

UMA ABORDAGEM CONVOLUTIVA PARA A FATORAÇÃO DE MATRIZ NÃO-NEGATIVA EM DADOS AMOSTRADOS IRREGULARMENTE

FABIO LOUVATTI DO CARMO, EVANDRO OTTONI TEATINI SALLES

*Programa de Pós-Graduação em Engenharia Elétrica
Universidade Federal do Espírito Santo
Av. Fernando Ferrari, 514, Prédio CT VI, 29075-910
Vitória - Espírito Santo, Brasil*

Emails: `fabio.carmo@aluno.ufes.br`, `evandro@ele.ufes.br`

Abstract— Non-negative Matrix Factorization (NMF) has gained popularity in audio tasks due its easy implementation and parts-based separation properties. Many studies has been adding extensions in order to condition the algorithm to the real characteristics of the sound. However, as a matrix-based method, it finds difficult to deal with data at non-regular grid, such as constant-Q transform (CQT). It was shown it is possible to adapt NMF to treat such data. This paper presents a convolutive NMF approach to work with irregularly-sampled data using CQT as spectral representation, here defined as NRCNMF. Two tests were performed and the results show the effectiveness of the proposed technique.

Keywords— Non-negative Matrix Factorization, Monaural separation, Constant-Q Transform.

Resumo— Fatoração de matriz não-negativa (NMF) ganhou popularidade em tarefas de áudio devido à sua fácil implementação e propriedades de separação baseada em partes. Muitos estudos vêm adicionando extensões a fim de condicionar o algoritmo às características reais do som. Entretanto, por ser um método baseado em matrizes, encontra dificuldade em tratar dados em grade não-regular, como a transformada Q constante (CQT). Foi mostrado que é possível adaptar a NMF para tratar dados deste tipo. Neste artigo é apresentada uma abordagem NMF convolutiva para tratar amostras em grade não-regular, utilizando a CQT como representação espectral, definida aqui por NRCNMF. Foram realizados dois testes e os resultados obtidos mostram a eficácia da técnica proposta.

Palavras-chave— Fatoração de Matriz Não-Negativa, Separação monaural, Transformada Q constante.

1 Introdução

A tarefa de separação de múltiplos sons a partir de uma gravação de áudio monaural é um problema difícil que tem sido amplamente abordado na literatura. Usualmente, os métodos propostos são baseados na decomposição matricial do espectrograma do sinal de áudio. Um passo importante na análise desses dados é buscar uma representação adequada a fim de encontrar as estruturas latentes ali contidas. A Fatoração de Matriz Não-negativa (NMF - *Non-negative Matrix Factorization*) é uma técnica que ganhou popularidade em (Lee and Seung, 2001) devido à sua natureza esparsa no seu processo de decomposição (baseada em partes) e pela simplicidade na implementação de suas regras multiplicativas. NMF tem sido utilizada em aplicações em áudio, como separação de voz monaural (Smaragdis, 2007), identificação de objetos auditivos com espectro variante no tempo (Smaragdis, 2004), e transcrição automática de música (Smaragdis and Brown, 2003).

A formulação original da NMF define que uma matriz $\mathbf{X}_{N \times M} \in \mathbb{R}^+$ seja aproximada pelo produto de duas matrizes $\mathbf{W}_{N \times R} \in \mathbb{R}^+$ e $\mathbf{H}_{R \times M} \in \mathbb{R}^+$, tal que seja minimizado o erro de reconstrução. Tal erro pode ser medido através de diversas funções custo, tais como norma de Frobenius, divergência de Kullback-Leibler (Lee and Seung, 2001) e divergência de Itakura-Saito (Fevotte et al., 2009). Entretanto, essa formulação é fraca

para tarefas em áudio, pois não considera a informação temporal dos dados, não levando em conta as posições relativas dos espectros ao longo do tempo. Em (Smaragdis, 2004) foi introduzido um modelo NMF convolutivo, onde a relação entre múltiplas observações sobre intervalos de tempo próximos é descrita por uma sequência de matrizes de bases $\mathbf{W}_t$ que corresponde à matriz de coeficientes $\mathbf{H}$ variando ao longo do tempo.

