APPLICATION OF NON-NEGATIVE MATRIX FACTORIZATION TO SIGNAL-ADAPTIVE AUDIO EFFECTS

Ryan Sarver
Centre for Digital Music
Queen Mary University of London, UK
rpsarver@gmail.com

Anssi Klapuri
Centre for Digital Music
Queen Mary University of London, UK
anssi.klapuri@eecs.qmul.ac.uk

ABSTRACT

This paper proposes novel audio effects based on manipulating an audio signal in a representation domain provided by non-negative matrix factorization (NMF). Critical-band magnitude spectrograms $\mathbf{Y}$ of sounds are first factorized into a product of two lower-rank matrices so that $\mathbf{Y} \approx \mathbf{BG}$. The parameter matrices $\mathbf{B}$ and $\mathbf{G}$ are then processed in order to achieve the desired effect. Three classes of effects were investigated: 1) dynamic range compression (or expansion) of the component spectra or gains, 2) effects based on rank-ordering the components (colums of $\mathbf{B}$ and the corresponding rows of $\mathbf{G}$) according to acoustic features extracted from them, and then weighting each component according to its rank, and 3) distortion effects based on controlling the amount of components (and thus the reconstruction error) in the above linear approximation. The subjective quality of the effects was assessed in a listening test.

1. INTRODUCTION

Audio effects can be viewed as processing modules that take in an audio signal and modify it according to certain control parameters to produce the desired audio output [1]. Typical examples include dynamic range compression, reverberation, and non-linear distortion for the electric guitar. The widespread use of audio effects in recorded music motivates the creation of new types of effects that produce musically interesting results and can be controlled by intuitive parameters.

During the last ten years, non-negative matrix factorization (NMF) has been actively studied for the purposes of audio content analysis [2, 3, 4, 5]. However, the potential of NMF for digital audio effects has not been properly investigated. NMF decomposes an input signal into a set of "components" that often correspond to physically distinct sources or sound events, and thereby opens a way towards applying effects on each source separately. For example, dynamic range compression can be applied on each component, instead of compressing the wideband signal or the signals within fixed subbands. In this paper, we propose three different strategies for manipulating an audio signal in the representation domain provided by the NMF before resynthesizing it back to a time-domain waveform. The results were evaluated in a listening test where the subjects described the differences they heard between the affected samples and the original ones and gave their opinions on whether the effect was interesting and useful. Overall, the results were positive and encourage further work in this area. Audio examples of the proposed effects are available at http://www.elec.qmul.ac.uk/people/anssik/NMFeffects/

2. METHOD

2.1. Data Representation

The effects discussed in this paper are based on factorizing the magnitude spectrograms of audio signals. The short-time Fourier transform (STFT) of a time-domain signal $x(n)$ is first calculated as

$$X_t(k) = \sum_{n=0}^{N-1} x(tH+n)w(n)e^{-j2\pi kn/N}, \quad (1)$$

where t is frame index, k is frequency index, N is the frame size, $H = N/2$ is the frame hop, and $w(n)$ is the hamming window.

The frequency resolution of STFT is linear, whereas the human auditory system carries out frequency analysis on a nonlinear scale. The equivalent rectangular bandwidths b_c of the critical bands in human hearing are given by [6]

$$b_c = 0.108 f_c + 24.7 \text{ Hz}, \quad (2)$$

where f_c and b_c denote the center frequency and bandwidth of critical band ("channel") c, and $c = 0, 1, \ldots, C-1$. The bandwidth b_c can be viewed as the frequency resolution of the peripheral auditory system at frequency f_c.

The perceptual quality of the audio effects obtained using NMF is greatly improved by warping the linear frequency resolution of the STFT to a critical-band resolution. This is achieved by simulating a bank of critical-band bandpass filters in the frequency domain. The center frequencies f_c of the filters that we use are distributed uniformly on the critical band scale (obtained by integrating the inverse of (2)),

$$f_c = 229 \left[10^{(a_1 c + a_0)/21.4} - 1 \right], \quad (3)$$

where $a_0 = 1.5$ determines the center frequency of the lowest band (40 Hz) and $a_1 = 0.79$ determines the band density in critical bandwidth units. We use a total of $C = 50$ subbands between 40 Hz and 20 kHz.

