2011-01-01

Clustering NMF Basis Functions Using Shifted NMF for Monaural Sound Source Separation

Rajesh Jaiswal
Dublin Institute of Technology

Derry Fitzgerald
Dublin Institute of Technology, derry.fitzgerald@dit.ie

Dan Barry
Dublin Institute of Technology, dan.barry@dit.ie

Eugene Coyle
Dublin Institute of Technology, eugene.coyle@dit.e

Scott Rickard
University College Dublin

Recommended Citation

Jaiswal, R., Fitzgerald, D., Barry, D., Coyle, E., Rickard, S.: Clustering NMF Basis Functions Using Shifted NMF for Monaural Sound Source Separation. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) , Prague, 2011.

CLUSTERING NMF BASIS FUNCTIONS USING SHIFTED NMF FOR MONAURAL SOUND SOURCE SEPARATION

Rajesh Jaiswal* *Derry FitzGerald** *Dan Barry** *Eugene Coyle** *Scott Rickard*†

* Audio Research Group, Dublin Institute of Technology, Kevin Street, Dublin 8, Ireland
†Dept of Electronic Engineering, University College Dublin, Belfield, Dublin 4, Ireland

ABSTRACT

Non-negative Matrix Factorization (NMF) has found use in single channel separation of audio signals, as it gives a parts-based decomposition of audio spectrograms where the parts typically correspond to individual notes or chords. However, a notable shortcoming of NMF is the need to cluster the basis functions to their sources after decomposition. Despite recent improvements in algorithms for clustering the basis functions to sources, much work still remains to further improve these algorithms. To this end we present a novel clustering algorithm which overcomes some of the limitations of previous clustering methods. This involves the use of Shifted Non-negative Matrix Factorization (SNMF) as a means of clustering the frequency basis functions obtained from NMF. Results show that this gives improved clustering of pitched basis functions over previous methods.

***Index Terms*—** NMF basis functions, Shifted-NMF, Sound Source Separation, Constant Q spectrogram

1. INTRODUCTION

Monophonic sound source separation (SSS) refers to a process that separates out audio signals produced by various sound sources from a single channel audio mixture. Many audio applications like automatic music transcription, remixing, chord estimation and pitch modification would benefit from the availability of segregated sound sources from the mixture of audio signals for further processing. Furthermore, these methodology once implemented on single-channel music recordings can be extended to the upmixing from mono to stereo or 5.1 surround sound recordings.

Monophonic SSS typically uses a time-frequency representation of the signal, such as a spectrogram. The most commonly used is the Short-time Fourier Transform (STFT) which transforms a discrete-time signal $x(n)$ into a complex spectrogram ($\hat{\mathbf{X}}$). From this, a magnitude spectrogram ($\mathbf{X}$) is obtained for analysis. In recent years, many factorisation techniques of spectrograms have been proposed to separate out sources which include Non-negative Sparse Coding (NNSC) and NMF [2, 3].

NMF is a widely used factorisation technique [1] that has found application in the decomposition of audio spectrograms due to its ability to give additive parts-based decompositions, where the parts typically correspond to notes or chords in the music. NMF attempts to approximate the magnitude spectrogram $\mathbf{X}$ by decompositions into factors $\mathbf{A}$ and $\mathbf{B}$ such that the input data vector in $\mathbf{X}$ can be approximated by a linear combination of the column vectors of $\mathbf{A}$, basis vectors, and the corresponding activation weights of $\mathbf{B}$. The equation can be expressed as:

$$\mathbf{X} \approx \mathbf{A}\mathbf{B} \tag{1}$$

where $\mathbf{A}$ is $n \times r$ matrix and $\mathbf{B}$ is $r \times m$ matrix, with $r < n, m$. The commonly used cost function for finding the basis vectors using NMF is the Kullback-Leibler (KL) divergence as proposed in [1]:

$$D(A||B) = \sum_{i,j}(A_{ij} log \frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij}) \tag{2}$$

An advantage of NMF is that there can be a single basis function for each note played by a given instrument, thereby capturing changes in timbre with pitch for each instrument or source. However, this results in multiple basis functions per instrument and so the number of basis functions is greater than the number of sources. Therefore, clustering of these basis functions is required for separating the sources. Supervised clustering methods have been discussed in [4] to map the separated signals into sources. Spiertz and Gnann [5] have implemented unsupervised clustering of separated basis functions by mapping the basis functions to the Mel frequency cepstral domain where clustering is performed. While this method represents a considerable advance over previous methods, there is still room for improvement in clustering the basis functions to sources.