Estes métodos tradicionais utilizam uma representação matricial do espectrograma, onde é imposto um espaçamento regular em tempo e frequência. A representação mais comumente utilizada é a transformada de Fourier de tempo curto (STFT - *Short Time Fourier Transform*). Entretanto, a STFT possui um espaçamento linear na frequência, e não mapeia eficientemente as frequências musicais, que são espaçadas geometricamente. Em contraste, a transformada Q constante (CQT - *Constant-Q Transform*), originalmente introduzido em (Brown, 1991), caracteriza uma resolução em frequência dependente das frequências centrais das janelas usadas para cada *bin* de frequência, e dessa forma, consegue-se um espaçamento geométrico. A CQT é importante no processamento de sinais de voz e música, pois é baseada na percepção humana do som. Em (M. Dörfler and Velasco, 2011) foi apresentado uma transformada CQT perfeitamente inversível e eficiente baseado em *frames* de Gabor não-estacionários.

Porém, a transformada retorna os dados amostrados em uma grade não-regular, não sendo possível a representação matricial.

Em (Smaragdis and Kim, 2013) foi introduzido uma formulação da NMF que pode ser aplicada em dados com índices reais, não tendo mais a necessidade de trabalhar apenas com representação matricial (índices inteiros). Com isso, é possível realizar tarefas de decomposição espectral utilizando representações como a CQT.

Neste artigo será introduzida uma abordagem convolutiva à NMF para dados irregularmente espaçados em tempo-frequência, a fim de considerar a dependência temporal dos dados, que é inerente aos sinais de áudio. Ademais, utilizando-se a CQT baseada em *frames* de Gabor não-estacionários como representação espectral e observando-se as características de espaçamento das amostras em tempo-frequência, foi possível adaptar a formulação do algoritmo, a fim de melhorar os resultados da fatoração.

Esse artigo se divide como segue. A seção 2 apresenta brevemente a reformulação da NMF para amostras espaçadas irregularmente (Smaragdis and Kim, 2013). A seção 3 propõe a abordagem convolutiva do algoritmo apresentado na seção 2. A seção 4 mostra duas aplicações de separação monaural, comparando o algoritmo da seção 2 com o proposto na seção 3. E a seção 5 traz as conclusões.

2 NMF para amostras espaçadas irregularmente

A formulação original da NMF define que a reconstrução dos dados seja da forma

$$\hat{\mathbf{X}} = \mathbf{W}\mathbf{H}, \quad (1)$$

onde $\hat{\mathbf{X}}_{N \times M} \in \mathbb{R}^+$ é uma matriz que contém a aproximação do espectrograma do sinal de áudio, $\mathbf{W}_{N \times R} \in \mathbb{R}^+$ e $\mathbf{H}_{R \times M} \in \mathbb{R}^+$ são as matrizes de bases e de coeficientes, respectivamente. A dimensão R é o número de partes que o espectrograma será fatorado.

Para dados amostrados irregularmente, (Smaragdis and Kim, 2013) propôs uma reformulação em que o espectrograma $\mathbf{X}$ é vetorizado¹, $\mathbf{x} = \text{vec}(\mathbf{X}) \in \mathbb{R}^+$, e são criados outros dois vetores, $\mathbf{t} \in \mathbb{R}^+$ e $\mathbf{f} \in \mathbb{R}^+$, da mesma dimensão de $\mathbf{x}$, que contém a localização em tempo-frequência de cada amostra. Com isso, a fatoração em (1) se torna

$$\hat{\mathbf{x}} = \sum_{r=0}^{R-1} \mathbf{v}_r \odot \mathbf{g}_r, \quad (2)$$

onde $\mathbf{v}_r \in \mathbb{R}^+$ e $\mathbf{g}_r \in \mathbb{R}^+$ correspondem aos valores de $\mathbf{W}$ e $\mathbf{H}$ avaliados nas posições de tempo