Warping from a linear frequency scale to the critical band scale is achieved using triangular sub-band responses (basis functions) that assign appropriately weighted STFT frequency bin values to the corresponding critical-band spectrogram bins. The basis functions are stored as rows in matrix $\mathbf{W}$ which maps the STFT magnitude spectrogram $|\mathbf{X}|$ of size $(K \times T)$ to a critical-band spectrogram $\mathbf{Y}$ of size $(C \times T)$ by

$$\mathbf{Y} = \mathbf{W}|\mathbf{X}|. \quad (4)$$

Figure 1 illustrates the structure of the basis matrix $\mathbf{W}$.

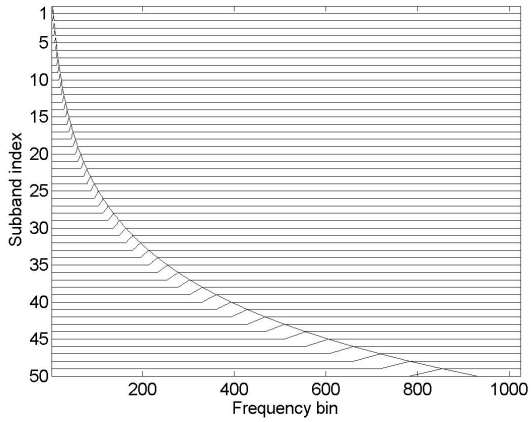

Figure 1: Illustration of the contents of the basis matrix **W** used to warp from a linear frequency scale to a critical band scale.

2.2. Non-negative Matrix Factorization

The idea of NMF is to approximate a non-negative matrix $\mathbf{Y} \in \mathbb{R}_+^{C\times T}$ as a product of two lower-rank (that is, smaller) matrices $\mathbf{B} \in \mathbb{R}_+^{C\times Z}$ and $\mathbf{G} \in \mathbb{R}_+^{Z\times T}$:

$$\mathbf{Y} \approx \mathbf{BG}. \tag{5}$$

The columns of matrix **B** contain the spectra of individual components z, $z = 1, 2, \ldots, Z$, and the rows of matrix **G** contain the corresponding time-varying gains. The number of columns in **B** (and rows in **G**) is here denoted by Z and determines the number of components that **Y** is broken into. Since magnitude spectra are inherently non-negative, a non-negativity restriction can be placed on these matrices [4]. Since the power spectra of many natural sounds (such as drum hits or individual notes) remains quite consistent across different occurrences, the factorization (5) often results in the separation of meaningful sound sources [4].

The algorithm we used for learning **B** and **G** is based on minimizing the Kullback-Leiber divergence between **Y** and **BG**. The algorithm works by initializing **B** and **G** with random positive values and updating them iteratively with multiplicative rules until the algorithm converges [2]. The value of the cost function is decreased at each update until a local minimum is reached. The update rules for **B** and **G** are given by

$$\mathbf{B} \leftarrow \mathbf{B}.\times \frac{(\mathbf{Y}./\mathbf{BG})\mathbf{G}^T}{\mathbf{1G}^T} \tag{6}$$

$$\mathbf{G} \leftarrow \mathbf{G}.\times \frac{\mathbf{B}^T(\mathbf{Y}./\mathbf{BG})}{\mathbf{B}^T\mathbf{1}} \tag{7}$$

where $\mathbf{1}$ is a K-by-T matrix of ones, and $.\times$ and $./$ denote element-wise multiplication and division, respectively [2].

2.3. Resynthesis

The proposed audio effects are based on manipulating the matrices **B** and **G** before resynthesis. Before discussing the actual effects, however, let us consider the resynthesis of a time-domain signal from the NMF representation.

2.3.1. Direct Resynthesis from the Linear Model

The most straightforward way of resynthesis is based on the linear signal model of NMF directly:

$$\hat{\mathbf{Y}} = \mathbf{BG} \tag{8}$$

This is followed by a warping of the critical-band scale back to the linear frequency scale, achieved using a transpose of the matrix of basis functions **W**:

$$|\hat{\mathbf{X}}| = \mathbf{W}^{\mathsf{T}}\hat{\mathbf{Y}} \tag{9}$$

The resulting magnitude spectrogram is combined with the phase spectrogram of the original mixture signal. Finally, inverse Fourier transform of each frame and 50% overlap-add is performed to obtain a time-domain signal.

