Drum Transcription in the presence of pitched instruments using Prior Subspace Analysis

Derry FitzGerald[ϕ], Bob Lawlor*, and Eugene Coyle[ϕ]

[ϕ]Music Technology Centre,
Dublin Institute of Technology,
Rathmines Road, Dublin,
IRELAND
E-mail:[ϕ]derry.fitzgerald@dit.ie ,eugene.coyle@dit.ie

** Department of Electronic Engineering,*
National University of Ireland,
Maynooth,
IRELAND
*E-mail: *rlawlor@eeng.may.ie*

Abstract -- **This paper demonstrates the use of Prior Subspace Analysis (PSA) as a method for transcribing drums in the presence of pitched instruments. PSA uses prior subspaces that represent the sources to be transcribed to overcome some of the problems associated with other subspace methods such as Independent Subspace Analysis (ISA) or sub-band ISA. The use of prior knowledge results in improved robustness for transcription purposes and enables the method to work more readily in the presence of pitched instruments than other subspace methods. The system presented in this paper attempts to extend the use of PSA to transcribe drum sounds in the presence of interfering pitched instruments.**

I INTRODUCTION

In the past few years a number of subspace methods such as Independent Subspace Analysis (ISA) and Prior Subspace Analysis (PSA) have been proposed for sound source separation in single channel mixtures [1,2]. The underlying assumptions of these methods make them particularly suited to attempting the task of transcribing drums from single channel audio mixtures, as has been shown in [2,3]. However these methods have dealt solely with the case where drums only are present. This system presented in this paper attempts to extend the PSA method to transcribe drums robustly in the presence of pitched instruments.

II SUBSPACE METHODS FOR SOUND SOURCE SEPARATION

The subspace methods described in [1,2,3] attempt to represent sound sources as low dimensional independent subspaces in the time-frequency plane.

These methods make a number of assumptions about the signal. An input signal containing a number of sound sources is transformed to a time-frequency representation such as a magnitude spectrogram. It is assumed that the mixture signal spectrogram **Y** can be decomposed into *l* statistically independent spectrograms Y_j. These spectrograms are assumed to be represented by the outer product of an invariant frequency basis function f_j, and a corresponding invariant amplitude basis function t_j which describes the variations in amplitude of the frequency basis function over time. This yields:

$$\mathbf{Y} = \sum_{j=1}^{l} Y_j = \sum_{j=1}^{l} f_j t_j^T \qquad (1)$$

These independent basis functions represent features of the individual sources. Each source is made up of a number of these basis functions which form a low dimensional subspace that represents the sound source.

Where the subspace methods differ is in how decomposition of the original mixture spectrogram **Y** into outer product basis functions is achieved. ISA achieves the decomposition by performing Principal Component Analysis on the mixture spectrogram. Components of low variance are then discarded to achieve low dimensionality. Independent Component Analysis (ICA) [4] is then performed on the remaining components to obtain independent subspaces. The above decomposition is performed in a totally blind manner and makes no use of information about sources known to be present in the

mixture. A detailed description of the above decomposition can be found in [1].

Though an effective means of separating sound mixtures there are significant limitations to the ISA method. Firstly the assumption that the basis functions are invariant means no pitch changes are allowed in the overall spectrogram. However this is not a problem when dealing with sources that can be considered stationary in pitch such as drum sounds, making ISA suited to dealing with drum sounds.

Secondly estimating the number of components to retain from PCA remains a problem. The number of components required for separation varies with the frequency and amplitude characteristics of the source sounds, and the threshold method proposed in [1] cannot adequately predict the required number of components. This results in the necessity of an observer to decide the number of components to retain. An attempt to overcome this problem by means of sub-band preprocessing is described in [3].

Despite these limitations ISA provides a method of overcoming the problem of identifying mixtures of drums encountered by Sillanpää et al when trying to identify and transcribe mixtures of drums [5].

On the other hand PSA assumes that there exists known prior frequency basis functions f_p that are good initial approximations to the actual basis functions of the sources of interest. Substituting these f_p for the f_j in equation 1 yields:

$$\mathbf{Y} \approx \sum_{j=1}^{l} f_p t_j^T \qquad (2)$$

Multiplying the overall spectrogram $\mathbf{Y}$ by the pseudoinverse of the prior frequency subspaces yields estimates of the amplitude basis functions, $\hat{\mathbf{t}}$:

$$\hat{\mathbf{t}} = \mathbf{f}_{\mathbf{pp}} \mathbf{Y} \qquad (3)$$

where $\mathbf{f}_{\mathbf{pp}}$ is the pseudoinverse of $\mathbf{f}_{\mathbf{p}}$. However the amplitude basis functions returned are not independent and so ICA is carried out on $\hat{\mathbf{t}}$ to give

$$\mathbf{t} = \mathbf{W}\hat{\mathbf{t}} \qquad (4)$$

where $\mathbf{W}$ is the unmixing matrix obtained using ICA. Improved estimates of the frequency basis functions can then be obtained from

$$\mathbf{f} = \left(\mathbf{Y}\mathbf{t}_{\mathbf{p}}\right)^{\mathbf{T}} \qquad (5)$$

