Improved Stereo Instrumental Track Recovery using Median Nearest-Neighbour Inpainting

Derry FitzGerald[†] **and Rajesh Jaiswal**[*]

Audio Research Group
Dublin Institute of Technology,
Kevin St, Dublin 8, IRELAND

E-mail: [†]`derry.fitzgerald@dit.ie` [*]`rajesh.enc@gmail.com`

Abstract — **Several algorithms have been proposed for vocal removal which operate by finding the position of the vocals in the stereo field and removing the time-frequency bins associated with that position. However, in many cases, there will be other instruments such as drums and bass guitar present in the same position. These instruments will be removed along with the vocals and so adversely affect the sound quality of the recovered instrumental track. We present a method for estimating the missing information in the removed time-frequency bins, while still suppressing vocals, allowing recovery of an improved stereo instrumental track.**

Keywords — **Sound Source Separation, Vocal Removal, Audio Inpainting**

I Introduction

The removal of vocals from a stereo mixture of a piece of music to create an instrumental backing track has numerous applications such as the creation of karaoke backing tracks,or as a preprocessing stage for chord estimation. As a first approximation, modern stereo recordings can be considered to have been created by recording each of the sources individually and then summing and distributing the sources across two channels to create a stereo mixture. Localisation of the sources is achieved typically by varying the gain of a given source in each of the channels. Several simple, but effective stereo sound source separation algorithms have been proposed which take advantage of this mixture model[1, 2, 3]. These algorithms are capable of extracting or removing vocals using panning information via user identification of the source position. With modern mixing techniques, the vocal is often found in the center position, i.e. the vocal is equally loud in both channels.

These algorithms all assume that different sound sources are positioned at unique points in the stereo field. They then transform each of the channels of the stereo mixture to the time-frequency domain via the Short-Time Fourier Transform (STFT). The algorithms then proceed to identify which time-frequency bins can be associated with a given point in the stereo field by evaluating a set of parameters obtained from comparison of the channel spectrograms. For example, [1] uses gain scaling and subtraction to identify which time-frequency bins can be associated with a given pan position, while Avendano [3] uses the ratio of the amplitudes of the bins in each channel.

As a result of time-frequency overlap with other sources, the estimated position of the time-frequency bins will usually not be exactly that of the actual source position. This necessitates choosing time-frequency bins which are found to be in a user-defined region around the chosen point in the stereo field to be separated. Inherent in this is a trade-off between capturing more bins belonging to the source and the increased presence of other sources in the separated signal. Too wide a region gives good recovery of the source but with other instruments becoming more audible, while too narrow a region results in less interference at the expense of a reduction in sound quality of the source to be separated.

Once the time-frequency bins associated with the vocals have been determined, a stereo backing track can be obtained by zeroing the relevant bins in the STFT of each channel and inverting the modified STFTs to yield time domain waveforms with the vocals removed. Traces of the vocals will still typically be present, particularly if a stereo reverberation or delay effect has been used on the vocals, but the loudness of the vocals will have been considerably reduced in comparison to the original mixture.

A considerable shortcoming of these approaches lies in the assumption that each source occupies a unique point in the stereo field. As already noted, in modern mixes, vocals are typically centred between the two channels. However, other sources are often positioned at that point in the stereo field too. In particular, snare and kick drums as well as bass guitar often occupy this point in the stereo field as well. This means that the bulk of the energy associated with these sources will be removed from the original mixture along with the vocals and this will adversely affect the quality of the recovered instrumental backing track. All of these sources will usually have significant low frequency energy and so a simple way to ameliorate this is the use of a high-pass filtering approach. Here any bins below a user determined cut-off frequency are assumed as being non-vocal, and are so allocated to the instrumental backing track. In the absence of prior knowledge about the pitches of the vocal melody, a cutoff frequency of 100Hz has been demonstrated to work well in most cases [4]. In particular, this helps recover the low-frequency information related to the drums and bass, but does not recover the higher harmonics of the bass guitar or the high frequency transient information related to the drum sources. Also due to their broadband noise-based nature, the drum sounds will suffer time-frequency overlap more than other types of sources and so some energy from these sources will also be present in the separated instrumental track, but at considerably reduced levels than in the original mixture. Further, the transients of the drums remaining in the separated signal are smeared due to the missing information, resulting in audible artefacts in the resynthesis.

The recovered instrumental STFTs will have a large number of bins with no energy as these bins have been determined to belong to the vocal source. It can be seen that a method capable of estimating the missing information in these time-frequency bins would be advantageous in attempting to improve the sound quality of the recovered instrumental backing track. Existing techniques for attempting recovery are discussed in the following section.