2.3.2. Perfect Reconstruction Resynthesis

Synthesising a time-domain signal using (8) leads to inevitable distortion if the number of components Z is insufficient to represent the input audio spectrogram accurately. A typical requirement for audio effects is that the user can control the amount of effect applied on the input signal, and when this "effect depth" parameter is set to zero, the output signal is identical to the input signal (perfect reconstruction).

Perfect reconstruction resynthesis is achieved by reconstructing the complex-valued STFT spectrogram of component z by

$$\mathbf{X}_z = \left[\mathbf{W}^{\mathsf{T}}\left(\frac{\mathbf{b}_z\mathbf{g}_z}{\mathbf{BG}}\right)\right].\times\mathbf{X} \tag{10}$$

where $\mathbf{b}_i$ and $\mathbf{g}_i$ denote the zth column of **B** and the zth row of **G**, respectively, and **X** is the complex-valued STFT spectrogram of the input signal. This is one form of the Wiener filter and leads to perfect reconstruction of the complex-valued STFT spectrogram of the input signal by

$$\mathbf{X} = \sum_z \mathbf{X}_z \tag{11}$$

Inverse Fourier transform of **X** followed by overlap-add can then be used to reconstruct the original input signal.

2.4. Audio Effects in the "NMF Domain"

The effects proposed in this paper are based on processing the parameter matrices **B** and **G** before resynthesizing the signal. For convenience in the following, we use the term "NMF domain" to refer to the parametric representation (5) of the input signal provided by the NMF.

2.4.1. Dynamic Range Compression and Expansion

Dynamic range compression and expansion involve multiplying the input signal by a slowly-varying gain factor that depends on the level of the input signal [7]. The operation of a dynamic range controller is typically described using a piece-wise linear curve that defines the desired output level (in decibels) as a function of the input level (in decibels). If the slope of this curve is $\frac{1}{3}$, for example, any change ΔL_i in the input level is mapped to a three times smaller change ΔL_o in the output level and the corresponding compression ratio $R = \Delta L_i/\Delta L_o$ would be 3. The term compression refers to $R > 1$ and expansion to $R < 1$.

A straightforward implementation of compression in the NMF domain can be achieved by raising the gains $g_z(t) \equiv \mathbf{g}_z$ of component z to power $1/R$. If the same amount of compression or expansion is to be applied on all components, then all elements of the matrix $\mathbf{G}$ are raised to power $1/R$. For example, compression by factor 3 is achieved by raising all elements of $\mathbf{G}$ to power $1/3$. This can be viewed as compression/expansion without a threshold (i.e., there is no threshold level below which the effect would be switched off).

Intuitively, compressing the component gains brings the less prominent sounds (at a given time) more to the foreground, since individual components tend to capture physical sound events or sound sources on the recording.

In the experiments to be described in Section 3, we investigated dynamics processing of not only the gain matrix but also the spectral basis matrix $\mathbf{B}$. Compression of the spectral basis matrix $\mathbf{B}$ results in spectra of the individual components that are either compressed (flattened) or expanded.

2.4.2. Effects Based on Ordering the Components

The factorization achieved by NMF suffers from permutation ambiguity: the order of the components (columns of $\mathbf{B}$ and the corresponding rows in $\mathbf{G}$) is arbitrary and depends on the random initialization of the matrices $\mathbf{B}$ and $\mathbf{G}$ before applying the multiplicative updates (6)-(7). In this sense, the individual components have no "identity".

The class of effects described in this subsection is based on ordering the components $z = 1, 2, \ldots, Z$ according to acoustic features calculated from the component spectra and gains. The components are then weighted differently before resynthesis, depending on their position on the ordered list.

