Outer product multiplication is denoted by $\circ$. Further, as all parameters are source specific, the subscript k is implicit in all parameters within the summation.

$\mathcal{G}$ is a tensor of size r, containing the gains of a given pitched source in each channel. $\mathcal{F}$ is of size $n \times n$, where the diagonal elements contain a filter which attempts to model the formant structure of an instrument, thus allowing the timbre of the instrument to alter with frequency. $\mathcal{H}$ is a tensor of size $n \times z_k \times h_k$ where z_k and h_k are respectively the number of allowable notes and the number of harmonics used to model the kth instrument, and where $\mathcal{H}(:, i, j)$ contains the frequency spectrum of a sinusoid with frequency equal to the jth harmonic of the ith note. $\mathcal{W}$ is a tensor of size h_k containing the harmonic weights for the kth source. $\mathcal{S}$ is a tensor of size $z_k \times m$ which contains the activations of the z_k notes associated with the kth source, and in effect contains a transcription of the notes played by the source. For the separation of signals containing pitched instruments only, best results were obtained when the lowest note played by each instrument was used as the lowest note in the source harmonic dictionary $\mathcal{H}$.

For unpitched instruments, $\mathcal{M}$ is a tensor of size r containing the gains of an unpitched source in each channel. $\mathcal{B}$ is of size n and contains a frequency basis function which models the timbre of the unpitched instrument. $\mathcal{C}$ is a tensor of size m which contains the activations of the lth unpitched instrument.

It can be seen that to obtain an estimate of the pitched sources only the first right hand side term of eqn 12 needs to be recontructed, and for the unpitched sources, only the second right hand side term needs to be used. The model can also be collapsed to the single channel case by eliminating both $\mathcal{G}$ and $\mathcal{M}$ from the model.

The generalised Kullback-Leibler divergence is used as a cost function to measure reconstruction of the original data as it has been shown to be effective for audio sound source separation [20]:

$$D\left(\mathcal{X} \parallel \hat{\mathcal{X}}\right) = \sum \mathcal{X} \log \frac{\mathcal{X}}{\hat{\mathcal{X}}} - \mathcal{X} + \hat{\mathcal{X}} \tag{13}$$

where summation takes place over all dimensions of $\hat{\mathcal{X}}$. Using this measure, iterative multiplicative update equations can be derived for each of the model variables. These are presented in [17] and, due to space limitations, are not presented here. From these, separation of pitched and unpitched instruments can be attempted. It was noted in testing this approach that the separation quality was better without the use of the gamma-chain priors used in [17], and so all parameters related to the gamma-chain priors have been set to zero, eliminating them from the update equations. This is because the gamma-chain priors favour continuity over time to capture pitched instruments, and that this does not hold well for singing voice.

When used as a post-processing step for vocal separation, it was noted that the lowest "source" of the separated pitched part of the signal contained mainly noise related to the kick drum and the bass guitar, and so this was not used when reconstructing the voice signal, but was instead added back to the instrumental track. With regards to the percussive part separated by the algorithm it was found that some of the noise or unpitched basis functions contained high frequency components of the vocals while others contained actual percussive events. If the high frequency vocal components were removed, the recovered voice sounded much less brighter. Further, the number of components required to capture the percussion events was found to vary from signal to signal, and it would require manual intervention to decide which noise components contained vocal information. As a result, it was decided to leave the noise part of the signal in the vocal separations, though in some cases improved separation can be obtained by manually eliminating percussive basis functions. As will be seen later, the tensor factorisation stage can considerably improve the separation of the vocals from the mixture signal. However, the downside of using the tensor factorisation-based approach lies in the fact that it is significantly more computationally intensive than the median filtering-based approach, taking between 5-10 times real-time to run.

V. Re-separation using non-negative matrix partial cofactorisation

Another approach of potential interest as a post-processing step to improve the separations is Non-negative matrix partial cofactorisation. Non-negative matrix partial cofactorisation was recently proposed as a means of separation of drum sounds from polyphonic music signals containing both percussion and pitched instruments [21]. This technique assumed that there existed some prior examples of drums or percussion instruments available. These were then used to create a "drums-only" spectrogram. The spectrogram of the mixture signal and the "drums-only" spectrogram were then decomposed simultaneously, while sharing some frequency basis functions between the two spectrograms, to force some basis functions to be associated with the drums only, thereby allowing the separation of the drums from the polyphonic music signal.

