2011-7

Upmixing from Mono : a Source Separation Approach

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

UPMIXING FROM MONO - A SOURCE SEPARATION APPROACH

Derry FitzGerald

Audio Research Group
Dublin Institute of Technology
Kevin St., Dublin 2, Ireland

ABSTRACT

We present a system for upmixing mono recordings to stereo through the use of sound source separation techniques. The use of sound source separation has the advantage of allowing sources to be placed at distinct points in the stereo field, resulting in more natural sounding upmixes. The system separates an input signal into a number of sources, which can then be imported into a digital audio workstation for upmixing to stereo. Considerations to be taken into account when upmixing are discussed, and a brief overview of the various sound source separation techniques used in the system are given. The effectiveness of the proposed system is then demonstrated on real-world mono recordings.

Index Terms— Upmixing, Sound source separation, vocal separation, percussion separation, pitched instrument separation, Non-negative Tensor Factorisation, Non-negative partial cofactorisation.

1. INTRODUCTION

Since the first issuing of commercial stereo gramophone recordings in the late 1950s, there have been numerous attempts to create stereo recordings from material originally issued as monophonic or single channel recordings. This typically involved taking two copies of the signal, and delaying one copy by 30-40 ms, and then high-pass filtering one copy and low-pass filtering the other copy. This was the basis of the Duophonic system used by Capitol Records to create pseudo-stereo recordings. Alternatively, approaches based on comb-filtering have been proposed [1],[2].

However, a notable problem with these recordings is that while they can give the appearance of spread or width to a monophonic recording, they do not allow for the placement of individual sources in the original recording at distinct points in the stereo field. The ability to do this would result in a much more natural sounding conversion or upmixing of mono recordings to stereo, or indeed 5.1 surround sound.

To this end, it is proposed to use sound source separation techniques to separate out sources or instruments from a mono recording. The separated source signals can then be used to create a stereo upmix of the mono recording. Sound source separation techniques have previously been used for the creation of stereo to 5.1 upmixes [3], but little or no work has been done on mono to stereo. The state of the art in single channel sound source separation has advanced considerably in recent years, with large amounts of work focusing on approaches using spectrogram factorisation in particular. Nevertheless, given the difficulty of the problem, there will typically still be imperfections in the separation, both in terms of residual traces of other sources, and in artifacts due to the separation process. This can be a limiting factor when attempting to use these separations in the context of new musical pieces, but, as will be discussed later, this is not as much of a problem when using the separated sources for the purposes of upmixing from mono to stereo.

The focus in this paper is on the use of blind source separation techniques to separate the signals for upmixing, where there is little or no information provided to the system about the nature of the sources to be separated. It is felt that this will result in a more general system which is capable of dealing with very different types of music, without recourse to optimising the techniques for each style.

There are a number of considerations to be taken into account when attempting to create a successful upmix from mono to stereo when using sound source separation techniques. Firstly, the upmix should be free from any audible artifacts. This can be achieved by ensuring that no information from the original signal is lost at any point in separating the sources, so that the separated signals summed together fully reconstitute the original mono signal. In this case, any artifacts in the individual separations will be usually be masked by the other sources, provided that the chosen pan positions for the separated sources are not too extreme. This has ramifications for the stereo width of the upmix, and is discussed in greater detail in section 3.

Secondly, another important consideration is the stability of the pan position of the separated sources. It is desirable to create upmixes in which the pan position of the sources remains fixed throughout their appearance in the piece. If the sources drift gradually out of position or occasionally jump position, then this can be distracting for the listener, particu-

This research was funded under the Science Foundation Ireland Stokes Lectureship scheme

larly for those using headphones.

The principal reason for a source to move position is due to incorrect separation of the sources, where, for example, part of a vocal track has not been correctly separated from the percussion instruments and so where the incorrect separation occurs, the vocal moves towards the position of the percussion instruments. This would be particularly noticeable if the sources were on opposite sides of the stereo field, such as the vocals panned hard right and the percussion instruments panned hard left. In cases such as this, the easiest way to ameliorate the problem is either to put both sources in the same position, or to put the sources in positions close to each other, such as putting the drums hard left and the vocals mid-left. This again can impose limitations on the positioning of sources in the stereo upmix.

