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 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 effectiveness and robustness of PSA as a tool for drum transcription in the presence of pitched instruments is demonstrated in a simple drum transcription algorithm.**

I INDEPENDENT SUBSPACE ANALYSIS

Independent Subspace Analysis (ISA) is a technique for separating sound sources from single channel mixtures [1]. Based on the concept of reducing redundancy in time-frequency representations it represents sound sources as low dimensional subspaces in the time-frequency plane. The signal is transformed to a magnitude spectrogram by means of a Short-Time Fourier Transform (STFT). ISA assumes that the mixture signal spectrogram **Y** can be decomposed into l statistically independent spectrograms Y_j:

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

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:

$$Y_j = f_j t_j \qquad (2)$$

Summing over the Y_j yields:

$$\mathbf{Y} = \sum_{j=1}^{l} f_j t_j^T \qquad (3)$$

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.

One method of decomposing a spectrogram into a sum of outer products is by means of Principal Component Analysis (PCA). PCA takes a set of correlated variables and linearly transforms them into a number of uncorrelated or orthogonal variables that are termed principal components. These principal components are ordered by the amount of variance of the original variables they contain. As the principal components are ordered by variance PCA can be used to reduce redundancy by discarding components of low variance. PCA is in this case performed using the singular value decomposition method, which decomposes **Y**, an m x n spectrogram, into:

$$\mathbf{Y} = USV^T \qquad (4)$$

where U is an m x m matrix, the columns of which contain the principal components of **Y** on a frequency basis, V is an n x n matrix whose columns contain the principal components on a time basis, and S is an m x n matrix of singular values. As the sound sources are assumed to be low-dimensional subspaces in the time-frequency plane dimensional reduction is carried out by discarding components of low variance. If the first *l* principal components are retained then equation 4 can be rewritten as:

$$\mathbf{Y} \approx \sum_{j=1}^{l} u_j s_j v_j^T \qquad (5)$$

By letting $u_j s_j$ equal h_j and v_j equal z_j it can be seen that the spectrogram has been decomposed into a sum of outer products as described in equation 3. In matrix notation this becomes:

$$\mathbf{Y} \approx \mathbf{hz^T} \qquad (4)$$

However PCA does not return a set of statistically independent basis functions. To obtain independent basis functions a further procedure, known as Independent Component Analysis (ICA), must be carried out [2].

Independent Component Analysis attempts to separate a set of observed signals that are composed of mixtures of a number of independent non-gaussian sources into a set of signals that contain the independent sources. As musical signals are non-gaussian in nature the assumption of non-gaussianity is valid. The independent sources are assumed to have been mixed linearly. Using vector-matrix notation this can be stated as:

$$\mathbf{x} = \mathbf{As} \qquad (5)$$

where **x** contains the observed mixture signals, **s** contains the independent non-gaussian sources, and **A** is the mixing matrix.

To recover the independent sources ICA makes use of a corollary of the central limit theorem. The central limit theorem states that mixtures of non-gaussian signals will tend towards a gaussian distribution as the number of signals increases. As a result the mixture signals in **x** will have probability density functions that are closer to gaussian than the source signals in **s**. It can then be seen that the original sources will have probability density functions more non-gaussian than any mixture of the sources. It can therefore be seen that finding an unmixing matrix which gives a set of signals that are as non-gaussian as possible will, in most cases, result in the recovery of the independent sources.

It should be noted that ICA cannot recover the signals at their original amplitudes or in the order in which the signals are presented. However in practice these restrictions do not affect the usefulness of ICA methods. There are numerous algorithms publicly available for performing ICA, such as FastICA and Jade [3,4]. Good reviews of ICA methods can to be found in [2,5].

Carrying out ICA on **h** to obtain basis functions independent on frequency yields:

$$\mathbf{f} = \mathbf{Wh} \qquad (6)$$

where **f** contains the independent frequency basis functions, and **W** is the unmixing matrix obtained from ICA. The associated amplitude basis functions can then be obtained from obtained by multiplying the spectrogram **Y** by the pseudoinverse of the frequency basis functions **f**, where $\mathbf{f_{pinv}}$ denotes the pseudoinverse of **f**. This yields:

$$\mathbf{t} = \mathbf{f_{pinv}Y} \qquad (7)$$

The independent spectrograms can then be obtained as described in equation 2. If independence on a time basis is required it can be obtained in a similar manner to that described above.