This approach can be formalised as follows, given a polyphonic music mixture spectrogram $\mathbf{X}$, and a "drums-only" spectrogram $\mathbf{Y}$ then simultaneously decompose these matrices as:

$$\mathbf{X} \approx \hat{\mathbf{X}} = \mathbf{A}_H \mathbf{S}_H + \mathbf{A}_P \mathbf{S}_P \tag{14}$$

$$\mathbf{Y} \approx \hat{\mathbf{Y}} = \mathbf{A}_P \mathbf{S}_{P1} \tag{15}$$

where $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$ are approximations to $\mathbf{X}$ and $\mathbf{Y}$ respectively, $\mathbf{A}_H$ contains the frequency basis functions associated with the harmonic or pitched instruments in the spectrogram and $\mathbf{S}_H$ contains the associated time activation basis functions. $\mathbf{A}_P$ contains the frequency basis functions associated with the drums or percussion instruments, and which is common to the factorisation of both matrices. $\mathbf{S}_P$ contains the time activation basis functions of the drums in the mixture signal, while $\mathbf{S}_{P1}$

contains the time activation basis functions in the "drums-only" spectrogram. The pitched part of the spectrogram can then be reconstructed as:

$$\mathbf{X}_H = \mathbf{A}_H \mathbf{S}_H \tag{16}$$

and the percussive part of the signal as:

$$\mathbf{X}_P = \mathbf{A}_P \mathbf{S}_P \tag{17}$$

In this case, prior knowledge of drum sounds, though not necessarily the exact drums in the mixture signal, was used to guide the factorisation of the mixture signal. This partial co-factorisation was carried out using the least-squares error between the factorisations and the spectrograms as a cost function, and gave separation results comparable with other state of the art approaches for drum separation. However, a potential problem lies in the possible mismatch in spectral characteristics of the drums available as prior knowledge, and the drums in the actual recording.

It can be seen that such a partial cofactorisation approach could be adapted to deal with other sources, given prior knowledge or examples of the other instruments to be separated. It is proposed to take advantage of this inherent flexibility in the partial cofactorisation approach in an attempt to further improve the separation of the vocals from the instrumental track. To this end, the existing separated vocal obtained from the previously described algorithms will be used as prior knowledge to drive the partial cofactorisation algorithm in order to separate the vocals and the instrumental backing track from the original mixture spectrogram. This is a novel use of partial cofactorisation in that an existing separation is being used as a guide to re-separate the original mixture.

As noted above, the partial cofactorisation approach described above made use of a least-squares cost function. However, for musical signals in general, the generalised Kullback-Liebler divergence has been found to give better separation performance. To this end, we present an algorithm for non-negative partial cofactorisation based on this divergence:

$$D = \sum \left(\mathbf{X} log \frac{\mathbf{X}}{\hat{\mathbf{X}}} - \mathbf{X} + \hat{\mathbf{X}} \right) + \sum \left(\mathbf{Y} log \frac{\mathbf{Y}}{\hat{\mathbf{Y}}} - \mathbf{Y} + \hat{\mathbf{Y}} \right) \tag{18}$$

with

$$\hat{\mathbf{X}} = \mathbf{A}_T \mathbf{S}_T + \mathbf{A}_V \mathbf{S}_V \tag{19}$$

and

$$\hat{\mathbf{Y}} = \mathbf{A}_V \mathbf{S}_{V1} \tag{20}$$

where $\mathbf{X}$ is the mixture spectrogram, $\mathbf{Y}$ is the separated vocal spectrogram, $\mathbf{A}_T$ and $\mathbf{S}_T$ contain the frequency and time basis functions for the instrumental track and $\mathbf{A}_V$ contains the common frequency basis functions between the two input matrices associated with the vocals. $\mathbf{S}_V$ and $\mathbf{S}_{V1}$ contain the time basis functions for the vocal frequency basis functions for matrices $\mathbf{X}$ and $\mathbf{Y}$ respectively. Further, the summations take place elementwise over all entries.