Taking these considerations into account, it can be seen that the separation of the various sources does not have to be perfect in order to achieve a realistic upmix from mono to stereo. Due to the effects of masking, most of the issues relating to the residual presence of other sources and artifacts due to separation can be overcome if due care is taken when reconstructing the sources. Therefore, the principal criteria for achieving a realistic upmix is that the sources are sufficiently separated to allow directionality to be imposed on the sources. This is a considerably less onerous requirement than achieving separations with minimal artifacts and bleed from other sources.

Finally, there is the issue as to whether to aim to recreate the original mono mix, but with enhanced width, or whether to attempt to remix the material by increasing or decreasing the gains of the individual sources. The approach taken here is to solely enhance width, while leaving the overall balance of the sources the same. It is felt that this results in upmixes which are more faithful to the original source material, though arguments could be made for remixing to reflect the tastes of modern listeners.

2. UPMIXING SYSTEM OVERVIEW

In recent years much research has been carried out on the topic of single channel sound source separation, which concerns itself with the extraction of individual sources or instruments from a single channel mixture of instruments. A wide range of different techniques have been used in attempting to solve this problem, some involving the use of training data to generate models of the sources to be separated, some using other prior knowledge such as pitch information about the sources, and others which attempt to perform the separations in a blind manner without any prior knowledge about the sources. However, for the purposes of this paper, this research on single channel separation can be broken into three broad categories, distinguished by the types of sources to be separated. These are outlined below.

The first category is research related to the separation of pitched instruments from other pitched instruments. In the context of upmixing, it is clear that this is a necessary part of any upmixing system, as it will allow pitched sources in a mixture to be placed at different points in the stereo field, such as in a recording containing piano, guitar and bass guitar. Techiques used for this purpose include techniques based on non-negative matrix and tensor factorisations [4, 5, 6], and others based on common trajectories of harmonics [7, 8].

The second category attempts to extract percussion instruments occuring in mixtures of other instruments including both pitched instruments and voice. This is necessary for upmixing in order to allow the placement of percussion sources at points in the stereo field. Again, a number of techniques used here are based on non-negative matrix and tensor factorisations [9, 10, 11], while others use template matching and adaptation [12, 13]. Other techniques make use of simple heuristics based on characteristics of both pitched and harmonic instruments [14, 15].

The third category deals with the problem of extracting vocals from mixtures of instruments including both pitched instruments and percussion instruments. Again, it is evident that a requirement of any upmixing system based on sound source separation requires the ability to place the vocal at a given place in the stereo field. Techniques used here include Bayesian approaches and matrix factorisation techniques [16, 17, 18, 19].

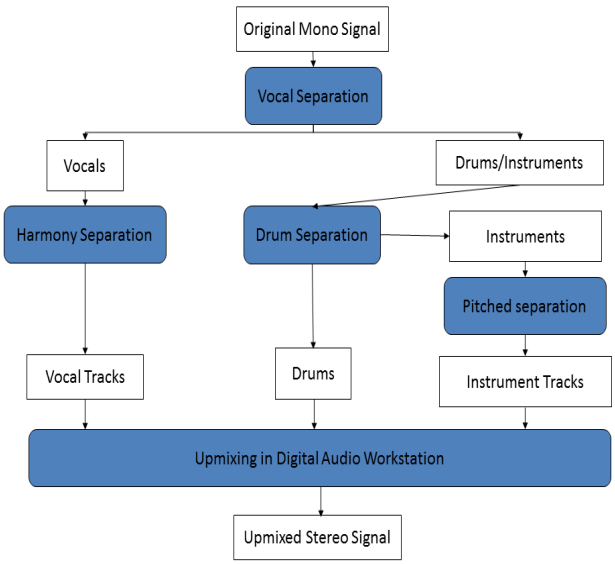

Fig. 1. *Upmixing system flowchart*

In order to create a separation-based upmixing system, techniques and algorithms from all of these categories will need to be used. The upmixing system proposed takes methods from all of these categories and applies them successively to generate the separated source signals. Figure 1 outlines the

order in which the algorithms are applied to decompose the original signal. Firstly, a vocal separation algorithm is employed to separate the vocals from the other instrumentation in the track. Where vocal harmonies are present the vocals can be further processed using the pitch separation algorithm to separate to some extent the harmonies and add stereo width.