Two different ordering criteria were investigated: spectral centroid and kurtosis. Spectral centroid is here defined as the first moment of the spectrum of a given component and conveys information about the "brightness" of that component. The spectral centroid S_z of component z is given by:

$$S_z = \frac{\sum_c f_c \mathbf{b}_z(c)}{\sum_c \mathbf{b}_z(c)} \tag{12}$$

where f_c is the center frequency of the critical band corresponding to bin c of matrix $\mathbf{B}$ and is given by (3).

The spectral centroids S_z were then utilized to produce an audio effect by weighting component z by $(r_z - 1)/(Z - 1)$ before resynthesis. Here r_z denotes the rank of component z on a list where components are sorted in either ascending or descending-centroid order.

Another criterion that we used for ordering the components was the kurtosis. The kurtosis of the gain function $g_z(t)$ of component z is given by

$$K_z = \frac{\frac{1}{T}\sum_{t=1}^{T}(g_z(t) - \bar{g}_z)^4}{\left(\frac{1}{T}\sum_{t=1}^{T}(g_z(t) - \bar{g}_z)^2\right)^2} - 3 \tag{13}$$

where $\bar{g}_z$ denotes the empirical mean of $g_z(t)$. Note that the term -3 has no effect on the order and can be discarded.

We investigated ordering the components according to the kurtosis of their gains as well as the kurtosis of their spectra. Kurtosis of the gains function characterizes the "transientness" of a component (peakiness of its gains). Kurtosis of the spectrum, in turn, tends to be higher for harmonic spectra (components representing musical notes) than for "noisy" spectra (components representing drum sounds for example). Similarly to the ordering based on spectral centroid, components were then scaled by a weight between 0 and 1 depending on their rank on the sorted list of components formed according to the kurtosis value.

2.4.3. Distortion as an Effect

The third class of effects is based on a controlled use of the reconstruction error caused by a direct resynthesis from the NMF model as described in Section 2.3.1. The distortion resulting from the NMF decomposition sometimes produces interesting effects in itself as will be discussed in the Results section. The effect was presented by cross-fading from the clean input signal to a signal reconstructed from an NMF model with eight components. This was then further cross-faded to a signal obtained using four, two, and finally just one component, and then back to the clean signal in the opposite order. The number of components used controls the amount of distortion introduced.

3. RESULTS

As the success of an audio effect cannot be assessed objectively, we conducted a listening test where the subjects rated and described the effects they heard. The current implementation of the method is non-causal and requires off-line processing of the input signals. Therefore the parameters of the effects in the listening test had to be fixed and the test stimuli calculated in advance, as opposed to allowing the subjects to tune the parameters in real-time. We chose parameters and music clips that were thought to be representative and interesting examples of each class of effects. The samples and the used parameter values are available on-line at http://www.elec.qmul.ac.uk/people/anssik/NMFeffects/

3.1. Stimuli

Four clips of music were chosen that were thought to best exemplify the investigated effects. The clips were from *Smells Like Teen Spirit* by Nirvana, *Billie Jean* by Michael Jackson, *Come Together* by the Beatles, and *I Turn My Camera On* by Spoon. These span music from hard rock to pop and years from the 60s (*Come Together*) up to a few years ago (*I Turn My Camera On*). For each of the four clips, four effects were presented: compression/expansion of spectra and/or gain curves, scaling of NMF components based on their spectral centroid, scaling the components based on the kurtosis of their spectra or gains, and the proposed distortion effect. Therefore the stimuli consisted of a total of twenty clips including the four original versions and all of the effects.

With the compression/expansion there was obviously a choice between compression and expansion, but there was additional variability in that either could be done to the spectra, the gain curves, or both. These different combinations resulted in drastically different effects. As it would be infeasible to have a sample for each possible combination, suitable parameter combinations were chosen subjectively to exemplify the possibilities of each effect type.

3.2. Subjects

There were ten subjects in total, eight male and two female, aged between 22 and 41. Seven of the subjects were musicians and three

Table 1: *Overall results (%) of whether the subjects found the effects interesting and would use them if they were available.*

	Interesting	Would use
Distortion	80	64
Comp/Exp	80	59
Spec Cent	55	24
Kurtosis	73	39

Table 2: *Overall results (%) of the subjects ranking the effects from the most to the least interesting.*

	Most	2nd Most	3rd Most	Least
Distortion	45	22.5	12.5	20
Comp./Exp.	22.5	32.5	27.5	17.5
Kurtosis	27.5	27.5	25	20
Spec Cent	10	12.5	35	42.5

were not. Seven out of the ten said they were familiar with the term "audio effect" and how they are used, and six of them said they had experience using them.