II Audio Inpainting

Audio Inpainting describes a family of techniques that attempt to estimate corrupted or missing values in audio signals [5]. This has numerous applications including estimating of missing portions of audio due to packet loss in audio streaming or recovery of portions of a signal corrupted due to clipping or impulse noise. Another application of audio inpainting occurs in the area of sound source separation using binary masking. This is the case with the techniques described in the introduction. Here, all time-frequency bins which are estimated as not belonging to the source to be separated are zeroed, even though in many cases, these bins may contain energy related to the source. This is as a result of estimation errors due to time-frequency overlap between sources.

The audio inpainting problem has been approached in a number of different ways, both in the time domain and the time-frequency domain. Time domain approaches include that of Adler et al, who split the signal into overlapping time domain frames [5]. The missing frames were estimated using a dictionary-based Orthogonal Matching Pursuit algorithm. With regards to time-frequency domain approaches, Smaragdis et al proposed a spectrogram factorisation-based approach [6], as did Le Roux et al, who use a convolutive factorisation model in conjunction with sparsity constraints to estimate the missing values [7]. A related approach, aimed at inpainting missing values in binary-masked spectrograms created using sound sound source separation algorithms was proposed in [8]. Here, non-negative matrix factorisation [9], with the addition of a sparsity constraint was used to estimate the missing values in the binary masked spectrogram. The algorithm was found to help recover missing harmonics in both vocals and pitched instruments, but was particularly useful at recovering transients associated with note onsets or drum instruments.

In the case featured in this paper, we are trying to create instrumental backing tracks by removing time-frequency bins associated with the position of the vocal in the stereo space. As noted previously, snare and kick drums as well as bass guitar typically occupy this space and so, even after using a high-pass filtering approach, large amounts of energy for these sources will still be missing. While this information can be recovered to a certain extent using the inpainting approaches described above, it comes at a price. Any traces of the vocals remaining in the backing track will be used as part of the inpainting process and so vocal energy will also be recovered as part of the inpainting process, resulting in increased vocal levels in the instrumental backing track. It can therefore be seen that a method capable of estimating the missing information in the backing track while still suppressing the vocals would be advantageous. Recently, a technique for vocal separation was proposed based on estimating the backing track [10]. It is proposed to investigate adapting this method for inpainting the missing backing track information. The vocal separation is described below in III, and its adaptation to inpainting described in IV.

III Vocal Separation using Nearest-Neighbours and Median Filtering

The vocal separation technique described in [10] assumes that the vocal is sparse in the time-frequency domain and that the instrumental backing track repeats more often than the vocal melody. In such a case, when calculating the distance between spectrogram frames, the effects of the backing track signal will predominate. The algorithm proceeds by calculating the squared Euclidean distance between all frames in the spectrogram:

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

where $\mathbf{X}$ is the mixture magnitude spectrogram of size $n \times m$ where n is the number of frequency bins and m is the number of time frames. Then $\mathbf{X}_k$ 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 complete set of distances is then stored in a symmetric matrix $\mathbf{D}$ of size $m \times m$.

$\mathbf{D}$ is then sorted in ascending order, and the p nearest neighbour frames to the kth frame are then obtained and stored in a $n \times p$ matrix $\mathbf{P}$. This matrix is then used to obtain an estimate of the backing track of the mixture signal at the kth frame. This is done by calculating the median value of $\mathbf{P}$ across all time frames:

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

where $\mathbf{Y}_k$ is the kth frame of the estimated background music spectrogram $\mathbf{Y}$, and where $\mathcal{M}$ denotes the median operator.The use of the median filter is motivated by the fact that when vocal energy appears in a given frequency bin, it will not follow the same pattern as the repeating background music. Therefore, frequency bins which contain vocal energy will appear as outliers which can be removed by median filtering. The estimated backing track was then further processed to eliminate bins which had energy greater than that of the original mixture signal:

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

where f denotes the fth frequency bin and k the kth time frame. This is based on the assumption that the backing track cannot have greater energy at a given time-frequency bin than the original mixture. The estimated backing track was then used to create a mask which allowed recovery of the separated vocal signal.

IV Median Nearest-Neighbour Audio Inpainting

While the nearest-neighbour median filtering algorithm has proven to be effective at separating vocals from single channel and stereo mixtures of vocals and background music, the separated backing track typically contains more artefacts than when the vocals are removed using pan-based techniques. This is as a result of the median filtering process which adversely affect the recovered sound quality. Nonetheless it can be seen that this algorithm provides an estimate of the background music in which the presence of vocals has been considerably reduced. It is proposed to adapt this algorithm for estimating missing information for inpainting when attempting to recover the backing track obtained from panning-based algorithms such as described in the introduction. In this case, the artefacts introduced due to median filtering should be considerably less audible than in the full vocal separation algorithm as we are only using the technique to infer the missing bins, rather than estimating all the bins in a frame.