Iterative multiplicative update equations can be derived for each of the model variables in a manner simliar to that of standard NMF [22]. These update equations take the form

$$\mathbf{R} = \mathbf{R} \otimes \frac{\bigtriangledown^{-}_{\mathbf{R},D}}{\bigtriangledown^{+}_{\mathbf{R},D}} \tag{21}$$

where $\mathbf{R}$ represents a given variable in the model to be updated, D denotes the generalised Kullback-Liebler divergence, and where $\bigtriangledown^{-}_{\mathbf{R},D}$ and $\bigtriangledown^{+}_{\mathbf{R},D}$ represent the negative part and the positive part respectively of the partial derivative of the reconstruction metric with respect to $\mathbf{R}$.

The update equations for each of the parameters are now given below:

$$\mathbf{A}_T = \mathbf{A}_T \otimes \frac{\mathbf{P}\mathbf{S}_T^{'}}{\mathbf{O}_X \mathbf{S}_T^{'}} \tag{22}$$

$$\mathbf{S}_T = \mathbf{S}_T \otimes \frac{\mathbf{A}_T^{'}\mathbf{P}}{\mathbf{A}_T^{'}\mathbf{O}_X} \tag{23}$$

$$\mathbf{A}_V = \mathbf{A}_V \otimes \frac{\mathbf{P}\mathbf{S}_V^{'} + \mathbf{Q}\mathbf{S}_{V1}^{'}}{\mathbf{O}_X \mathbf{S}_V^{'} + \mathbf{O}_Y \mathbf{S}_{V1}^{'}} \tag{24}$$

$$\mathbf{S}_V = \mathbf{S}_V \otimes \frac{\mathbf{A}_V^{'}\mathbf{P}}{\mathbf{A}_V^{'}\mathbf{O}_X} \tag{25}$$

$$\mathbf{S}_{V1} = \mathbf{S}_{V1} \otimes \frac{\mathbf{A}_V^{'}\mathbf{Q}}{\mathbf{A}_V^{'}\mathbf{O}_Y} \tag{26}$$

where $\otimes$ indicates elementwise multiplication and $'$ indicates matrix transpose. $\mathbf{P} = \mathbf{X}/\hat{\mathbf{X}}$, $\mathbf{Q} = \mathbf{Y}/\hat{\mathbf{Y}}$, $\mathbf{O}_X$ is an all-ones matrix with the same dimensions as $\mathbf{X}$, and $\mathbf{O}_Y$ is an all-ones matrix with the same dimensions as $\mathbf{Y}$.

The re-separated vocal spectrogram can then be obtained from:

$$\mathbf{X}_V = \mathbf{A}_V \mathbf{S}_V \tag{27}$$

and the instrumental spectrogram obtained from:

$$\mathbf{X}_T = \mathbf{A}_T \mathbf{S}_T \tag{28}$$

Rather than resynthesise directly from these spectrograms, the spectrograms are used to generate masks in the manner described earlier in the paper, as this leads to better quality resynthesis of the re-separated sources.

The motivation for using the previously separated vocal to re-separate the vocals using partial co-factorisation is that, despite the good quality separations obtained using the algorithms presented in the previous sections, there will still be artifacts from the other instruments in the vocal separation. However, these artifacts will be low in volume in comparison to the vocal in the separated signal. Therefore, when the algorithm attempts partial cofactorisation, these low volume artifacts should end up being captured in the basis functions that belong to the instrumental track as opposed to the vocal basis functions, thereby reducing artifacts in the re-separated vocal, and improving the quality of the re-separated instrumental track.

The validity of this argument is evinced by the improved quality separation results obtained, as will be seen in the next section. However, as with the use of tensor factorisation, the downside of using partial cofactorisation lies in the increased computational demand and time taken to perform the cofactorisation.

VI. Separation Performance

A. Test Materials

In order to test the effectiveness of the algorithm, a set of test signals is required. To this end, pieces of music where the vocals and instrumental track are available separately are required. Fortunately, within the back catalogue of the Beach Boys, such a set of recordings is available. In particular, a number of Beach Boys tracks are available as split stereo recordings where all the vocals are in one channel and the instrumental track in the other channel [18]. Further, there are a number of tracks for which the vocals and the instrumental tracks are available separately [19]. These were manually resynchronised in a digital audio editor to allow the creation of mono mixes from these source materials.