¹o operador $\text{vec}(\cdot)$ concatena os vetores para cada *bin* de frequência que a CQT retorna em um único vetor coluna.

e frequência em $\mathbf{t}$ e $\mathbf{f}$, e $\odot$ denota o produto matricial de Hadamard. Através dessa reformulação, as regras multiplicativas são

$$\begin{aligned} \mathbf{P} &= \frac{\mathbf{V} \odot \mathbf{G}}{(\mathbf{V} \odot \mathbf{G}) \cdot \mathbf{1}_{R \times R}} \\ \mathbf{V} &= \mathbf{D}_f \cdot (\mathbf{P} \odot (\mathbf{x} \cdot \mathbf{1}_{1 \times R})) \\ \mathbf{G} &= \mathbf{D}_t \cdot (\mathbf{P} \odot (\mathbf{x} \cdot \mathbf{1}_{1 \times R})), \end{aligned} \quad (3)$$

onde $\mathbf{P}$, $\mathbf{V}$ e $\mathbf{G}$ são matrizes de dimensão $K \times R$, e cada coluna é referente a uma variável latente r . Para o caso de amostras com índices inteiros, as matrizes $\mathbf{D}_t$ e $\mathbf{D}_f$ são definidas como

$$\begin{aligned} \mathbf{D}_f(i, j) &= \begin{cases} 1, & \mathbf{f}(i) = \mathbf{f}(j) \\ 0, & \mathbf{f}(i) \neq \mathbf{f}(j) \end{cases} \\ \mathbf{D}_t(i, j) &= \begin{cases} 1, & \mathbf{t}(i) = \mathbf{t}(j) \\ 0, & \mathbf{t}(i) \neq \mathbf{t}(j). \end{cases} \end{aligned} \quad (4)$$

No caso de índices reais, os vetores de tempo e frequência têm grandes chances de serem formados por elementos únicos, e dessa forma, a segunda e terceira equação em (3) não capturarão as correlações das posições em tempo e frequência. Para amenizar este problema, as matrizes $\mathbf{D}_t$ e $\mathbf{D}_f$ são geradas através de *kernels* gaussianos

$$\begin{aligned} \mathbf{D}_f(i, j) &= \exp\left(\frac{-|\mathbf{f}(i) - \mathbf{f}(j)|^2}{\sigma_f^2}\right) \\ \mathbf{D}_t(i, j) &= \exp\left(\frac{-|\mathbf{t}(i) - \mathbf{t}(j)|^2}{\sigma_t^2}\right). \end{aligned} \quad (5)$$

3 NMF convolutivo para amostras espaçadas irregularmente

Na NMF tradicional, cada objeto do espectrograma é descrito por seu padrão espectral (colunas de $\mathbf{W}$) e sua ativação correspondente no tempo (linhas de $\mathbf{H}$). No caso da NMF convolutiva, cada objeto tem uma sequência sucessiva de espectros e um padrão de ativação ao longo do tempo. É possível introduzir uma abordagem convolutiva para a NMF com amostras espaçadas irregularmente. Para isso, o modelo em (2) se torna

$$\hat{\mathbf{x}} = \sum_{t=0}^{T-1} \sum_{r=0}^{R-1} \mathbf{v}_{tr} \odot S_t(\mathbf{g}_r), \quad (6)$$

onde T é o tamanho da sequência de espectros e a função $S_t(\cdot)$ realiza o deslocamento relativo dos coeficientes, onde $S_0(\mathbf{g}_r) = \mathbf{g}_r$, $t > 0$ denota deslocamento para a direita e $t < 0$ para a esquerda, preenchendo com zero. A Figura 1 mostra um exemplo de grade irregular em tempo-frequência, e pode-se notar que o deslocamento no tempo deve ser realizado de forma que as amostras em baixa frequência se desloquem de forma mais lenta que as de alta frequência, ou seja, para um deslocamento de uma amostra em baixa frequência, é preciso realizar vários em alta frequência para não deformar o espectrograma. A Figura 2 mostra um

exemplo do deslocamento relativo realizado pela função $S_t(\cdot)$.