The signal containing the remaining instrumentation, including both pitched and percussion instruments will then be split into separate percussion and pitched instrument signals, through use of techniques from the second category. An optional stage, depending on the upmixing requirements includes the further decomposition of the percussion signal into sources containing separated percussion instruments, to allow further flexibility in the positioning of the percussion sources. Finally, the pitched instruments are further decomposed to separate the pitched instruments from each other. The proposed system is quite computationally intensive and it can take between a half an hour to an hour to process a 3 minute pop song. Upon completion of the separation process, the separated source signals can then be imported into a digital multitrack editor to create a stereo upmix of the original material by panning the separated signals to various points in the stereo field.

The following sections deal with details of the upmixing separation system. Section 3 discusses how the sources are reconstructed as this is common to all the source separation techniques used. Then the following three sections describe the algorithms deployed for vocal separation, percussion separation and pitched instrument separation. However, the techniques will not be presented in the order shown in Figure 1, as the technique used for vocal separation makes use of the algorithms related to separating pitched and percussion sources, as well as that for separating pitched sources. Therefore, it is necessary to explain both of these algorithms in brief before proceeding to describe the vocal separation algorithm.

3. SIGNAL RECONSTRUCTION

The upmixing system makes use of a number of source separation algorithms, based on widely different techniques to estimate separated source spectrograms. Regardless of the separation method, the same method is used to reconstruct the separated source signals. Rather than apply the original phase information to the separated mixture spectrograms, these spectrograms are instead used to create masks which are applied to the original complex valued spectrogram $\mathbf{Y}$ generated from the input signal:

$$\mathbf{Y}_k = \mathbf{Y} \otimes \frac{\mathbf{Q}_k^2}{\mathbf{Q}_1^2 + \cdots + \mathbf{Q}_z^2} \tag{1}$$

where $\mathbf{Y}_k$ is the complex-valued spectrogram of the kth of z sources, $\mathbf{Q}_k$ is the estimated magnitude spectrogram of the kth source, $\otimes$ denotes elementwise multiplication, and all other operations are also performed elementwise. In effect, these masks perform a version of Wiener filtering on the input signal.

There are a number of reasons for the use of the above approach. Firstly, the above approach has been observed to give more natural sounding separations than applying the original phase to the estimated spectrograms. Secondly, the above approach ensures that the separated signals sum together to yield the original input signal. Given the multipass nature of the upmixing process, where separated signals are typically passed through another separation stage, the Wiener filtering approach insures that no information from the original signal is lost at any point. The final set of separated signals obtained from the various stages will sum together to yield the original input signal. This has important effects on the perceptual quality of the final upmix created from the separated signals.

Regardless of the techniques used, there will be artifacts in the separated signal. These will often be quite noticeable when the separated signals are played in isolation. However, if care is taken in the upmix, when played in conjunction with the other separated signals, the human auditory system will reintegrate the artifacts to their correct sources, yielding a stereo signal where no artifacts are audible.

However, it should be noted that in certain situations, if the chosen source position is too extreme, such as panning hard left or hard right, artifacts will become noticeable in the separation. This is because the positioning is too extreme for the human auditory system to successfully reintegrate the artifacts present in the other sources with the correct source. This varies from source to source, depending on the separation quality, and can sometimes result in limitations on the stereo width of the upmix for certain sources. The audible presence of artifacts can usually be avoided by panning the desired source as far as possible towards the desired location while listening to ensure no artifacts can be heard. For sources panned to center positions this is not an issue, unless significant gain changes are involved.

Another by-product of the reconstruction technique is the fact that it ensures that there will be no phase issues between the recovered signals. This means that the effect of source panning is solely created using differences in intensity. It should be noted that this also applies to using the original mixture phase with the separated source spectrograms. The lack of phase issues also ensures that the upmixes are fully mono compatible, i.e a mono-ed version of the upmix is equivalent to the original mono signal.

4. DRUM SEPARATION USING MEDIAN FILTERING

Drum separation is performed using a median filtering based approach [20]. The underlying idea behind this method is that percussion instruments can be regarded as forming vertical ridges in spectrograms, while the harmonics of pitched instruments form horizontal ridges in the spectrogram. In the

case of vertical ridges associated with percussion instruments, peaks associated with pitched instruments can be regarded as outliers. Therefore, removing these outliers will reduce the presence of pitched instruments from the spectrogram. Similarly, the transients associated with the onset of percussion instruments will be outliers in the horizontal ridges associated with the pitched instruments and removing these outliers will reduce the effects of percussion instruments in the spectrogram.