Let $\hat{\mathbf{X}}$ denote the left backing track spectrogram obtained after binary masking of the original left channel spectrogram to eliminate the vocal region of the spectrogram. A distance matrix $\mathbf{D}$ is then calculated as per eqn. 2. This matrix is again sorted in ascending order and the p nearest neighbours to frame k gathered to form $\mathbf{P}$ as before. The kth frame of $\hat{\mathbf{X}}$ is then estimated as:

$$\hat{\mathbf{X}}_{f,k} = \begin{cases} \hat{\mathbf{X}}_{f,k} & \text{if } \hat{\mathbf{X}}_{f,k} > 0 \\ \mathcal{M}(\mathbf{P} > 0) & \text{if } \hat{\mathbf{X}}_{f,k} = 0 \end{cases} \tag{4}$$

Here, only the non-zero values of $\mathbf{P}$ are used to calculate the median, as the zero-valued bins contain no information about the backing track to be estimated. All values of $\hat{\mathbf{X}}_{f,k}$ which are greater than the original mixture spectrogram value are replaced following eqn. 3. The process is then repeated for the right backing track spectrogram,after which both channels are inverted to the time domain by performing an inverse STFT.

The advantage of this approach over other inpainting techniques is that in this case, vocal suppression is built into the inpainting technique. This means that the inpainting technique will only recover the missing portions of the backing track such as the bass guitar, snare drum and kick drum.

V Inpainting Examples

To illustrate the utility of the proposed inpainting algorithm, an excerpt was taken from a real world recording. Here, bass guitar, snare drum and kick drum were positioned in the centre of the stereo space, along with the vocals. Figure 1 shows the spectrogram of the left channel from 5000Hz to 15000Hz. This range was chosen so that the effects of the inpainting using nearest neighbours and median filtering could be seen clearly in the figures. The drums can clearly be seen as a set of vertical ridges, while the upper harmonics of the vocals are visible as a set of wavy lines, most clearly seen in the 5000-8000 Hz frequency range.

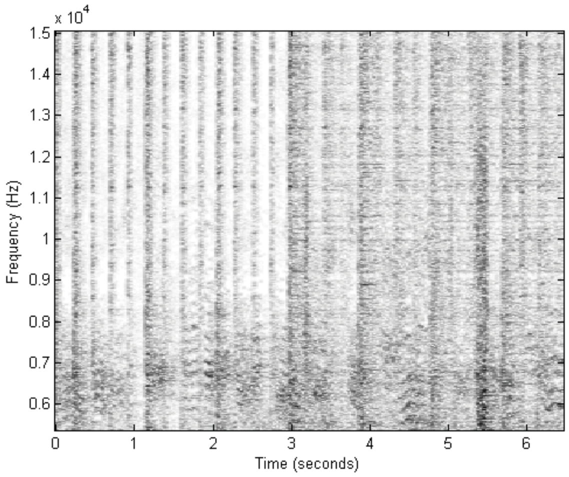

Fig. 1: Spectrogram of left channel of mixture signal

The stereo excerpt was then passed through the Adress algorithm to remove the vocals, along with the other instruments occupying that position in the stereo field. Figure 2 shows the resulting spectrogram after bins associated with the centre stereo position have been removed. It can be seen that the transients associated with the drums have been considerably degraded, with whole regions of the spectrogram associated with the drums showing no energy. This has audible effects, with the drums being considerably reduced in volume, and the sharpness of the transients associated with the drum onsets has been lost. Further, it can be seen that the vocals have been considerably reduced in volume, with the wavy lines associated with the vocals being difficult to see in the spectrogram.

Figure 3 then shows the spectrogram obtained after performing audio inpainting using the nearest-neighbour median filtering approach. It can clearly be seen that energy has been restored to the drums after the inpainting, with the visible gaps in the spectrogram having been filled in. Further, there is still little or no trace of the wavy vocal lines in the inpainted spectrogram. On listening to the reconstructed signal, the drums are louder than previously, though still not at loud as in the original mixture spectrogram, while traces of the vocals are no louder than in the spectrogram obtained via Adress. This demonstrates the utility of the proposed inpainting approach for the creation of instrumental backing tracks, with the other instruments being restored without increasing the level of the vocals.