In an effort to avoid the need for clustering of basis functions, FitzGerald et al, proposed an algorithm [6], Shifted Non-negative Matrix Factorisation (SNMF) which assumes that timbre of a note is constant for the entire range of pitch produced by an instrument. Thus, a translated instrument basis function $\mathcal{D}$ can be used to approximate the spectra of all notes played by the instrument in consideration. The activation of each of the shifted basis functions is denoted by $\mathcal{H}$. With this assumption, a logarithmic frequency resolution of the spectrogram is required to exploit the shift-invariant property of the instrument basis functions. This can be achieved by a Constant Q transform (CQT) [7]. According to the even tempered chromatic scale [9], the pitch of each half tone is spaced by a factor of twelfth root of 2 ($\sqrt[12]{2}$). As a result, one note can be used to approximate another note a half tone higher or lower by translating the frequency basis function of the note up or down by one frequency bin, if a semitone bin spacing is used.

The outline of the paper is as follows. Section 2 gives an overview of the proposed SNMF clustering algorithms for NMF basis functions. Simulation experiments are explained in Sections 3 followed by the discussion of results in Section 4.

This work is funded by ABBEST PhD Research Scholarship at Dublin Institute of Technology, Dublin, Ireland.

2. SYSTEM OVERVIEW

2.1. Shifted NMF

The parameters used in the SNMF model [6] are defined as per the conventions used in [12]. A tensor of any dimension is donated by calligraphic upper case letters, such as $\mathcal{R}$. Indexing of elements within a tensor is donated by $\mathcal{R}(i, j)$. A contraction of two tensors of finite dimension spaces is defined as a bilinear mapping of the elements of two tensors into a new dimension space. Let a tensor $\mathcal{R}$ be of dimension $I_1 \times \cdots \times I_S \times L_1 \times \cdots \times L_P$ and tensor $\mathcal{D}$ be of dimension $I_1 \times \cdots \times I_S \times J_1 \times \cdots \times J_N$ then equation 3 denotes the contracted tensor multiplication of $\mathcal{R}$ and $\mathcal{D}$ along the first S modes.

$$\langle \mathcal{R}\mathcal{D}\rangle_{\{1,\ldots,S;1,\ldots,S\}} = \sum_{i_1=1}^{I_1} \cdots \sum_{i_1=1}^{I_1} \mathcal{R} \times \mathcal{D} = \mathcal{Z} \tag{3}$$

The subscripts specified in the curly brackets indicates the dimensions along the tensors $\mathcal{R}$ and $\mathcal{D}$ respectively to be multiplied together. The resultant tensor $\mathcal{Z}$ will be of dimension $L_1 \times \cdots \times L_P \times J_1 \times \cdots \times J_N$.

For r sources in a Constant Q spectrogram $\mathbf{Y}$ of size $n \times m$, where n is the number of frequency bins and m is the number of time frames, can be approximately decomposed as:

$$\mathbf{Y} \approx \langle\langle \mathcal{R}\mathcal{D}\rangle_{\{3,1\}} \mathcal{H}\rangle_{\{2:3,1:2\}} \tag{4}$$

where $\mathcal{R}$ is a translation tensor of dimension $n \times k \times n$ for k possible translations. $\mathcal{R}$ translates the instrument basis functions in $\mathcal{D}$ up or down to approximate various notes played by an instrument in question. Tensor $\mathcal{D}$ of size $n \times r$ contains frequency basis functions for each source. $\mathcal{H}$ is a tensor of size $k \times r \times m$ such that $\mathcal{H}(i\ ,\ j\ ,\ :)$ represents the time envelope for the i^{th} translation of the j^{th} source, which informs when a given note is played by a given instrument.