As ISA works on the magnitudes of the STFT coefficients there is no phase information available to allow resynthesis of the separated sounds. A fast but crude way of obtaining phase information for resynthesis is to reuse the phase information from the original STFT. Phase information for resynthesis can be also be obtained via an iterative phase estimation method such as that described by Griffin and Lim [6].

II LIMITATIONS OF ISA METHODS

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. To overcome this the signal can be broken up into small sections which can be assumed to be stationary in pitch. The basis functions forming a source across the sections are then clustered using a mean-field clustering algorithm such as described in [7]. However this is not necessary when the sources can be considered stationary in pitch such as with drum loops.

Secondly the quality of separation also depends on the length of the signal input. For instance a signal containing just one hi-hat and snare played simultaneously will not separate correctly. For the hi-hat/snare separation 2-4 events are typically required, depending on the frequency and amplitude characteristics of the drums used.

Thirdly 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. There is also a trade-off between the number of components retained and the recognisability of the resulting basis functions. Keeping a large number of components results in basis functions that support small regions of the frequency spectrum. Using a small number of components results in recognisable basis functions with support across the entire frequency spectrum

As a result of this trade-off ISA works best on signals with less than five sources. This trade-off also means that it is necessary to choose carefully the number of components retained to achieve optimal source separation. As the number of components required varies with the signal being analysed this means that an observer is necessary to determine the required number of components. Methods such as sub-band ISA have been proposed in an effort to overcome this indeterminacy for the purposes of drum transcription [8].

Fourthly, in the case of broadband noise-based nature of drum sounds there will be regions of frequency overlap between the sounds, and as a result sometimes other drums show up as small peaks in the amplitude envelopes of the separated drums. However when good separation is obtained a simple thresholding operation is usually sufficient to identify the required events.

Finally due to fact that the ICA step is indeterminate with regards to ordering of the input components it is necessary to identify a given source by some means such as their frequency characteristics or amplitude envelopes after ISA has been completed.

However, 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 mixtures of drums [9].

III PRIOR SUBSPACE ANALYSIS

As noted above there are a number of problems associated with the ISA method. In particular estimating the optimal number of components required for separation is a source of much difficulty. While methods such as sub-band ISA go some way to overcoming this problem, a more efficient method lies in the use of prior knowledge of the sources to be separated. Prior Subspace Analysis (PSA) makes use of prior models of the sources to obtain sound source separation in single channel mixtures [10]

ISA arose out of attempts to create a signal representation that could characterise, and allow further manipulation of, everyday sounds such as a coin hitting the floor [11]. The method looked for invariants that characterised sounds and involved performing PCA followed by ICA on a spectrogram of a sound in a similar manner to that of ISA. This method was later incorporated into the MPEG 7 specification for sound classification [12]. The success of this method suggests that it can be adapted to create a set of prior subspaces that can characterise a given sound source such as a snare drum. These prior subspaces can then be used to carry out an initial analysis of the spectrogram of a mixture signal.

Stated formally, PSA assumes that the overall spectrogram can be decomposed in the manner described by equation 3. It then assumes that there exists known prior frequency basis functions f_p that are good initial approximations to the actual basis functions. Substituting the f_p for the f_j in equation 3 yields:

$$\mathbf{Y} \approx \sum_{j=1}^{l} f_p t_j^T \qquad (8)$$

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

$$\hat{\mathbf{t}} = \mathbf{f}_{\mathbf{pp}}\mathbf{Y} \qquad (9)$$

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

$$\mathbf{t} = \mathbf{W}\hat{\mathbf{t}} \qquad (10)$$

Improved estimates of the frequency basis functions can then be obtained from

$$\mathbf{f} = \left(\mathbf{Y}\mathbf{t}_{\mathbf{p}}\right)^{\mathbf{T}} \qquad (11)$$

The independent spectrograms can then be individually obtained in the manner shown in equation 2. Resynthesis of the separated sound sources can then be carried out in a manner similar to that of ISA.

