2007-01-01

Resynthesis Methods for Sound Source Separation using Shifted Non-negative Factorisation Models

Derry Fitzgerald
Cork Institute of Technology

Matt Cranitch
Dublin Institute of Technology

Recommended Citation

Fitzgerald, D. & Cranitch, M. Resynthesis methods for sound source separation using shifted non-negative factorisation models. Paper given at the *Irish Signals and Systems Conference, Derry, Ireland, 2007.*

Resynthesis methods for Sound Source Separation using shifted Non-negative Factorisation Models.

Derry FitzGerald, Matt Cranitch[*], Eugene Coyle[**]

[*] Dept. of Electronic Engineering,
Cork Institute of Technology,
IRELAND
derry.fitzgerald@cit.ie

[**] School of Control Systems and Electrical Engineering,
Dublin Institute of Technology,
IRELAND
eugene.coyle@dit.ie

Abstract— **Recently, techniques such as shifted 2D non-negative matrix factorisation and shifted 2D non-negative tensor factorisation have been proposed as methods for separating harmonic musical instruments from single and multi-channel mixtures. However, these methods require the use of a Constant Q transform, for which no true inverse exists. This has adverse effects on the quality of the resynthesis of the separated sources. In this paper, a number of different resynthesis methods are investigated in order to determine the best approach to resynthesis.**

***Keywords* – Sound Source Separation, Non-negative matrix and tensor factorisation**

I INTRODUCTION

In recent years, methods such as shifted 2D non-negative matrix factorisation (2DNMF) and tensor factorisation (2DNTF) have been proposed as a means of separating mixtures of harmonic pitched instruments in the single and multi-channel cases respectively [1],[2]. These techniques overcome some of the problems associated with the use of standard non-negative matrix and tensor factorisations (NMF and NTF respectively) [3],[4] for the purposes of musical instrument source separation, such as the problem of grouping the basis functions to their sources automatically. This is achieved by incorporating shift invariance in both the frequency and time basis functions recovered by the algorithms, thereby modelling each source or pitched instrument as translations of successive spectra in both frequency and time, thereby allowing time-varying spectra and fundamental frequencies.

Taking 2DNMF as an example, the decomposition model can be expressed as:

$$\mathcal{X} \approx \hat{\mathcal{X}} = \left\langle \langle \mathcal{T}\mathcal{A} \rangle_{\{3,1\}} \langle \mathcal{S}\mathcal{P} \rangle_{\{3,1\}} \right\rangle_{\{2:4,1:3\}} \qquad (1)$$

where $\mathcal{X}$ is a tensor of size *n* x *m*, containing a magnitude spectrogram of the mixture signal and $\hat{\mathcal{X}}$ is an approximation to $\mathcal{X}$. $\mathcal{T}$ is an *n* x *z* x *n* translation tensor, which translates the frequency basis functions in $\mathcal{A}$ up or down in frequency, thereby approximating different notes played by a given source. $\mathcal{A}$ is a tensor of size *n* x *K* x *p*, where *p* is the number of translations across time. $\mathcal{S}$ is a tensor of size *z* x *K* x *m* and $\mathcal{P}$ is a translation tensor of size *m* x *p* x *m*, which translates the amplitude envelopes contained in $\mathcal{S}$ across time, thereby allowing time-varying source spectra. $\langle\ \rangle$ denotes contracted tensor multiplication along the modes indicated in the angle brackets.