Unfortunately, the assumption that the timbre of any note played by an instrument, does not change with pitch does not hold. The spectral envelope of a note changes with pitch. This change in timbre should be dealt with to recover the correct timbre of the instrument which will result in improved sound separation. Also, for synthesis of the sound sources in [6], an approximate mapping from constant Q to linear spectrogram was required. As a result, the separation quality is compromised. Despite these shortcomings, the algorithm proved successful in separating simple mixtures of pitched instruments, though at reduced sound quality.

However, due to the shift-invariant properties of the algorithm, it can potentially be used to cluster the individual note basis functions obtained from a standard NMF decomposition. This use of SNMF is explored in the remainder of the paper.

2.2. Clustering of Basis Functions using SNMF

Having obtained a set of basis functions using NMF, SNMF can be used to cluster the basis functions as follows. The matrix $\mathbf{A}$, which contains the frequency basis functions, is treated as a type of spectrogram. This matrix is transformed to the Constant Q domain:

$$\mathbf{C} = \mathbf{T}\mathbf{A} \tag{5}$$

where $\mathbf{T}$ is a transform matrix used for mapping of the linear-frequencies to the Constant Q domain.

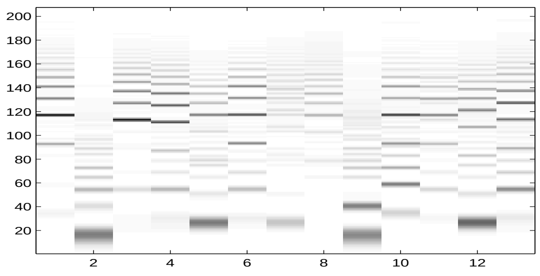
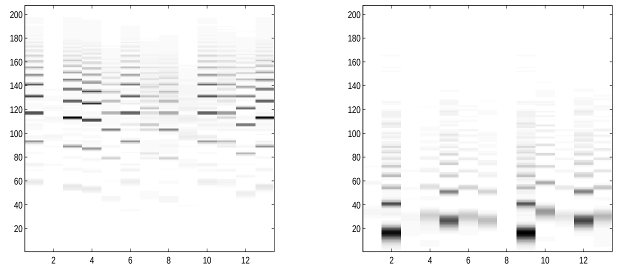


Fig. 1. NMF basis function of input mixture in constant Q domain.



(a) Source 1

(b) Source 2

Fig. 2. Clustering of NMF basis function in constant Q domain.

The spectrum $\mathbf{C}$ is then passed as input to SNMF:

$$\mathbf{C} \approx \langle\langle \mathcal{R}\mathcal{D}\rangle_{\{3,1\}} \mathcal{H}\rangle_{\{2:3,1:2\}} \tag{6}$$

Given the number of sources r, then SNMF will look for instrument basis functions that can be used to approximate $\mathbf{C}$. The number of translations k of an instrument basis function is appropriately chosen to cover the melodic range of the instrument. The cost function used to optimize tensors $\mathcal{D}$ and $\mathcal{H}$ in equation (4) is the same as described in [1] for standard NMF. Therefore, the multiplicative updates for the decomposition of spectrogram $\mathbf{C}$ can be derived as done in [6]. Assuming that each basis function in $\mathbf{C}$ corresponds to an individual note played by an individual instrument, the activations of the SNMF model should indicate which basis functions in $\mathbf{C}$ are associated with an individual source. To this end, we have introduced two different approaches used for clustering the basis functions.

2.3. Winner Takes All

Here, we will define how the shift-invariant property of SNMF model is used to cluster the NMF basis functions. In practice, clustering is carried out by reconstructing the individual source spectrograms $\mathbf{C}_r$ and comparing the energy in each source at each frame, where r is the number of sources. As mentioned in section 2.1, one of the limitations of the SNMF method is that, in general, a single basis function is not sufficient to cover all the notes played by the given instrument as the frequency information (timbre) changes with pitch. To overcome this drawback,we used a novel approach to cluster the basis functions in $\mathbf{C}$. After the optimization of tensors $\mathcal{D}$ and $\mathcal{H}$, the source spectrograms $\mathbf{C}_r$ is constructed by using the slices of tensors, $\mathcal{D}(:, r)$ and $\mathcal{H}(:, r, :)$, associated with the given source.