As can be seen from the above PSA differs from ISA in the manner in which decomposition of the spectrum is carried out. ISA carries out the decomposition of the spectrogram in a blind manner, using PCA to obtain what is considered the most important information in the spectrogram. In contrast PSA uses prior knowledge to obtain the most important information on the sources of interest. As noted previously ISA does not always yield the required information on the sources of interest, PSA overcomes this through the use of prior knowledge about the sources of interest.

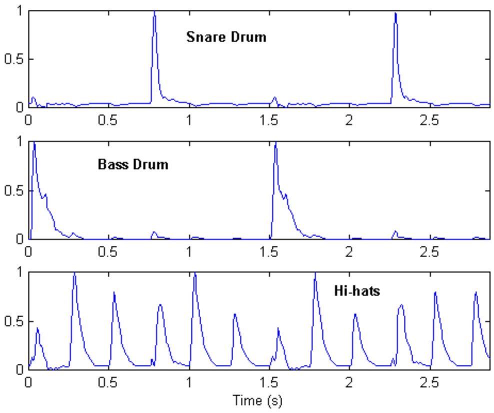

Figure 1. Drum loop separation using PSA

PSA is demonstrated in Figure 1, which shows the amplitude envelopes obtained from analysing a drum loop using PSA. The prior subspaces were created by analysing large numbers of each type of drum. An ISA-type analysis such as described in [11] 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. 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.

IV 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 should be 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 4096 and 512 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.

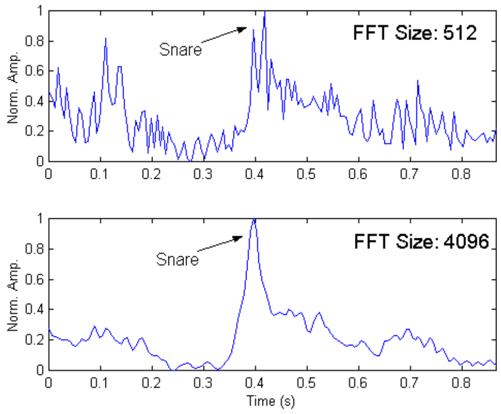

Figure 2. Snare envelopes at different frequency resolutions

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 [13]. 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.

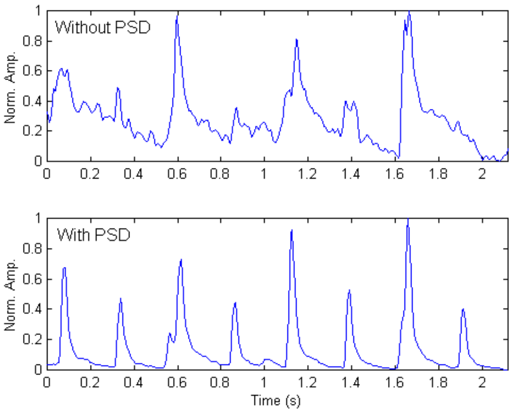

Figure 3. Hi-hat amplitude envelopes with/without PSD step

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.

V 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.

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

[3] FastICA package for Matlab, http://www.cis.hut.fi/projects/ica/fastica/

[4] Jade algorithm for ICA, http://www.tsi.enst.fr/icacentral/algos.html

[5] Cardoso, J.F., Blind Signal Separation: statistical Principles, Proceedings of the IEEE, Vol.9, No. 10, pp. 2009-2025, Oct 1998,1

[6] Griffin, D., & Lim, J. S. "Signal Estimation from Modified Short-Time Fourier Transform", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-32, pp. 236-243, 1984.

[7] Hofmann, T., and Buhmann, J.M., "Pairwise data clustering by deterministic annealing.", IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(1):1-14,1997

[8] 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.

[9] 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

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

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

[12] Casey, M., "Generalized Sound Classification and Similarity in MPEG-7", *Organized Sound*, 6:2, 2002

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