A simple technique for the removal of outliers is the use of median filtering, where a given sample is replaced by the median of the values taken from a window around the sample. Where $x(n)$ is the input vector, then the output after median filtering $y(n)$ is defined as:

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

where l is the length of the median filter, and l and is odd. For even lengths the mean of the two values in the sorted list is used. For an input magnitude spectrogram $\mathbf{X}$, let X_i denote the ith time frame and X_h denote the hth frequency slice containing the values of the hth frequency bin across time. Then, percussion enhanced frames and harmonic enhanced slices can be obtained from:

$$P_i = \mathsf{M}\{X_i, l_{perc}\} \quad (3)$$

$$H_h = \mathsf{M}\{X_h, l_{harm}\} \quad (4)$$

where M denotes median filtering,and where l_{perc} and l_{harm} are the length of the percussion-enhancing and harmonic-enhancing median filters respectively. The percussion enhanced frames P_i are then combined to yield a percussion enhanced spectrogram $\mathbf{P}$. Similarly the frequency slices H_h are combined to yield a harmonic-enhanced spectrogram $\mathbf{H}$. These spectrograms can then be used to generate masks which are applied to the original complex valued spectrogram before inversion to the time domain to recover the separated sources in the manner described in Section 3

5. PITCHED INSTRUMENT SEPARATION

The algorithm for pitched instrument separation is based on the tensor factorisation approaches described in [21] and [22]. This is a general separation algorithm, capable of separating pitched and percussive instruments from n-channel mixtures though it is not effective at separating vocals from pitched instruments. Here we deal with a slightly simplified single channel case of these algorithms, where shifts in time have been eliminated from the model.

In the following, tensors are denoted by upper-case calligraphic letters such as $\mathcal{A}$. Contracted tensor multiplication is denoted by $\langle \mathcal{AB} \rangle_{\{a,b\}}$ where $\mathcal{A}$ and $\mathcal{B}$ are the tensors to be multiplied. Here, contracted tensor multiplication takes place on the dimensions a and b of $\mathcal{A}$ and $\mathcal{B}$ respectively. Indexing of elements within a tensor is notated by $\mathcal{A}(i,j)$ as opposed to using subscripts, following the conventions used in the Tensor Toolbox for Matlab, which was used in the implementation of the algorithm [23]. Given an input spectrogram $\mathbf{X}$ of size $n \times m$, then $\mathbf{X}$ is modelled as:

$$\mathbf{X} \approx \hat{\mathbf{X}} = \sum_{k=1}^{K} \langle \langle \langle \mathcal{FH} \rangle_{\{2,1\}} \mathcal{W} \rangle_{\{3,1\}} \mathcal{S} \rangle_{\{2,1\}} + \sum_{l=1}^{L} \langle \mathcal{BC} \rangle_{\{2,1\}} \quad (5)$$

The first right-hand side term models pitched instruments, and the second unpitched or percussion instruments. K denotes the number of pitched instruments and L denotes the number of unpitched instruments. As all tensors are instrument-specific, for ease of notation, the subscripts k and l are implicit in all tensors within the respective summations.

The pitched instrument model is a source-filter model where each source is modelled as a weighted sum of sinusoids. These are then filtered by a filter which attempts to mimic the the formant structure of the instrument. Here the formant filter is modelled by $\mathcal{F}$, an $n \times n$ diagonal matrix. $\mathcal{H}$ contains a dictionary of sinusoids, where $\mathcal{H}(:,i,j)$ contains the frequency spectrum of a sinusoid with frequency equal to the jth harmonic of the ith note of the instrument. $\mathcal{H}$ 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. $\mathcal{W}$ contains a set of weights which describe how the sinusoids which make up a note played by an instrument are weighted to approximate the instrument timbre and is of size h_k. $\mathcal{S}$ is a tensor of size $z_k \times m$ which contains the activations of the z_k notes played by the instrument.

The unpitched instruments are modelled as the product of a source frequency basis function $\mathcal{B}$ of size $n \times 1$ with a time activation basis function $\mathcal{C}$ of size $1 \times m$. For separation of pitched instruments only, this part can obviously be eliminated from consideration. However, as this part of the model is used as part of the vocal separation algorithm, it is included here for completeness.