Furthermore, the energy of individual frame in each spectrogram $\mathbf{C}_r$ is compared with the corresponding frame of the other sources and the basis function in the original matrix $\mathbf{C}$ is allocated to the source which has the highest energy at that frame. The resultant $\mathbf{C}_r$ for each source can be combined into tensor denoted by $\mathcal{W}$ of size $n \times k \times r$.

$$\mathbf{C}_r = \mathcal{W}(:,:,r) = \langle\langle \mathcal{R}\mathcal{D}(:,r)\rangle_{\{3,1\}} \mathcal{H}(:,r,:)\rangle_{\{2:3,1:2\}} \qquad (7)$$

Let E be a energy matrix of size $k \times r$, then the summation of each frame along n frequency bins of tensor $\mathcal{W}$ can be represented by equation 8. Subsequently, the basis functions are indexed by δ_k corresponding to the sources hence clustering of the Constant Q basis functions.

$$E(k,r) = \sum^{n} \mathcal{W}(:,k,r) \qquad (8)$$

$$\delta_k = argmax\,(E(k,:)) \qquad (9)$$

Figure 1 shows the NMF basis functions in Constant Q domain of a input mixture of two sources. Figure 2 shows the separated basis function of source 1 and source 2 respectively. The x-axis shows the number of basis functions for individual notes to cover the highest pitch range played by the instrument in the test mixture. The figure shows the clear separation of basis functions associated with the different sources, hence these clustered basis function can be used to segregate the sources in question.

SNMF model requires the use of Constant Q transform to obtain the log-frequency resolution. Therefore, another drawback of using shift invariance property this SNMF model is the need of inverse CQT to transform the log-frequencies to obtain corresponding linear-frequencies. However, in this case there is a one to one correspondence between the basis function in $\mathbf{C}$ and $\mathbf{A}$. Therefore, the clustering obtained for $\mathbf{C}$ is equally valid for clustering $\mathbf{A}$ which can be further partitioned into individual $\mathbf{A}_r$, where $\mathbf{A}_r$ denotes the basis functions associated with the r^{th} source.

2.4. SNMF Masking

An alternate approach to map each $\mathbf{C}_r$ back in linear domain yielding $\mathbf{A}_r$ is also implemented. $\mathbf{T}^{'}$ (see equation 5) is multiplied with source spectrograms $\mathbf{C}_r$ to obtain corresponding $\mathbf{A}_r$.

$$\mathbf{A}_r = \mathbf{T}^{'}\mathbf{C}_r \qquad (10)$$

The recovered source frequency basis functions $\mathbf{A}_r$ are used to generate a mask which is applied to $\mathbf{A}$. In this case, the individual $\mathbf{A}_r$ are used to create individual source filters. Then, $\mathbf{A}$ is passed through these filters to obtain the source frequency basis functions $\hat{\mathbf{A}}_r$. A masking parameter $\hat{\mathbf{A}}_\mathbf{r}$ is calculated using the following equation:

$$\hat{\mathbf{A}}_r = \mathbf{A}\left(\frac{\mathbf{A}_r^p}{\sum^r \mathbf{A}_r^p}\right) \qquad (11)$$

$$\mathbf{X}_r = \hat{\mathbf{A}}_r\mathbf{B}_r \qquad (12)$$

where the power parameter p is set to 2 ($p = 2$) for this algorithm. Note that, in equations 11 and 13, all operations are done elementwise. As a result, NMF basis functions for each source are separated in the linear spectral domain. Then, each individual magnitude spectrogram $\mathbf{X}_r$ is retrieved using equation 12.

For both methods, resynthesis is carried out by using the separated source spectrograms, denoted by $\mathbf{X}_r$, to mask the original complex valued spectrogram, in the manner shown below:

$$\hat{\mathbf{X}}_r = \hat{\mathbf{X}}\left(\frac{\mathbf{X}_r^p}{\sum^r \mathbf{X}_r^p}\right) \qquad (13)$$

The recovered complex spectrograms are then inverted to the time domain using the inverse STFT.