In total, 30 mono signals of approximately 45 seconds duration were created from excerpts from 10 Beach Boys tracks. This length was chosen due to the memory and computational constraints of some of the algorithms used. Three different scenarios were considered, firstly the case where the vocals and instrumental tracks were mixed as they were, these are refered to as the 0dB mixes and secondly, where the amplitude of the vocals was raised by 6dB relative to the instrumental track, these are refered to as the 6dB mixes. Finally another set of mixes were prepared where the amplitude of the vocals was dropped by 6dB relative to the instrumental track, refered to as the -6dB mixes. The use of these mixes will allow the performance of the algorithms to be measured in a range of different conditions, thereby giving a better idea of the overall performance of the algorithms.

B. Algorithms and Parameters

As already noted in Section III there are four proposed ways to perform multipass median filtering-based separation (MMFS), depending on whether the high resolution pass is performed first or second, and on whether a linear spectrogram or a Constant Q spectrogram is used for the low frequency resolution pass is used. Further, for each of these four ways, the performance was measured for four different algorithms, the first is the basic MMFS algorithm. In all four ways to perform MMFS, the high frequency resolution FFT size was 16384 samples, with a hopsize of 2048 samples, with median filters of length 17 frames and 17 frequency bins were used for both the harmonic and percussive filters respectively. The low frequency resolution was 1024 samples with a hopsize of 256 samples, with median filters of length 17 again being used. For the CQT spectrogram, a resolution of 24 bins per octave was used, and median filters of length 7 frequency bins and 17 frames were used for the percussive and harmonic median filters respectively. Here, 7 frames were used for the percussive filter due to the low frequency resolution of the CQT.

The second algorithm considers is MMFS with high pass filtering during the masking stage, with a cutoff of 100Hz (MMFS+H). Next, MMFS in conjunction with a tensor factorisation approach was considered (MMFS+T). Here an FFT of 4096 samples and a hopsize of 1024 samples was used. The tensor factorisation approach divided the frequency range of the signal into four overlapping bands covering different pitched notes, the first covering an octave from 55 Hz, the second covering two octaves from 110 Hz, the third another two octaves from 220 Hz, while the fourth covered two octaves from 880 Hz, with 10 harmonics used to approximate the timbre of each note within each band. Three noise-based basis functions were also used.

Finally, the separated vocal from MMFS+T was fed to the partial co-factorisation algorithm (MMFS+T+CF). Here, better results were obtained with an FFT size of 16384 and a hopsize of 2048. 250 basis functions each were used to approximate the vocal and instrumental tracks.

These four algorithms in conjunction with the four ways to perform MMFS result in 16 different methods to perform vocal separation. Each of these methods are then tested for the three mixing scenarios described above.

C. Evaluation metrics

In order to quantitatively measure the quality of the separations obtained, a set of separation performance metrics must be used. A commonly used set of metrics are those defined by Vincent et al [23]. Here the recovered time domain signal is decomposed into the sum of three terms, with reference to the original unmixed source signal:

$$s_{rec} = s_{tar} + e_{int} + e_{art} \tag{29}$$

where s_{rec} is the recovered source signal, s_{tar} is the portion of the recovered signal that relates to the original or target source, e_{int} is the portion that relates to interference from other sources, and e_{art} is the portion that relates to artifacts generated by the separation technique and/or the resynthesis method. Based on this decomposition, source separation metrics were then defined.

The first of these, Signal to Distortion ratio (SDR), provides a measure of the overall quality of the sound source separation:

$$SDR = 10log_{10}\frac{\|s_{tar}\|^2}{\|e_{int} + e_{art}\|^2} \tag{30}$$

The Signal to Interference ratio (SIR) provides a measure of the presence of other sources in the separated source:

$$SIR = 10log_{10}\frac{\|s_{tar}\|^2}{\|e_{int}\|^2} \tag{31}$$

Finally, the Signal to Artifacts ratio (SAR) provides a measure of the artifacts present in the signal due to separation and/or resynthesis:

$$SAR = 10log_{10}\frac{\|s_{tar} + e_{int}\|^2}{\|e_{art}\|^2} \tag{32}$$

These metrics are invariant to scaling factors and were calculated using the BSS_EVAL toolbox available at [24].

However, a shortcoming of these metrics is that they do not necessarily correlate well with the perceptual quality of the separated signals. Nevertheless, SIR in particular provides a good measure of the rejection of the other sources in comparison to the other sources present.

In the context of vocal separation and suppression, these metrics are used to measure individually the separation quality of the isolated vocal, and the instrumental track with the vocal suppressed.