Multiplicative update equations can be derived using the generalised Kullback-Liebler divergence as a cost function in a manner similar to that shown in [21]. The individual source spectrograms can then be recovered by combining the tensors associated with the kth source to yield a source spectrogram $\mathbf{X}_k$. These source spectrograms are then used to create Wiener filters which are applied to the original complex-valued spectrogram in the manner described in section 3.

6. VOCAL SEPARATION

The vocal separation algorithm used is the algorithm described in [24]. This algorithm makes use of the drum

separation technique described in section 4 and the tensor factorisation technique described in section 5, as well as a further cofactorisation stage. The vocal separation algorithm takes advantage of the fact that for magnitude spectrograms obtained from large window sizes with high frequency resolution, vocals will appear as locally broadband noise, while for low frequency resolution the vocals will appear as harmonic in nature. This is in contrast to percussion instruments which will appear as broadband noise regardless of what frequency resolution is used, and to pitched instruments which will appear as harmonic regardless of the frequency resolution used.

It follows then that using the median filter based separation algorithm at high frequency resolution, such as with an FFT size of 16384 samples, on a signal containing drums, vocals and pitched instruments, will result in two signals, one containing mainly drums and vocals, and another containing mainly pitched instruments. The drums and vocals signal can then again be processed by the median filtering separation algorithm, this time at low frequency resolution, yielding a separated drum signal and a separated vocal signal. However, for best results, it was observed that using a Constant Q transform [25] typically gave better separations than using a low frequency resolution STFT.

After this vocal separation has been obtained, there will still be artifacts in the separated vocal, In the case of a typical pop song this would include elements of the drums and often some trace of the bass guitar. To reduce the effects of these artifacts, the separated vocal is then passed through the tensor factorisation based algorithm described in section 5 resulting in an improved vocal separation.

Further improvements can be obtained by using this separated vocal to perform a partial cofactorisation on the original input signal. Partial non-negative matrix cofactorisation was originally proposed as a means of separating drums from mixtures of pitched instruments [26] and was adapted in [24] for the purposes of vocal separation. Here, the original mixture spectrogram and a spectrogram of the separated vocal were decomposed simultaneously, while sharing some frequency basis functions between the two spectrograms, thereby forcing some basis functions to be associated with vocals only, while allowing the remaining basis functions to adapt to other sources in the mixture. This model can be expressed as:

$$\hat{\mathbf{X}} = \mathbf{A}_T\mathbf{S}_T + \mathbf{A}_V\mathbf{S}_V \tag{6}$$

and

$$\hat{\mathbf{D}} = \mathbf{A}_V\mathbf{S}_{V1} \tag{7}$$

where $\mathbf{X}$ is the mixture spectrogram and $\mathbf{D}$ is the separated vocal spectrogram. $\mathbf{A}_T$ and $\mathbf{S}_T$ contain the frequency and time basis functions for the instrumental track containing both pitched instruments and percussion. $\mathbf{A}_V$ then 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 $\mathbf{D}$ respectively. Then, again using the generalised Kullback-Liebler divergence as a cost function, update equations can be derived for the model parameters. The resulting vocal separation and separated backing track obtained using this technique typically contains less artifacts than the original vocal separation used as input.

At each stage in the vocal separation process, the Wiener filtering technique is again used to generate the source signals. Once the final separated vocal has been obtained, the separated instrumental track is then separated using firstly the drum separation algorithm and then the individual instruments are separated using the tensor factorisation algorithm, yielding the separated sources for use for upmixing.

7. UPMIXING EXAMPLES

Having described the various sound source separation techniques used in the upmixing separation system, we now present real-world examples of upmixes created using the separations obtained from the system. The separated signals were imported into a digital audio workstation, and were panned to various positions to create the stereo upmix.

Figure 2 shows the spectrogram of an excerpt taken from the original mono recording of "Good Vibrations" by the Beach Boys. Here the instrumentation consists of piano, mixed percussion, theremin and bass guitar. Figure 3 shows the spectrogram of the left channel of the stereo upmix, while figure 4 shows the spectrogram of the right channel of the upmix. In this case, the piano was panned to a mid-left position, while the percussion was panned hard right, the theremin to mid right and the bass guitar to the centre of the stereo field.