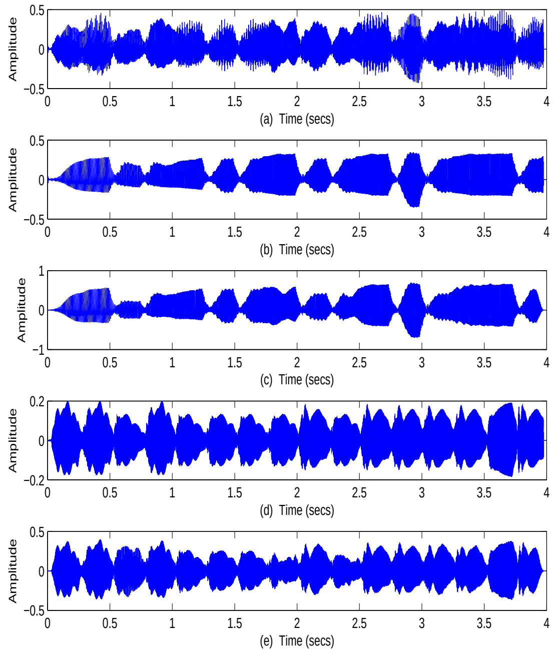


Fig. 3. Original and synthetic signals in the time domain. The figure shows (a) the original test signal, (b) the original source 1, (c) the reconstructed source 1, (d) the original source 2, (e) the reconstructed source 2.

3. EXPERIMENTS

The algorithm was implemented in Matlab for single channel audio mixtures to separate out sound sources. The SNMF model was tested for 25 monaural input mixtures of 2 instruments from a total of 15 different orchestral instruments taken from a sample library [11] including brass, woodwind and strings. The test set were of roughly 4 to 8 seconds in length with a sampling frequency of $44.1kHz$. To imitate the real world melodies, the notes played by individual instruments in the input mixture were in harmony and covered pitch from as low as $87Hz$ to $1500Hz$. The source signals were mixed with unity gain for the single channel mixture. More details on how the database was created can be found in [8].

The magnitude spectrogram of the time-domain signal were obtained using the STFT with a 75% overlapping Hann window, 4096 samples in length. Although, the number of NMF basis functions used in the algorithm can vary from 10 to 16 to improve the quality of separation, 13 basis functions were used for all the tests. NMF was run for 300 iterations. Constant Q transform used 24 frequency bins per octave covering frequencies ranging from $55Hz$ to $22.05kHz$. Before passing the Constant Q spectrogram in SNMF model both $\mathcal{D}$ and $\mathcal{H}$, in equation 4, are initialised randomly with positive values. As discussed in section 2.2 the cost function used for SNMF decomposition is the commonly used KL divergence as in equation 2. The multiplicative updates and positive initialization for $\mathcal{D}$ and $\mathcal{H}$ ensures the positive tensor factorisation. The algorithm is set for number of sources equal to 2 and it ran for 50 iterations. The number of time shifts i.e. allowable translations, k, was set to 7. The linear domain basis functions were then reconstructed using the techniques outlined in sections 2.3 and 2.4.

Examples of audio waveforms in time domain are shown in Figure 3. The waveforms of the original and synthesized music signals were found to match closely for pitched music signals. It can be seen through visual inspection of the waveforms that the sources have been separated well with a small interference of one source's melody on the other and vice versa. This shows that the algorithm can be used for separating pitched instruments for monaural mixtures. However, the proposed clustering method is sensitive to the chosen number of basis functions which defines all the notes played by the instruments in the mixture. The quality of separation is evaluated in the following section.