Using a cost function which encourages sparseness in $\mathcal{A}$ and $\mathcal{S}$ results in a factorisation where the basis functions in $\mathcal{A}$ and $\mathcal{S}$ correspond to perceptually meaningful features, such as typical frequency spectra of individual instruments and their associated amplitude envelopes. Such a sparse factorisation can obtained by minimising the generalised Kullback-Liebler divergence between $\mathcal{X}$ and $\hat{\mathcal{X}}$. This is defined as:

$$D\left(\mathcal{X}\middle\|\hat{\mathcal{X}}\right) = \sum_{i,j} \mathcal{X}\log\frac{\mathcal{X}}{\hat{\mathcal{X}}} - \mathcal{X} + \hat{\mathcal{X}} \qquad (2)$$

where *i* and *j* index over frequency bin and time frame respectively. Update equations for $\mathcal{A}$ and $\mathcal{S}$ can be derived in a manner similar to that presented in [2]. For a given number of sources, the free parameters are *z*, the number of frequency translations and *p*, the number of time translations. Separation is performed by estimating individual source spectrograms from:

$$X_k = \left\langle \langle \mathcal{T}\mathcal{A}_{:k} \rangle_{\{3,1\}} \langle \mathcal{S}_{:k}\mathcal{P} \rangle_{\{3,1\}} \right\rangle_{\{2:4,1:3\}} \qquad (3)$$

where X_k is the estimated log-frequency spectrogram of the k^{th} source, and $:k$ denotes the tensor slice associated with the k^{th} source.

Both 2DNMF and 2DNTF have proved successful in separating mixtures of pitched instruments in both single and multi-channel cases. However, introducing shift invariance in frequency requires the use of a time-frequency spectrogram that has log-frequency resolution, such as the constant Q transform [5]. Alternatively, log-frequency resolution can be obtained using weighted sums of a linear time-frequency spectrogram such as obtained via a short-time Fourier transform (STFT). This can be expressed as:

$$X = CY \qquad (4)$$

where Y is the linear time-frequency spectrogram of size f frequency bins and t time frames, C is the weighting matrix of size cf x f, which maps the f linear frequency bins to cf log-frequency resolution bins, with $cf < f$ and X is a log frequency spectrogram of size cf x t. As C is a rectangular matrix, no true inverse exists, and so any mapping back from log-frequency resolution to linear frequency resolution will only be an approximate inverse. Typically, the pseudoinverse of C can be used to obtain a least squares approximation of the inverse. A similar problem arises when using the Constant Q transform.

The approximate nature of the inverse has an adverse effect on the sound quality of the separated sources, and so alternative methods for resynthesis have been suggested, such as in [1], where spectrogram masks were constructed for each instrument by assigning each bin to the instrument with the highest power at that bin. These spectrogram masks were then mapped back to the linear frequency domain, and used to filter the original complex spectrogram. The resultant spectrogram was then inverted back to the time domain.

The type of binary masking described above is equivalent to assuming that the instruments are disjoint orthogonal, i.e. that the sources do not overlap in time or frequency. This is the assumption used in the DUET algorithm, which has proved successful in separating speech signals, which can be considered to be approximately disjoint orthogonal [6]. However, this assumption does not hold well for musical signals where the instruments typically play in harmony with one and other, resulting in overlapping partials. As a result, this type of binary masking is less than optimal as a means of resynthesis of musical signals. The remainder of this paper explores different methods of resynthesising the separated signals, with a view to determining a more effective method of resynthesising the separated sources.

II MAPPING FROM LOG TO LINEAR FREQUENCY

The resynthesis methods explored in this paper can be divided into three main groups according to how the mapping of the source spectrograms from the log-frequency domain back to the linear frequency domain is performed. The first method used for obtaining this mapping is the pseudoinverse of C, which, as already noted, provides the best mapping in a least squares sense. This can be expressed as:

$$\hat{Y} = C^{+} X_k \qquad (5)$$

where $\hat{Y}$ is a linear frequency domain spectrogram and $^{+}$ denotes pseudoinverse. However, the use of the pseudoinverse can result in negative values in the recovered magnitude spectrogram. This runs contrary to the definition of a magnitude spectrogram and can result in artifacts in the resynthesis.