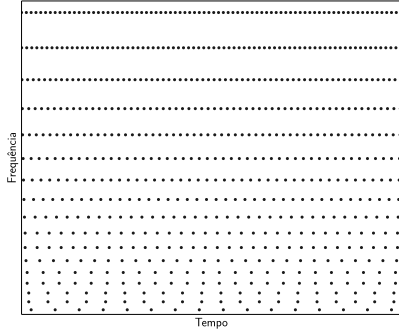


Figura 1: Exemplo de grade irregular no plano tempo-frequência para a CQT utilizando *frames* de Gabor não-estacionários.

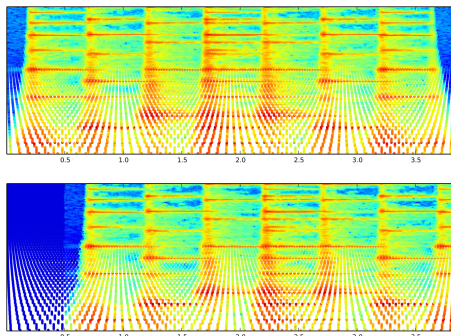


Figura 2: Exemplo de deslocamento de 0.5 segundos em um espectrograma da CQT.

As novas regras multiplicativas da fatoração em (6) podem ser vistas como um conjunto de T operações NMF em (3) que são somadas ao final. Com isso, o algoritmo precisa atualizar $T + 1$ matrizes a cada iteração. Então as regras multiplicativas se tornam

$$\begin{aligned} \mathbf{P}_t &= \frac{\mathbf{V}_t \odot S_t(\mathbf{G})}{\left(\sum_{t=0}^{T-1} \mathbf{V}_t \odot S_t(\mathbf{G}) \right) \cdot \mathbf{1}_{R \times R}} \\ \mathbf{V}_t &= \mathbf{D}_f \cdot S_{-t}(\mathbf{P}_t \odot (\mathbf{x} \cdot \mathbf{1}_{1 \times R})) \\ \mathbf{G} &= \mathbf{D}_t \cdot S_{-t}(\mathbf{P}_t \odot (\mathbf{x} \cdot \mathbf{1}_{1 \times R})). \end{aligned} \quad (7)$$

A cada iteração, todos $\mathbf{V}_t$ e $\mathbf{G}$ são atualizados, onde $\mathbf{G}$ é calculado como a média de suas atualizações para cada $\mathbf{V}_t$. Nota-se que para o caso $T = 1$, o algoritmo em (7) se torna (3).

Para utilizar NMF convolutiva para dados espaçados irregularmente em amostras obtidas a partir da CQT baseada em *frames* de Gabor não-estacionários, é possível perceber algumas características e fazer algumas considerações sobre as matrizes $\mathbf{D}_t$ e $\mathbf{D}_f$. Apesar dos *bins* de frequências serem espaçados geometricamente, seus valores são bem definidos. Sendo assim, pode-se utilizar o *kernel* binário em (4) para calcular $\mathbf{D}_f$, ou seja, os índices de frequência são valores inteiros.