PSA uses prior knowledge to obtain the most important information specifically on the sources of interest and so overcomes the problem of estimating the amount of information needed for separation that is associated with ISA. PSA also relaxes the assumption that no pitch changes are allowed in the overall spectrogram. Instead it assumes that only the sources of interest are stationary in pitch. This makes PSA suitable for attempting to transcribe drums in the presence of pitched instruments. PSA is demonstrated in Figure 1, which shows the amplitude envelopes obtained from analysing a drum loop using PSA.

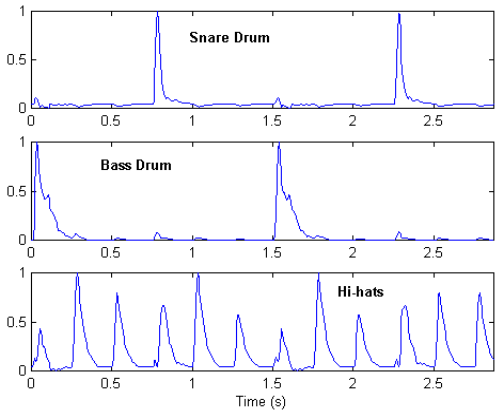

Figure 1. Drum loop separation using PSA

The prior subspaces used in this example were created by analysing large numbers of each type of drum. An ISA-type analysis such as described in [6] was carried out on each example. As mentioned previously this amounts to carrying out PCA followed by ICA on the spectrogram of the example. The first three principal components retained from the PCA step were passed to the ICA algorithm and the resulting independent frequency subspace with the largest projected variance was taken to represent the example. K-means clustering was then carried out on the frequency subspaces for a given drum type to yield a single subspace that best characterised a given drum type.

PSA was initially tested on 15 drum loops containing snares, kick drums and hi-hats. It achieved an overall success rate of 92.5% in successfully identifying the drums present. This represents an improvement over the 89.5% success rate achieved using sub-band ISA on the same signals [2]. PSA was found to be better than sub-band ISA in correctly identifying hi-hats and was also significantly faster than ISA or sub-band ISA due to the fact that PSA does not require the use of PCA. In tests on the same signals PSA was found to be approximately ten times faster than sub-band ISA and five times faster than ISA.

III PSA IN THE PRESENCE OF PITCHED INSTRUMENTS

It was previously noted that as the basis functions obtained by ISA are invariant no pitch changes are allowed within the sources present. It was also noted that PSA provides a relaxation of this assumption in that this restriction now only applies to the sources being searched for. As already noted drum sounds meet this criterion, making PSA a valuable tool for drum transcription. As it is no longer required that all the sources present be stationary in pitch, only the sources being searched for, it is possible to extend PSA to work in the presence of pitched instruments.

However a number of issues must be addressed before PSA can be used to transcribe drums in the presence of pitched instruments.

The first of these is to note that the presence of a large number of pitched instruments will cause a partial match with the prior subspace used to identify a given drum. This causes interference in the recovered amplitude envelope, which can in turn make detection of the drums more difficult. However it should be noted that pitched instruments have harmonic spectra with resulting regions of low intensity between partials. Furthermore due to the rules of harmony used in popular music many of the pitches played simultaneously will be in harmonic relation to each other and so will have many overlapping partials.

As a result every time pitched instruments occur there will be regions in the frequency spectrum where little or no energy is present due to pitched instruments. It can therefore be seen that using a higher frequency resolution reduces the interference due to the pitched instruments, and as a result improves the likelihood of recognition of the drums.

This is demonstrated in Figure 2, which shows the snare amplitude envelopes obtained from spectrograms of an excerpt from a pop song. The spectrograms had FFT sizes of 512 and 4096 respectively. The interference due to other instruments can be seen to be greatly reduced at the higher frequency resolution, and as a result the snare drum is more easily identified at the higher frequency resolution. However the use of higher frequency resolution comes at the price of a reduction in the time resolution, which leads to inaccuracies in the detected onset times of the drum events.