D. Test Results

The separation performance results for the separation of the vocals from polyphonic audio are presented in Table I. It can be seen that, as expected, the algorithm performs worse for the -6dB mixes, which represent a "worst-case" scenario where the vocals are very low in the instrumental mix. It can be seen that the baseline MMFS algorithm is capable of some degree of separation, even in this case, particularly when using the CQT for the low-res pass, where with improvements of SIR of around 3dB possible. The use of high pass filtering improves the SIR results by a further 2dB, while MMFS+T results in a 5dB increase in SIR over that of MMFS+H. The use of cofactorisation improves this result by on average 2.5 dB, resulting in a maximum SIR of 13.25 dB for the -6dB mixes. This is a very good level of rejection considering the adverse mixing conditions presented to the algorithms. In all cases the SAR and SDR scores are quite low, this is to be expected due to the low level of the vocals in the mixture signals, which can make it difficult to isolate the vocals without the presence of artifacts. Also to be noted is that there is a trade-off between improving SIR and reducing SAR. As SIR performance increases, it results in increasing artifacts due to the separation algorithm, thereby reducing SAR and SDR.

On listening to the separations obtained from the 0dB mixes, the principal artifacts in the vocal separation are due to the presence of percussion instruments, with some traces of the pitched instruments in the separated vocals. Nonetheless, it can be clearly heard that the algorithms have still managed to separate the vocals to some degree, even under adverse separation conditions, with the vocals still predominant in the separated sources.

As expected, it can be seen that there is a large jump in separation performance across all metrics for the 0dB mixes. Again, as the complexity of the algorithms increases, so does the separation quality obtained. Further, the methods using the CQT again outperform those using a linear spectrogram for the low resolution pass. In the 0dB mixes, it should be noted that both MMFS and MMFS+H are capable of obtaining very good separation of the vocal tracks, obtaining an average SIR of 11.44 dB for MMFS+H when using a CQT with the low resolution pass performed before the high resolution pass. This shows that these simple algorithms with low computational load are capable of giving good separation results without recourse to the computationally intensive tensor and matrix factorisation separation stages. Nevertheless, when these additional stages are used, there is a large jump in performance with the SIR metric improving by a further 10-12 dB. Again the trade-off between improved SIR and reduced SAR and SDR can be noted. On listening to the separated vocals it can be noted that there is a notable improvement in sound quality of the separated vocals, and that the presence of artifacts due to drums has been considerably reduced.

Finally, the separation performance again improves when the 6dB mixes are presented to the algorithm. Of interest here is the fact that for the first time, when using MMFS+T, the use of a linear low-res spectrogram outperforms the use of a CQT, and that the cofactorisation stage does not improve performance. This suggests that under ideal conditions where the vocals are very high in the mix, there is no requirement for the cofactorisation stage when separating the vocals. However, as will be seen later, the use of cofactorisation in this case does improve the separation of the instrumental track.

Table II shows the separation performance for the instrumental tracks. It can be seen that, as would be expected, the separation performance is worst for the 6dB mixes and improves as the level of the instrumental track rises. It can also be observed that as the algorithms increase in complexity, the separation performance of the instrumental track consistently improves in terms of SIR, though not to the same extent as the vocal separation. It can be seen that SIR is consistently lower for the separated instrumental tracks than for the separated vocals, with consistently more of the vocals found in the separated instrumental tracks than vice-versa. Unlike the separated vocal tracks, there is no trade-off between improved separation and increasing amount of artifacts, with both SAR and SDR improving along with SIR. Also of note is the fact that the use of a linear spectrogram for the low-resolution pass consistently outperforms that of the CQT for separating the instrumental tracks. On listening to the separated tracks, traces of the vocals can be heard in the separated instrumental track, though the level of the vocals is clearly reduced in all cases. In particular, the use of cofactorisation results in a noticeable improvement in the separation of the instrumental tracks, in general reducing the amount of the vocals heard in the separated tracks.