4. RESULTS

The performance of the SNMF algorithms were evaluated using the quality measures signal-to-distortion ratio (SDR), the signal-to-interference ratio (SIR), and the signal-to-artifacts ratio (SAR). These measures are widely used for the evaluation of separation quality and the details of these matrices can be found in [10]. SDR determines the overall sound quality of the recovered signal, SIR measures the interference of other sources in the separates sound source and SAR calculates the artifacts present in separated signal. The original source signals were used as a reference for the performance evaluation.

clustering	SDR	SIR	SAR
C_{MFCC}	0.80	10.96	3.30
C_{NMF}	2.89	12.72	4.59
C_{SNMF}	5.40	15.27	6.90
$C_{SNMF,mask}$	8.94	23.69	9.72

Table 1. Mean SDR, SIR and SAR for separated sound sources using SNMF clustering

C_{NMF} and C_{MFCC} are the two other clustering methods used for comparison. The algorithm for C_{NMF} and C_{MFCC} are implemented as documented in [5]. C_{NMF} represents NMF clustering with divergence cost function and C_{MFCC} is MFCC (Mel frequency cepstrum coefficient) clustering. All the clustering algorithms are tested by the same set of input mixtures to compare the results. All the results for mean SDR, SIR and SAR are shown in dB. The performance of two proposed clustering algorithm C_{SNMF}, Shifted NMF with one-to-one mapping, and $C_{SNMF,mask}$, shifted NMF with masking, are shown in the Table 1. It can be seen from the data that masking method gave better results than '*winner takes all*' SNMF model. It is also evident from the Table 1 that both the proposed clustering algorithms C_{SNMF} and $C_{SNMF,mask}$ in the paper outperform these other clustering techniques. We tested the clustering algorithm discussed in [4] for the same set of audio mixtures. However, the results were poor and so were not included.

5. CONCLUSION

In this paper, we presented two SNMF based clustering algorithms for single channel blind source separation which used tensor factorisation to separate the sound sources. We dealt with the change in the timbre with pitch by assigning separate basis function for each note being played by the individual instruments. For the first algorithm, we used one-to-one mapping from Constant Q domain to linear spectrogram to eliminate the need of inverse Constant Q transform. Alternatively, we used an approximate inverse transform followed by masking of the original spectrogram (containing basis functions) with the recovered basis functions to obtain the clustered basis function in linear domain. We tested the algorithm on various test input mixtures of two sources. The tests show a significant improvement on the sound quality as compared to unsupervised clustering done by Spiertz [5]. Furthermore, these clustering algorithms can be extended for input mixtures of n sources. Therefore, clustering using SNMF is an effective way to cluster pitched basis function to separate out harmonic instruments.

6. REFERENCES

[1] D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorisation, " *Advances in Neural Information Processing System*, 2000, pp. 556-562.

[2] S. A. Abdullah and M. D. Plumbley, "Polyphonic transcription by non-negative sparse coding of power spectra," *Proceedings of the* 5^{th} *International Conference on Music Information Retrieval (ISMIR 2004)*, Barcelona, Spain, Oct. 2004.

[3] P. Smaragdis, and J. C. Brown, "Non-negative matrix factorisation for polyphonic music transcription," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 2003, pp. 177-180.

[4] T. Virtanen, "Sound Source Separation Using Sparse Coding with Temporal Continuity Objective," *International Computer Music Conference*, 2003.

[5] M. Spiertz and V. Gnann, "Source-Filter based clustering for monaural blind source separation," *Proceedings of the* 12^{th} *International Conference on Digital Audio Effects*, Italy, 2009.

[6] D. FitzGerald, M. Cranitch and E. Coyle, "Shifted Non-negative matrix factorisation for sound source separation," *IEEE Workshop of Statistical Signal Processing, Bordeaux*, France, 2005.

[7] J. C. Brown, "Calculation of a Constant Q spectral transform," *Journal of the Acoustic Society of America*, vol. 89, no.1, pp 425-434, 1991.

[8] D. FitzGerald, M. Cranitch and E. Coyle, "Extended Non-negative Tensor Factorisation Models for Musical Sound Source Separation," *Computational Intelligence and Neuroscience*, Hindawi Publishing Corp., 2008.

[9] E. M. Burns, "Intervals, Scales and Tuning," *The Psychology of Music*, D. Deutsch, Ed. Academic Press, 1999.

[10] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, 2006.

[11] P. Siedlaczek, "Advanced Orchestra Library Set," 1997.

[12] B. W. Bader and T. G. Kolda, "Algorithm 862: MATLAB tensor classes for fast algorithm prototyping," *ACM Transactions on Mathematical Software*, vol. 32, no. 4, pp. 635-653, 2006.