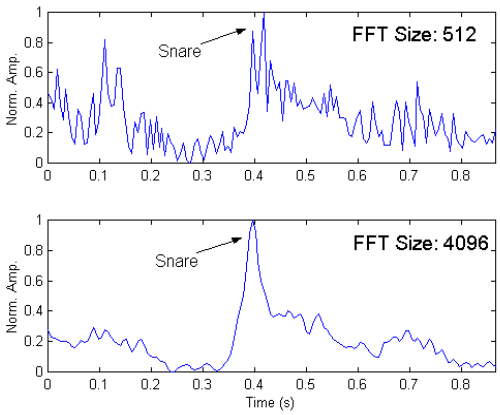

Figure 2. Snare envelopes at different frequency resolutions

Despite the use of high frequency resolution the interference present in the hi-hat subspace was in some cases found to be considerably greater than that in the bass drum or snare subspaces. This caused problems in trying to identify hi-hat events. The extra interference appears to be as a result of the fact that the hi-hat prior subspace has its energy spread out over a greater range of the spectrum than the snare and kick drum, making it more sensitive to the presence of pitched instruments.

However by noting that most of the energy of pop songs is contained in the lower region of the spectrum, it is possible to overcome this problem. The power spectral density (PSD) of a signal gives an estimate of the average power at each point in the spectrum [6]. Dividing a spectrogram by the PSD will emphasise those regions of the spectrum where there is less power, in this case the upper regions of the spectrum. This results in improved recognisability of the hi-hats. This is demonstrated in Figure 3 which shows the hi-hat amplitude envelopes obtained from an excerpt from a pop song both with and without PSD normalisation. The PSD was obtained using an eigenvector method using a small number of eigenvectors to capture only the broad regions where most of the energy occurs.

During testing of the modified PSA algorithm it was discovered that while successful in many cases, in some cases the algorithm did not perform correctly. Further analysis revealed that this was as a result of the sensitivity of the ICA algorithm to the interference or noise due to the presence of pitched instruments remaining in the snare and kick drum amplitude envelopes.

To overcome this problem all values in the amplitude envelope below a set threshold are set to zero. A normalised amplitude of 0.4 was found to be a suitable threshold for both the snare and kick drum. This operation is not carried out on the hi-hats as the interference was found to have been sufficiently eliminated by the PSD normalisation step.

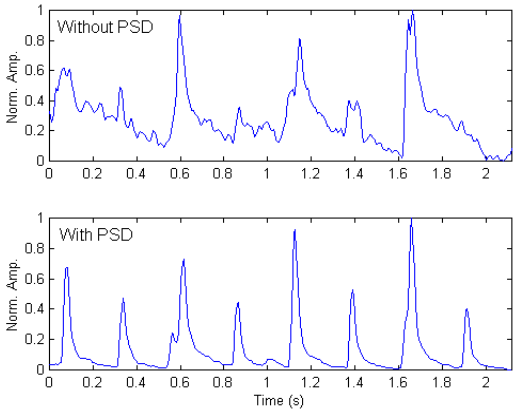

Figure 3. Hi-hat amplitude envelopes with/without PSD step

However the thresholding operation was found to have another consequence. The resulting snare and kick drum envelopes contained large areas of no activity, with sudden and sharp peaks occurring when a snare or kick occurred. This contrasts with the more natural peaks and decays occurring in the hi-hat envelope. When these very different amplitude envelopes were input to an ICA algorithm the

resulting independent signals contained unusual artifacts such as numerous sudden large amplitude modulations which were detected as events where none was present. To eliminate this problem it was necessary to carry out ICA on only the snare and bass drum amplitude envelopes, as they are comparable in that they both contain sharp peaks and large areas of no activity. This resulted in the correct separation of bass drums and snare drums in most cases. The hi-hat envelope is passed directly to the onset detection algorithm. While this gives good results in general it can result in extra errors in detection of hi-hats. As the hi-hat amplitude envelope no longer undergoes ICA the algorithm loses the ability to distinguish between a snare occurring on its own and a snare and hi-hat occurring simultaneously. However in many cases a hi-hat does occur simultaneously with the snare, so this only results in a small reduction in the efficiency of the transcription algorithm.

IV DRUM TRANSCRIPTION IN THE PRESENCE OF PITCHED INSTRUMENTS

To test the ability of PSA to transcribe drums in the presence of pitched instruments a drum transcription system was implemented in Matlab. The system implemented deals only with snares, bass drums and hi-hats. Due to the source signal ordering problem in the ICA step it is assumed that the bass drum has a lower spectral centroid than the snare. The system was tested on 20 excerpts taken at random from pop songs from as wide a range of styles as possible ranging from pop to disco and rock. The drum patterns from these excerpts were transcribed by an expert listener.