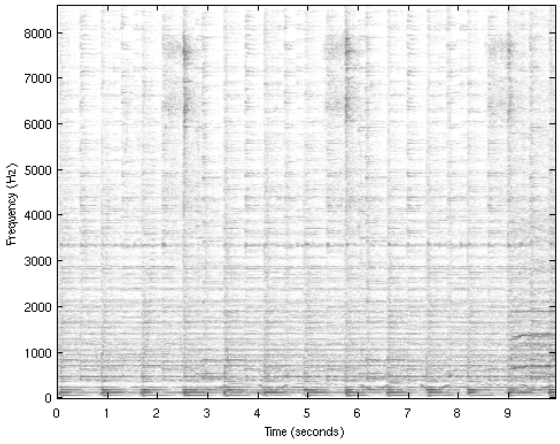

Fig. 2. *Spectrogram of an excerpt from "Good Vibrations" - Original Mono Recording. Instrumentation includes piano, mixed percussion, theremin and bass guitar.*

It can be seen that the presence of percussion is very much

reduced in the left channel spectrogram in comparison with the original mono signal, while the presence of the percussion can be clearly seen in the right channel. The harmonics of the piano can clearly be seen in both left and right spectrograms, though it is clearly louder in the left channel than the right. The theremin is visible as a modulating sinusoid near the bottom right corner of both the original spectrogram and the right channel of the upmix, and indeed can be more clearly seen in the right channel of the upmix than in the original mono mix. The low frequency energy of the bass guitar can be seen to be equally loud in both left and right spectrograms of the upmix.

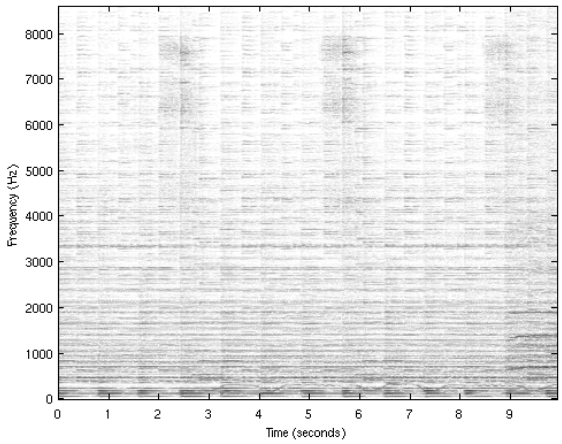

Fig. 3. *Spectrogram of an excerpt from "Good Vibrations" - Stereo Upmix Left Channel. The piano and bass are the predominant instruments in this channel*

On listening to the stereo upmix, a wide stereo image can be heard. This is particularly noticeable when swapping between the original mono mix and the stereo upmix. No artifacts can be heard in the created upmix, and the sources can clearly be identified as having come from a given position in the stereo field. The stability of the positioning of the sources is quite good throughout the upmix, with little or no source drift evident upon listening. This demonstrates the potential of the upmixing system in creating realistic stereo upmixes from mono real-world recordings. A longer version of the excerpt used in these figures is available for listening at `http://eleceng.dit.ie/derryfitzgerald/index.php?uid=489&menu_id=54`. Other upmixing examples are also available for listening from this page.

8. CONCLUSIONS

Having outlined the reasons for using sound source separation for the purposes of upmixing from mono to stereo, we then dealt with considerations that need to be taken into account in order to achieve successful upmixing using sound source separation. Following from this, an overview of the

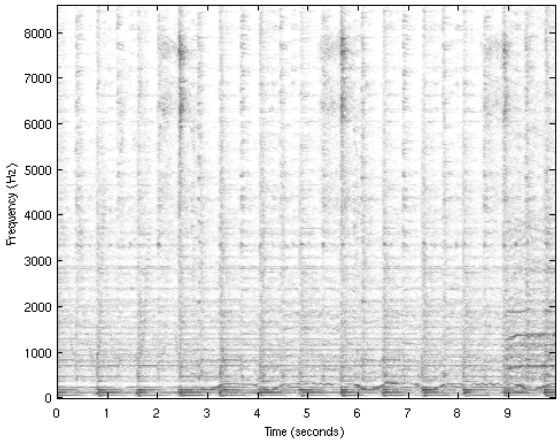

Fig. 4. *Spectrogram of an excerpt from "Good Vibrations" - Stereo Upmix Right Channel. The percussion, theremin and bass are the predominant instruments in this channel.*

upmixing separation system was presented, and a description of the signal reconstruction method presented. The various sound separation technologies used in the system were then briefly discussed. Finally the effectiveness of the system for upmixing was demonstrated on real-world mono recordings.