The second method explored is to simply use the transpose of C to do the mapping. This can be expressed as :

$$\hat{Y} = C' X_k \qquad (6)$$

where ′ denotes matrix transpose

This has the advantage of ensuring the non-negativity of the recovered magnitude spectrogram, though the spectrogram will now be scaled differently to the original. Fortunately, the measures of signal separation and quality used (See section III for details) are invariant to gain changes in the recovered signals, and for playback, the signals can be rescaled to the desired level.

The third method aims at arriving at a compromise between the previous two mappings, namely finding the best least squares approximation, subject to the constraints that the recovered magnitude spectrogram is non-negative. Such a mapping can be determined using a simplified version of NMF, using the Euclidian distance as a cost function. The iterative multiplicative update rule for determining $\hat{Y}$ is then given by:

$$\hat{Y} = \hat{Y}.*(CX_k)./(C'C\hat{Y}) \qquad (7)$$

where .* denotes elementwise multiplication, and ./denotes elementwise division $\hat{Y}$ is randomly initialised and the algorithm run to convergence. This method is more computationally expensive than the other methods and so takes longer to run.

III RESYNTHESIS METHODS

For each of these mappings, a number of different methods of resynthesis are then implemented. The first of these is to apply the original phase information to $\hat{Y}$ and to invert the resultant spectrogram to the time domain. The second is the method used by Schmidt et al in [1], which was described previously in section I.

The third method used is to use the recovered source spectrogram to filter the original spectrogram. This can be written as:

$$\hat{S}_i = \hat{Y}_i F \qquad (8)$$

where $\hat{S}_i$ is the estimated complex spectrogram of the i^{th} source, F is the original complex mixture spectrogram, and $\hat{Y}_i$ is the estimated magnitude spectrogram of the i^{th} source. This method has an advantage over the second method in that it is no longer based on binary masking and so should give better results when dealing with musical signals.

The fourth method can be described as source cancellation, where the source of interest is estimated by elementwise division of the original spectrogram with the sum of the estimated spectrograms of the other sources. This can be written as:

$$\hat{S}_i = \frac{F}{\sum\limits_{j(j\neq i)} \hat{Y}_j} \qquad (9)$$

where $\hat{Y}_j$ is the estimated magnitude spectrogram of the j^{th} source. This is similar to the cancellation approach used in [8]. A problem with this method is that in carrying out the cancellation it does not take into account regions where the recovered source is strong or weak. Therefore a potentially better method of cancellation is given by:

$$\hat{S}_i = \frac{\hat{Y}_i}{\sum\limits_{j} \hat{Y}_j} .* F \qquad (10)$$

where j now indexes over all sources. This method can be viewed as a combination of the second and third methods.

The final method tested can be written as:

$$\hat{S}_i = \frac{\hat{Y}_i^{\,2}}{\sum\limits_{j} \hat{Y}_j^{\,2}} .* F \qquad (11)$$

where the square operation is carried out elementwise. This is equivalent to the adapted Wiener filtering approach proposed by Benaroya et al in [9] for stationary Gaussian sources. While audio signals can be considered approximately stationary on a frame by frame basis, musical signals are non-gaussian in nature. Nevertheless, it represents a simple method of attempting a Wiener filtering type approach.

IV EVALUATION METHODOLOGY

In order to evaluate the performance of the previously discussed methods, a set of 5 test signals was created. These test signals all consist of single channel mixtures of two instruments, and the original individual source waveforms were retained to allow evaluation of the separation performance. As the mixtures are single channel, the separation algorithm used is 2DNMF, previously described in Section I. The separation performance of these algorithms is particularly sensitive to the choice of *z*, the number of frequency translations, and so the separation algorithm was run several times to determine the optimal choice of *z* for each example. The choice of *p* was fixed at 5 for each example, as the separation algorithm is not as sensitive to the choice of *p*.