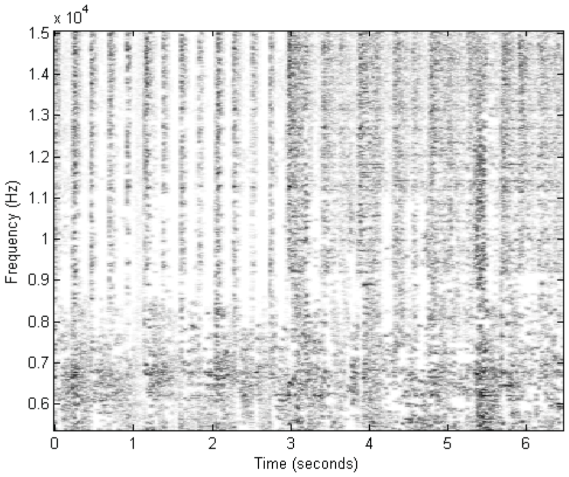

Fig. 2: Spectrogram of left channel after center position removal using Adress

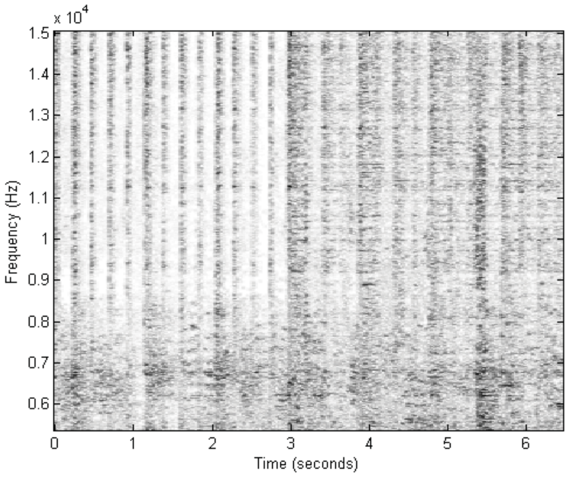

Fig. 3: Spectrogram of left channel after inpainting using nearest neighbours and median filtering

Audio examples related to the inpainting algorithm can be found at [11].

VI Conclusions

A novel inpainting algorithm has been proposed based on nearest-neighbours and median filtering. The algorithm is designed specifically for the case where an instrumental backing track has been created by removing time-frequency bins associated with the region in stereo space where the vocals are present. As noted previously, this region will often contain other instruments such as bass guitar and drums. Standard audio inpainting techniques can recover the missing values, but at the expense of inpainting of missing vocals too. The proposed algorithm overcomes this by using a technique which allows estimation of the other instruments while still suppressing the vocals. The utility of this technique has been demonstrated on real-world examples, and has also been shown to recover improved vocal separations as well.

VII Acknowledgements

Derry FitzGerald was supported in this research by the Science Foundation Ireland Stokes Lectureship programme.

References

[1] D. Barry, E. Coyle, and B. Lawlor, “Sound Source Separation: Azimuth Discrimination and Resynthesis”, *Proc. 7th International Conference on Digital Audio Effects*, 2004

[2] O. Yilmaz and S. Rickard, “Blind Separation of Speech Mixtures via Time-Frequency Masking”, *IEEE Transactions on Signal Processing*, Vol. 52, No. 7, 2004.

[3] C. Avendano “Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppression and re-panning applications”, *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA2003)*, pages 55 58, October 2003.

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

[5] A. Adler, V. Emiya, M. Jafari, M. Elad, R. Gribonval, and M. Plumbley. “Audio Inpainting”. *IEEE Trans. on Audio Speech and Signal Processing*, Vol. 20, No. 3, 2012.

[6] P. Smaragdis, B. Raj, and M. Shashanka. “Missing Data Imputation for Time-Frequency Representations of Audio Signals”. *Journal of Signal Processing Systems*, 2010.

[7] J. Le Roux, H. Kameoka, N. Ono, A. de Cheveigne, S. Sagayama. “Computational auditory induction as a missing-data model-fitting problem with Bregman divergence”. *Speech Communication*, 53 (2011) 658-676.

[8] D. FitzGerald and D. Barry “On Inpainting the Adress Algorithm”. *23nd IET Irish Signals and Systems Conference*,NUI Maynooth, 2012.

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

[10] D. FitzGerald “Vocal Separation Using Nearest Neighbours and Median Filtering”. 23nd IET Irish Signals and Systems Conference,NUI Maynooth, 2012.

[11] Inpainting Examples `http://eleceng.dit.ie/derryfitzgerald/index.php?uid=489&menu_id=71`.