Overall, the presented set of algorithms are capable of extracting the vocal tracks well from polyphonic music. In general, the use of the CQT results in improved performance for the separation of vocal tracks, while the use of the linear low-res pass improves that of the instrumental tracks. Further, the results are slightly better for the case where the low resolution pass is performed before the high resolution pass. It can be seen that the low complexity algorithms (MMFS and MMFS+H) are capable of good vocal separation results, and so could find application as a lightweight separation algorithm for use as preprocessing for other tasks, such as predominant melody estimation. However, for remixing purposes the use of both the tensor factorisation and partial cofactorisation stages result in improved separation quality. This is most noticeable in the quality of the separated instrumental tracks, where the use of partial cofactorisation results in much better separation quality and reduced artifacts. Audio examples of vocal separations obtained from real-world recordings can be found at `http://eleceng.dit.ie/derryfitzgerald/index.php?uid=489&menu_id=46`.

VII. Conclusions and Future Work

Previous methods for the separation of singing voice from single-channel recordings of polyphonic music have been discussed and problems with existing methods highlighted. In particular, many of the existing approaches require use of prior knowledge about the signal or sources to be separated. Many algorithms require either knowledge of the vocal melody to aid the separation, or attempt to estimate this knowledge from

		MMFS			MMFS+H			MMFS+T			MMFS+T+CF		
		SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
-6dB	CLH	-2.30	3.46	1.76	-1.87	5.16	1.42	-1.23	10.5	0.30	-1.40	12.99	-0.32
	CHL	-2.51	2.88	1.88	-2.02	4.67	1.50	-0.87	10.84	0.46	-1.00	**13.25**	-0.06
	LLH	-3.88	1.10	1.31	-2.99	3.38	0.99	-1.48	8.71	0.88	-1.38	10.41	0.74
	LHL	-3.80	1.32	1.24	-2.93	3.60	0.92	-1.25	9.94	0.41	-1.14	11.35	0.29
0dB	CLH	1.14	9.71	2.84	1.34	11.44	2.70	1.10	20.03	1.52	0.88	**22.30**	1.18
	CHL	1.04	9.04	2.93	1.28	10.86	2.76	1.17	21.22	1.62	1.10	22.29	1.39
	LLH	0.28	7.48	2.56	0.75	9.75	2.42	1.54	18.98	2.29	1.56	19.28	2.30
	LHL	0.25	7.59	2.47	0.71	9.86	2.34	1.49	18.24	1.88	1.83	19.81	2.12
6dB	CLH	2.74	15.96	3.30	2.80	17.88	3.24	1.58	23.69	1.73	1.09	23.89	1.20
	CHL	2.75	15.22	3.38	2.83	17.17	3.31	1.72	25.12	1.88	0.97	23.97	1.08
	LLH	2.40	13.81	3.16	2.55	16.14	3.10	1.89	**25.85**	2.03	1.54	24.27	1.66
	LHL	2.33	13.83	3.08	2.48	16.14	3.03	2.07	23.97	2.18	1.88	24.12	1.97

TABLE I

VOCAL SEPARATION PERFORMANCE FOR THE VARIOUS ALGORITHMS PROPOSED IN THIS PAPER. HERE -6DB,0DB AND 6DB INDICATE THE AVERAGE RESULTS OBTAINED FOR THE -6DB,0DB AND 6DB MIXES RESPECTIVELY. CLH INDICATES THE USE OF A CQT SPECTROGRAM WITH THE LOW FREQ. RESOLUTION PASS PERFORMED BEFORE THE HIGH FREQ. PASS, LLH INDICATES THE SAME CONFIGURATION EXCEPT WITH A LINEAR SPECTROGRAM FOR THE LOW FREQ. PASS. CHL INDICATES THE USE OF A CQT SPECTROGRAM, WITH THE HIGH FREQ. RESOLUTION PASS FIRST, AND CLH INDICATES THE USE OF A CQT, WITH THE LOW FREQ. RESOLUTION PASS PERFORMED FIRST. MMFS INDICATES MULTIPASS MEDIAN FILTER-BASED SEPARATION, MMFS+H INDICATES THE ADDITION OF A HIGH PASS FILTER TO MMFS, MMFS+T INDICATES MMFS IN CONJUNCTION WITH A TENSOR FACTORIZATION-BASED SEPARATION PASS, AND MMFS+T+CF INDICATES THE ADDITION OF A PARTIAL COFACTORISATION PASS TO THE PREVIOUS METHOD.