Once the optimal *z* for each example was identified, the separation algorithm was run for each example, and the source log-frequency spectrograms recovered. Each of the three mappings to from log to linear frequency was then performed, and each of the six resynthesis methods performed, giving in total 18 different resynthesised waveforms for each source. As all 18 resynthesis methods are performed on the same log-frequency source spectrogram, the resulting waveforms give a reliable indication of the effectiveness or otherwise of the resynthesis methods.

The performance metrics used for evaluation of the various resynthesis methods are those defined in [10]. In this case, the recovered time-domain source is decomposed, with reference to the original unmixed sources, into the sum of three terms:

$$s_{recov} = s_{target} + e_{interf} + e_{artif} \qquad (12)$$

where s_{recov} is the recovered source, s_{target} is the portion of the recovered signal that relates to the original or target source, e_{interf} is the portion of the recovered signal that relates to other interfering sources, and e_{artif} is the portion of the recovered signal that relates to artifacts generated by the separation algorithm and/or the resynthesis method. The performance metrics used are the Signal to Distortion ratio (SDR), which provides an overall measure of the quality of the sound source separation:

$$SDR = 10\log_{10} \frac{\left\| s_{target} \right\|^2}{\left\| e_{inter} + e_{artif} \right\|^2} \qquad (13)$$

the Signal to Interference ratio (SIR), which provides a measure of the presence of the other sources in the separated sounds:

$$SIR = 10\log_{10} \frac{\left\| s_{target} \right\|^2}{\left\| e_{interf} \right\|^2} \qquad (14)$$

and the Signal to Artifacts ratio (SAR) which provides a measure of the artifacts present in the recovered signal due to separation and resynthesis:

$$SAR = 10\log_{10} \frac{\left\| s_{target} + e_{interf} \right\|^2}{\left\| e_{artif} \right\|^2} \qquad (15)$$

These measures were designed to be used with separation techniques such as Independent Component Analysis [7], where the signals could be

	Signal 1			Signal 2			Signal 3			Signal 4			Signal 5		
Method	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
P-inv	4.1	37.4	4.5	2.3	23.0	2.4	-1.7	**54.1**	-1.5	2.3	14.5	3.0	-3.4	25.5	-3.4
P-Sch	4.8	35.1	4.9	-2.1	28.7	-2.0	4.2	31.2	4.3	-0.8	**37.2**	-0.8	-4.8	22.5	-4.7
P-Mask	5.2	34.6	5.4	4.2	25.6	4.3	3.7	34.8	3.8	3.9	17.7	4.1	-3.6	24.3	-3.5
P-Cross	-8.1	28.9	-8.0	-14	20.4	-13	-9.5	19.5	-9.4	-19	14.8	-19	-17	25.6	-17
P-Wien	**7.9**	30.7	8.2	12.0	28.7	12.2	10.0	30.8	10.2	6.1	17.6	6.7	-3.0	**45.0**	-3.0
T-inv	3.2	19.8	3.6	-0.2	16.3	-0.0	0.2	26.2	0.4	2.4	13.0	3.1	-4.2	21.7	-4.1
T-Sch	6.1	22.4	6.5	10.0	22.5	10.3	6.5	29.4	7.0	5.7	16.7	6.5	-3.2	19.9	-3.1
T-Mask	5.1	24.7	5.4	6.8	21.0	7.0	3.8	31.7	4.2	4.4	16.2	4.8	-4.0	21.8	-3.9
T-Can	0.6	26.7	0.8	3.1	20.8	3.3	-9.7	44.2	-9.7	-2.6	13.3	-2.1	-6.5	34.6	-6.5
T-Cross	7.7	20.3	**8.4**	11.5	18.3	12.7	**10.6**	24.5	**11.2**	5.6	13.1	**7.1**	-3.1	23.2	-3.1
T-Wien	7.7	28.6	8.0	12.0	26.5	12.2	10.3	31.4	10.6	**6.0**	17.4	6.7	-3.0	35.0	-3.0
L-inv	6.5	34.3	6.9	5.3	24.1	5.4	4.1	27.5	4.3	4.2	14.3	5.2	**-2.9**	27.0	**-2.9**
L-Sch	5.9	25.7	6.1	6.2	**53.0**	6.3	6.8	41.2	6.9	2.4	19.8	2.7	-3.6	37.0	-3.5
L-Mask	5.2	35.4	5.4	4.2	27.0	4.3	3.7	34.6	3.9	3.9	17.4	4.2	-3.6	25.6	-3.6
L-Cross	7.8	24.4	8.3	11.7	20.5	**12.6**	10.0	27.7	10.5	5.2	12.8	6.7	-3.1	25.5	-3.1
L-Wien	**7.9**	**42.0**	8.2	**12.1**	26.3	12.3	9.9	33.1	10.2	5.6	15.4	6.6	-3.1	31.3	-3.1