3.3. Experimental Setup

The listening test was conducted completely on-line. The order of presentation was randomized for each clip. The first question asked simply whether the listener found the effect "interesting." The next question asked the listeners to describe in their own words the differences they heard between the original and the affected clips. The third question asked the listener whether they would be interested in using the effect were it available as a commercial product. After all of the effects were evaluated for each clip, the subject was asked to rank the four effects for that clip from the most to the least interesting.

3.4. Results

Table 1 shows the percentages of subjects that a) found the effects interesting and b) would consider purchasing or using the effect if it were available to them. For the latter, responses were excluded from subjects who reported "I don't regularly use audio effects". In each case the majority of the subjects found the effect to produce interesting results, albeit a slight majority in the case of the spectral centroid effect. A majority of the subjects who use audio effects would be interested in using the distortion and compression/expansion effects. This was not the case, however, with the spectral centroid and kurtosis effects.

Table 2 shows how people ranked the effects from the most to the least interesting. Again the results have been averaged over all the four clips. It is clear that the distortion effect leans towards being the one considered most interesting and the spectral centroid effect the least, with the compression/expansion and kurtosis effects having a fairly even spread.

When describing the differences the subjects heard between the original and affected versions it was common for the subjects to describe the spectral centroid effect as sounding like a simple filter was applied. This would explain the poor results, as listeners familiar with audio effects might find it trivial. On the other hand, the subjects generally seemed to be intrigued by the distortion effect, as if it were something they had never encountered before.

The responses for the compression/expansion and kurtosis effects were more mixed but still generally positive, and this was also reflected in the written responses. These effects, similar to the distortion effect, found the subjects coming up with more sophisticated descriptions of the things they heard, even in some cases stating that they could not really describe what was going on. For example with the compression/expansion effect for "Smells Like Teen Spirit" two separate responses were received in which the effect was described as making the clip sound more "industrial"; other responses described the clip as sounding like it was "recorded underwater" and "playing inside a can".

4. CONCLUSIONS

Non-negative matrix factorization provides a musically meaningful representation for audio signals that has not been fully utilized for audio effects. Three different types of effects were investigated in this paper: compression/expansion of component gains and/or spectra, scaling components based on ordering them according to extracted acoustic features, and distortion inherent to the NMF approximation. The distortion effect produced the best results, with the subjects consistently ranking it as one of the more interesting effects. The results concerning the compression/expansion and kurtosis ordering effects were fairly mixed but generally positive.

Future work involves a real-time implementation of the proposed effects. NMF is inherently a non-causal method since the component spectra and gains are estimated jointly. However, a causal (real-time) implementation can be achieved by keeping the spectral bases $\mathbf{B}$ fixed and updating only the gains $\mathbf{G}$ for each incoming audio frame. The component spectra are then updated only occasionally, for example every 5 seconds based on the preceding 10 second segment. A real-time implementation would be useful for more efficient exploration of the parameter space of the effects.

5. REFERENCES

[1] U. Zolzer, *DAFX - Digital Audio Effects*, J. Wiley & Sons, West Sussex, UK, 2002.

[2] D. Lee and H. Seung, "Algorithms for non-negative matrix factorization," in *Neural Information Processing Systems*, Denver, USA, 2001, pp. 556–562.

[3] Paris Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, USA, 2003.

[4] T. Virtanen, *Signal Processing Methods for Music Transcription*, chapter Unsupervised Learning Methods for Source Separation, pp. 267–298, Springer, NY, USA, 2006.

[5] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 538–549, 2010.

[6] B. C. J. Moore, Ed., *Hearing—Handbook of Perception and Cognition*, San Diego, California, 2nd edition, 1995.

[7] U. Zolzer, *Digital Audio Signal Processing*, J. Wiley & Sons, West Sussex, UK, 1997.