		MMFS			MMFS+H			MMFS+T			MMFS+T+CF		
		SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
-6dB	CLH	6.71	8.24	12.78	6.91	8.46	12.94	7.37	8.74	13.84	7.24	11.25	9.86
	CHL	6.54	8.16	12.43	6.81	8.41	12.70	7.47	8.88	13.82	7.42	11.50	9.99
	LLH	5.70	8.67	9.46	6.22	9.14	9.99	7.11	9.75	11.18	6.71	11.48	8.86
	LHL	5.77	7.99	10.54	6.34	8.41	11.27	7.42	9.05	13.21	7.09	**11.79**	9.30
0dB	CLH	1.65	2.61	10.78	1.85	2.82	10.88	2.15	3.04	11.45	2.62	5.61	6.96
	CHL	1.52	2.50	10.63	1.77	2.74	10.79	2.18	3.08	11.39	2.78	5.83	6.93
	LLH	1.19	2.84	8.23	1.65	3.30	8.54	2.50	4.04	9.38	2.75	6.14	6.55
	LHL	1.13	2.45	9.11	1.57	2.85	9.52	2.47	3.57	10.78	3.19	**6.59**	6.89
6dB	CLH	-3.84	-3.02	8.83	-3.64	-2.82	8.88	-3.53	-2.75	9.15	-3.08	-0.58	3.99
	CHL	-3.96	-3.15	8.84	-3.72	-2.91	8.91	-3.57	-2.83	9.31	-3.18	-0.93	4.42
	LLH	-4.26	-2.98	6.67	-3.82	-2.55	6.82	-3.31	-2.11	7.30	-2.86	-0.30	4.03
	LHL	-4.18	-3.11	7.50	-3.78	-2.72	7.67	-3.09	-2.11	8.29	-2.57	**0.16**	3.88

TABLE II

INSTRUMENTAL TRACK SEPARATION PERFORMANCE FOR THE VARIOUS ALGORITHMS PROPOSED IN THIS PAPER. ALL ABBREVIATIONS ARE AS IN TABLE VI-D

the signal, which can lead to erroneous results where the pitch is not detected properly. Further, many methods also require techniques that can distinguish regions containing vocals from regions without vocals in order for separation to proceed. Other methods require training data, such as a large amount of previously recorded vocal excerpts to generate models of the singing voice. The problem with such a model lies in the wide varity of timbres that vocalists can produce, making it difficult for the training data to adequately capture a given voice, particularly if the vocal timbre is not similar to an example in the training database. Further, all of the above methods are designed to work with solo voice or singing and are not designed to deal with vocal harmony.

Following on from this, a simple but effective median filtering-based harmonic-percussive separation algorithm was described, and it was shown that the performance of this algorithm in the presence of singing voice varied with the frequency resolution of the spectrogram used. High frequency resolution led to the separation of the voice with the percussion instruments, while low frequency resolution resulted in the vocal being separated with the pitched instruments.

It was then proposed to take advantage of this fact to perform single channel voice separation by using a novel multipass version of the harmonic-percussive separation algorithm. Four versions of this algorithm were proposed, depending on whether the high frequency resolution pass was performed first or second, and on whether a CQT or a low frequency linear spectrogram was used for the low resolution pass. All four versions were found to peform well in the separation of vocals, with the use of a CQT giving better results for vocal extraction, but a linear spectrogram performing better for the separation of the instrumental track.

However, there are still artifacts, principally due to the percussion instruments, present in the separations. These can be ameliorated to some extent through the use of high pass filtering, but improved results were obtained through the addtion of a tensor factorisation-based separation algorithm, which considerably reduced the artifacts obtained in the separation.

Finally a novel use of non-negative partial cofactorisation was proposed in order to re-separate the vocals from the original polyphonic music mixture. Here, the vocal separation obtained from the previous algorithm was used as a guide when factorising the original signal into vocal parts and instrumental parts, with the vocal part of the original mixture and the existing separation sharing a common set of frequency basis functions. This resulted in further improvements in separation performance, particularly in the case of separating the instrumental track from the vocals.

The proposed algorithms were tested on a real-world dataset and found to give good separation of vocals, including vocal harmonies, which represents an advance over existing research on single channel singing voice separation. It was noted that the inital multipass median-filtering based algorithms are computationally efficient and simple to implement while still capable of giving good separation, making them suitable as a preprocessing stage for other tasks such as predominant melody estimation. The factorisation-based extensions are considerably more computationally intensive than the median filter based algorithms, but do result in considerably improved separation, and can be used where better quality is required.