Future work will concentrate on improving the quality of the separations obtained from the various algorithms, particularly with a view to the considerations highlighted in this paper. Work will also be carried out in performing listening tests on upmixes obtained from a mono version of a multitrack recording, in comparison with actual stereo mixes made from the multitrack. It is hoped that this will allow quantification of the performance of the upmixing system against actual stereo mixes.

9. REFERENCES

[1] M. Schroder, "An artificial stereophonic effect obtained from a single audio signal," *Journal of the Audio Engineering Society*, vol. Vol.6, no.2, pp. 74–80, 1958.

[2] R. Orban, "A rational technique for synthesizing pseudo-stereo from monophonic sources," *Journal of the Audio Engineering Society*, vol. Vol.18, pp. 157–165, 1970.

[3] D. Barry and G. Kearney, "Localization quality assessment in source separation-based upmixing algorithms," in *AES 35th International Conference*, London, 2009.

[4] E. Vincent and X. Rodet, "Underdetermined source separation with structured source priors," in *5th International Conference on Independent Component Analysis and Blind Signal Separation*, 2004.

[5] T. Virtanen, "Separation of sound sources by convolutive sparse coding," in *Proc. of ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing*, 2004.

[6] P. Smaragdis, "Discovering auditory objects through non-negativity constraints," in *Proc. of ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing*, 2004.

[7] Zhiyao Duan, Yungang Zhang, Changshui Zhang, and Zhenwei Shi, "Unsupervised single-channel music source separation by average harmonic structure modeling," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 4, pp. 766 –778, May 2008.

[8] T. Virtanen, *Sound Source Separation in Monaural Music Signals*, Ph.D. thesis, Tampere University of Technology, Tampere, Finland, 2006.

[9] O. Gillet and G. Richard, "Transcription and separation of drum signals from polyphonic music," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 3, pp. 529–540, 2008.

[10] M. Helen and T. Virtanen, "Separation of drums from polyphonic music using non-negative matrix factorisation and support vector machine," in *Proc. European Signal Processing Conference*, Anatalya, Turkey, 2005.

[11] D. FitzGerald, E. Coyle, and M. Cranitch, "Using tensor factorisation models to separate drums from polyphonic music," in *Proc.of 12th International Conference on Digital Audio Effects, (DAFX09)*, Como, Italy, Sept. 2009.

[12] K. Yoshii, M. Goto, and H. Okuno, "Drum sound recognition for polyphonic audio signals by adaptation and matching of spectrogram templates with harmonic structure suppression," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 333–345, 2007.

[13] K. Yoshii, M. Goto, and H. Okuno, "Inter:d a drum sound equaliser for controlling volume and timbre of drums," in *Proc. European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies*, 2005.

[14] D. Barry, D. FitzGerald, and E. Coyle, "Drum source separation using percussive feature detection and spectral modulation," in *Proc. IEE Irish Signals and Systems Conference*, 2005.

[15] 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 16th European Signal Processing Conference*, 2008.

[16] 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.

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

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

[19] 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.

[20] D. FitzGerald, "Harmonic/percussive separation using median filtering," in *Proc.of 13th International Conference on Digital Audio Effects, (DAFX10)*, Graz,Austria, Sept. 2010.

[21] D. FitzGerald, M. Cranich, and E. Coyle, "Extended non-negative tensor factorisation models for musical sound source separation," *Computational Intelligence and Neuroscience, vol. 2008, Article ID 872425*, 2008.

[22] D. FitzGerald, M. Cranich, and E. Coyle, "Musical source separation using generalised non-negative tensor factorisation models," in *Workshop on Music and Machine Learning, International Conference on Machine Learning*, Helsinki.

[23] B. Bader and T. Kolda, "Matlab tensor toolbox version 2.2," 2007.

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

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

[26] J. Yoo, M. Kim, K. Kang, and S. Choi, "Nonnegative matrix partial co-factorization for drum source separation," in *Proc. of the IEEE Conference on Acoustics, Speech, and Signal Processing*, 2010.