Table 1: Resynthesis sound quality results for each mixture signal. P denotes pseudoinverse mapping, T denotes transpose mapping, while L denotes non-negative least squares mapping. Inv signifies the direct use of the inverse for resynthesis, Sch, the method used by Schmidt et al, Mask, the use of the inverse to filter the original spectrogram, Can, source cancellation, Cross, the hybrid of the previous two methods, and Wien, the modified Wiener filtering approach. The best scores are highlighted in bold.

recovered up to a scaling factor, and so these measures are invariant to the scale of the signals. These metrics were calculated using the BSS_EVAL toolbox for Matlab, available at [11].

These metrics are obtained for each source, and so do not provide an overall measure of the separation and resynthesis across each of the test signals. This was obtained by averaging the metrics of each source, and overall results for each of the 18 resynthesis methods were obtained by averaging across the test signals.

V RESULTS & DISCUSSION

The results obtained for each of the mixture signals are presented in Table 1, with the best result for each metric highlighted in bold. Table 2 presents an overall summary of the results for each method averaged across all of the mixture signals, with the final column giving the average result across all metrics. The source cancellation approach for both pseudoinverse and least squares mappings has been removed from the results as informal listening tests showed that for these methods the recovered signals are unrecognisable to the listener, and are therefore unusable for time-domain resynthesis.

The degree of separation and the resultant scores obtained vary widely across the different test signals. The separation quality depends on a number of different factors, such as the degree of overlap in time and frequency of the sources and the similarity of the instrument timbres in the mixtures. Despite this, it is still possible to make a number of observations with regards to the resynthesis methods.

It can be seen that the SIR is quite high for practically all of the methods; this demonstrates that the 2DNMF method is capable of separating mixtures of pitched instruments to a high degree. It can also be noted that the source cancellation approach provides lower quality of resynthesis than the other methods, with low SDR and SAR scores, regardless of what mapping from log to linear frequency was used.

Further, it can be seen that the pseudoinverse based methods have lower SDR and SAR than the transpose and non-negative least squares mapping methods. This highlights the need to use a mapping which reflects the non-negative nature of magnitude spectrograms. This is further supported by the fact that the pseudoinverse Wiener filtering approach, where the filter is non-negative, outperforms the other pseudoinverse approaches in terms of SDR and SIR.

The best SDR is achieved through the use of the adapted Wiener filtering approach, regardless of the mapping method, followed closely by the masking/cancellation hybrid approach for both transpose and non-negative least squares mappings. Similarly, these methods give good SAR scores, showing that the resynthesis achieved is relatively free of artefacts. Further, the best scores for both SDR and SAR in all but one of the test signals are obtained using these methods, and even in this case, these methods still give the second best score.

In terms of the trade-off between the separation of the sources and the quality of the resynthesis, it can be seen that the adapted Wiener filtering approach represents the best method for resynthesising the separated signals, with the pseudoinverse mapping slightly outperforming the other mapping methods. However, it should be noted that informal listening tests on the resynthesised separated sources indicate that it can be hard for the listener to discriminate between the adapted Wiener filtering approaches and those obtained from the masking/cancellation hybrid approach for both transpose and non-negative least squares mappings.