Future work will concentrate on the use of this algorithm in the context of upmixing old single channels from mono to stereo or to 5.1 surround sound, as well as investigating other ways of improving the separation quality obtained from the vocal extraction algorithms. For example, the system as currently implemented makes no attempt to distinguish between regions where vocals are present, and where vocals are not. The incorporation of such information should further improve the vocal separation capabilities of the algorithms in this paper. Also, the ability to automatically detect which noise basis functions belong to drum sounds in the tensor factorisation stage would further improve results.

REFERENCES

[1] S. Sofianos, A. Ariyaeeinia and R. Polfreman, *Singing Voice Separation based on Non-Vocal Indpendent Component Subtraction and Amplitude Discrimination*, Proc. of the 13th International Conference on Digital Audio Effects (DAFx-10), Graz, Austria 2010.

[2] P. Comon, *Independent Component Analysis - a new concept?*, Signal Processing, 36, pp. 287-314, 1994

[3] D. Barry, E. Coyle and R. Lawlor, *Sound Source Separation: Azimuth Discrimination and Resynthesis*. Proc. of the 7th International Conference on Digital Audio Effects (DAFx-04), Naples, Italy 2004.

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

[5] G. Hu and D. Wang, *Monaural Speech Segregation based on pitch tracking and amplitude modulation*, IEEE Transactions on Neural Networks, 2004.

[6] 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.

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

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

[9] C. Hsu and J. Jang, *Singing Pitch Extraction by Voice Vibrato/Tremelo estimation and instrument partial deletion*, International Society for Music Information Retrieval Conference, 2010.

[10] 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

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

[12] 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.

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

[14] J. Brown, *Calculation of a Constant Q Spectral Transform*, J. Acoust. Soc. Am. 89 425-434, 1991.

[15] D. FitzGerald, M. Cranitch,and E. Coyle, *Shifted Non-negative Matrix Factorisation for Sound Source Separation*, Proc. of the IEEE conference on Statistics in Signal Processing, Bordeaux, France, July 2005.

[16] C. Schoerkhuber and A. Klapuri,*Constant-Q transform toolbox for music processing* 7th Sound and Music Computing Conference, Barcelona, Spain 2010.

[17] D. FitzGerald, M. Cranitch,and E. Coyle, *Using Tensor Factorisation Models to Separate Drums from Polyphonic Music*, Proc. of the 12th International Conference on Digital Audio Effects (DAFx-09), Como, Italy 2009.

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

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

[20] D. FitzGerald, M. Cranitch,and E. Coyle, *Extended Nonnegative Tensor Factorisation Models for Musical Sound Source Separation*, Computational Intelligence and Neuroscience, 2008.

[21] J. Yoo, M. Kim, K. Kang, S. Choi, *Nonnegative matrix partial cofactorization for drum source separation*, Proc. of the IEEE Conference on Acoustics, Speech, and Signal Processing, Dallas, 2010.

[22] D. Lee, and H. Seung, *Algorithms for non-negative matrix factorization*, Adv. Neural Info. Proc. Syst. 13, 556-562 (2001).

[23] 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.

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

Derry FitzGerald graduated in Chemical Engineering from Cork Institute of Technology in 1995. Having worked as a chemical engineer for a number of years, he returned to college to complete an M.A. in Music Techology at Dublin Institute of Technology in 1999. Following on from that he completed his PhD in 2004, again at Dublin Institute of Technology, on the topic of the automatic separation and transcription of percussion instruments. Since then, he has worked as a post-doctoral researcher at Cork Institute of Technology, before taking up his current position as Stokes Lecturer at Dublin Institute of Technology in 2008. His research interests lie in the areas of sound source separation and automatic music transcription.

Mikel Gainza graduated from the university of Zaragoza (Spain) and the Dublin Institute of technology with an honours degree in Electrical/Electronic Engineering. Following the completion of his undergraduate studies he joined EDSN in Paris, where he worked as a training engineer in the radio communications domain. In 2002, he returned to the Dublin Institute of technology and completed his PhD research in Digital Audio Signal Processing in 2006. He is currently involved in the Institute's Audio Research Group, where he works as a senior researcher in several projects. This includes the EU Framework project EASAIER (Enabling Access to Sound Archives through Integration, Enrichment and Retrieval)and the IMAAS project, which is funded by Enterprise Ireland.