Because of the imperfect separation of the ICA step the amplitude envelopes were normalised and onsets over a given threshold were taken to be a drum onset. The same threshold was used for both snare and kick drums while a lower threshold was used for the hi-hats. This reflects the fact that the amplitude of the hi-hats in real world examples can vary widely depending on the style of drumming. The results obtained are outlined in Table 1. Though the results demonstrate the effectiveness of PSA as a method for transcribing drums in the presence of pitched instruments a greater number of errors occur than for PSA with drums only. Possible reasons for this are discussed below.

Type	Total	Missing	Incorrect	%
Snare	57	1	9	82.5
Kick	84	4	7	86.9
Hi-hats	238	14	30	81.5
Overall	379	19	46	82.8

Table 1. Drum Transcription Results

In the case of the bass drums, six snare events were incorrectly identified as bass drums. These errors occurred in excerpts where a “disco” style of drumming was employed. In these excerpts the snare drum is typically less bright than in the other genres of music, and so a greater chance of incorrect identification is the result. Only one of the incorrect bass drum detections was as a result of a bass guitar note being identified as a bass drum. The missing four undetected bass drum events were visible on the amplitude envelope of the excerpts in question, but were below the threshold for detection. The bass drums at these points were audibly lower than the other bass drum events in the excerpts.

In the case of the snare drum, five of the incorrect snares were as a result of the combination of a bass drum and a hi-hat occurring simultaneously being mistaken for snares. This happened in two excerpts. The remaining errors occurred as a result of noise due to pitched instruments.

With regards to the hi-hats the majority of incorrect identifications were as a result of interference that had not been eliminated in the PSD normalisation step. In two cases an event with the characteristics of a hi-hat was clearly visible in both the spectrogram and the recovered amplitude envelope, but no event of this type was audible to the listener. These events may be genuine hi-hat events that have been masked by other audio events, but as there is no way of determining this for excerpts from commercial recordings, these onsets have been classed as incorrect detections. In the case of the undetected hi-hats the majority of the hats were clearly visible in the amplitude envelopes, but below the threshold required for identification. Further improvements may be possible by adjusting the thresholds for detection but there is a trade-off between reducing the number of incorrect identifications and increasing the number of missed events.

Due to the limitations in the time resolution of the STFT, the detection of onset times had an average error of 10ms. It should be noted that this error tended to be consistent across all the drums in a given loop, so that inter-onset intervals remained consistent within a given loop. However it is still desirable to improve the accuracy of onset detection in PSA.

It should be noted that these results were obtained without the use of any form of rhythmic modelling to predict when a given drum was most likely to occur.

V CONCLUSIONS & FUTURE WORK

Prior Subspace Analysis has been shown to be a viable approach for the transcription of drums in the presence of pitched instruments, overcoming some of the problems associated with Independent Subspace Analysis. Further work needs to be done to improve the correct identification of the drums and to increase the accuracy of the onset times. It is also proposed to

generalise the method to deal with an increased number of drum types.

REFERENCES

[1] Casey, M. & Westner, A., "Separation of Mixed Audio Sources By Independent Subspace Analysis" *Proceedings Of ICMC 2000*, pp. 154-161, Berlin, Germany, 2000.

[2] FitzGerald, D., Lawlor, B., Coyle, E., "Prior Subspace Analysis for Drum Transcription", 114th AES Conference Amsterdam March 22nd–25th 2003

[3] FitzGerald, D., Coyle E, Lawlor B., "Sub-band Independent Subspace Analysis for Drum Transcription", *Proceedings of the. Digital Audio Effects Conference (DAFX02)*, Hamburg, pp. 65-69, 2002.

[4] Hyvärinen A. & Oja E., "Independent Component Analysis: Algorithms and Applications". *Neural Networks*, 13(4-5): pp. 411-430, 2000.

[5] Sillanpää J., Klapuri A., Seppänen J., Virtanen T., "Recognition of acoustic noise mixtures by combining bottom-up and top-down processing". In proc. European Signal Processing Conference, EUSIPCO 2000.

[6] Casey, M., "Auditory Group Theory: with Applications to Statistical Basis Methods for Structured Audio", *Ph.D. Thesis*, MIT Media Lab, February 1998.

[7] Vaseghi, Saeed V., *Advanced Digital Signal Processing and Noise Reduction*, 2nd ed. John Wiley & Sons Ltd. pp. 270-290.