This highlights a problem with the performance metrics used, namely that they are not perceptually based metrics, and so may not reflect what the listener perceives. This is further borne out by the fact that the informal listening tests indicated that the quality of the separated sources from test signal 5 is comparable to that of the other signals, despite the fact that they have lower SDR and SAR scores. Nevertheless, the scores obtained do in general provide a means of determining which resynthesis methods are the most effective. Despite this, it is felt that perceptually based performance metrics would give an overall better indication of the separations obtained. The development of such metrics remains an open issue.

Method	SDR	SIR	SAR	AVG
P-inv	0.7	30.9	1.0	10.9
P-Sch	0.3	30.4	0.3	10.3
P-Mask	2.7	27.4	2.8	11
P-Cross	-13.5	21.8	-13.4	-1.7
P-Wien	**6.6**	30.6	6.9	**14.7**
T-inv	0.3	19.4	0.6	6.8
T-Sch	5.0	22.2	5.5	10.9
T-Mask	3.2	23.1	3.5	9.9
T-Can	-3.0	27.9	-2.8	7.4
T-Cross	6.5	19.9	**7.3**	11.2
T-Wien	**6.6**	27.8	6.9	13.8
L-inv	3.4	25.4	3.8	10.9
L-Sch	3.5	**35.2**	3.7	14.1
L-Mask	2.7	28.0	2.8	11.2
L-Cross	6.3	22.2	7.0	11.8
L-Wien	6.5	29.6	6.9	14.3

Table 2: Average results for each resynthesis method

IX CONCLUSIONS

The problem of obtaining good quality resynthesis from sound source separation techniques such as 2DNMF and 2DNTF has been highlighted. This is as a result of the necessity of mapping from log to linear frequency resolutions. Several different mappings from log to linear frequency have been suggested, and a number of different resynthesis methods have been explored. The results obtained suggest that the best method for resynthesis of the separated sources is an adapted Wiener filtering approach, followed by a hybrid masking/source cancellation approach. This result is borne out in informal listening tests. However, the need for perceptually based performance metrics has also been demonstrated, and this remains an area for future research.

REFERENCES

[1] M. Schmidt and M. Morup, “Nonnegative Matrix Factor 2D Deconvolution for Blind Single Channel Source Separation”, Proceedings of the International Conference on Independent Component Analysis 2006.

[2] D. FitzGerald, M. Cranitch, E. Coyle, “Shifted 2D Non-negative Tensor Factorisation”, Proceedings of the Irish Signals and Systems Conference, Dublin, Ireland, 2006

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

[4] D. FitzGerald, M. Cranitch and E. Coyle, “Non-negative Tensor Factorisation for Sound Source Separation”, Proceedings of the Irish Signals and Systems Conference, Dublin 2005.

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

[6] S. Rickard and O. Yilmaz, “On the W-Disjoint Orthogonality of Speech”, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2002), Orlando, Florida, USA, 2002.

[7] P. Comon, “Independent component analysis - a new concept?” Signal Processing, 36 pp. 287_314, 1994

[8] D. Barry, D. FitzGerald, E. Coyle and B. Lawlor, "Single Channel Source Separation using Short-Time Independent Component Analysis", 119th AES Convention, October 2005 New York.

[9] L. Benaroya, F. Bimbot, and R. Grimbonval, “Audio source separation with a single sensor”, IEEE Trans. Audio, Speech and Language Proc., vol. 14, no. 1, January 2006, pp 191--199.

[10] E. Vincent, R. Gribonval and C. Févotte. “Performance measurement in Blind Audio Source Separation,” *IEEE Trans. Speech and Audio Processing*, vol. 14, no. 4, pp. 1462-1469, Jul. 2006.

[11] http://bass-db.gforge.inria.fr/bss_eval/