Para o cálculo de $\mathbf{D}_t$, nota-se que as posições das amostras no tempo não são bem definidas, como pode ser visto na Figura 1. Na CQT, o tamanho do *atom* em tempo-frequência é adaptativa. Os parâmetros de deslocamento do *atom* no tempo dependem da largura de banda de sua posição em frequência. Com isso, em baixas frequências, consegue-se uma maior resolução em frequência e menor resolução no tempo, e para altas frequências, maior resolução no tempo e menor resolução em frequência. Esse fenômeno é conhecido como incerteza de *Heisenberg*. Sabendo disso, é conveniente usar o *kernel* gaussiano em (5) para o cálculo de $\mathbf{D}_t$. Entretanto, (5) define um valor de variância fixo para todas as amostras em $\mathbf{t}$. Devido ao problema da incerteza de *Heisenberg*, a variância fixa pode provocar um borrimento espectral no resultado do algoritmo, pois não é possível encontrar um valor fixo de variância que atenda satisfatoriamente às amostras em baixa e alta frequências. Com isso, é introduzido em (5) uma variância que depende do *bin* de frequência que a amostra se encontra. Então o valor de σ_t^2 em (5) passa a ser

$$\sigma_{t,\text{bin}}^2 = \frac{-d_{\text{bin}}^2}{\ln \gamma}, \quad 0 < \gamma < 1, \quad (8)$$

onde d_{bin} é a distância entre amostras vizinhas para cada *bin* de frequência, e γ é o valor do *kernel* gaussiano avaliado na amostra vizinha. A Figura 3 exemplifica a projeção dos *kernels* gaussianos de cada amostra em cada *bin*.

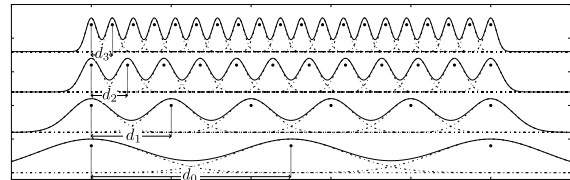


Figura 3: Exemplo de projeção de amostras com *kernels* gaussianos utilizando variâncias diferentes para cada *bin* de frequência. ($\gamma = 10^{-3}$)

Em Figura 4 e Figura 5 estão os pseudocódigos da técnica proposta e da função de deslocamento temporal.

4 Resultados experimentais

Esta seção destaca as vantagens da abordagem NMF convolutiva proposta para dados amostrados irregularmente (NRCNMF). Desta forma, usa-se dois exemplos de aplicação em espectro de áudio utilizando a CQT baseada em *frames* de Gabor não-estacionários como representação espectral. Em seguida, faz-se uma comparação com o algoritmo apresentado em (Smaragdakis and Kim, 2013), aqui chamado de NRNMF.

Entrada: $\mathbf{x}, \mathbf{t}, \mathbf{f} \in \mathbb{R}_{K \times 1}$ - vetores contendo as amostras e as posições em tempo e frequência, $R, iter_{max} \in \mathbb{Z}$ - o número de partes a fatorar e o número máximo de iterações, $t_s \in \mathbb{R}$ - deslocamento temporal em segundos.

Saída: As matrizes $\mathbf{V}_t, \mathbf{G} \in \mathbb{R}_{K \times R}$.

```

1 início
2   Calcula o número de deslocamentos
    $T \in \mathbb{Z}$  para  $t_s$  em segundos.
3   Inicializa aleatoriamente as  $T$  matrizes
    $\mathbf{V}_t$  e a matriz  $\mathbf{G}$ .
4    $\mathbf{D}_f(i, j) = \begin{cases} 1, & \mathbf{f}(i) = \mathbf{f}(j) \\ 0, & \mathbf{f}(i) \neq \mathbf{f}(j) \end{cases}$ 
5    $\mathbf{D}_t(i, j) = \exp\left(\frac{|\mathbf{t}(i) - \mathbf{t}(j)|^2}{\sigma_{t, bin}^2}\right)$ 
6   para cada  $i$  de 1 até  $iter_{max}$  faça
7     Calcula a reconstrução
       
$$\mathbf{R} = \left( \sum_{t=0}^{T-1} \mathbf{V}_t \odot \mathbf{S}_t(\mathbf{G}) \right) \cdot \mathbf{1}_{R \times R}$$

8     para cada  $t$  de 1 até  $T$  faça
9        $\mathbf{P}_t = \frac{\mathbf{V}_t \odot \mathbf{S}_t(\mathbf{G})}{\mathbf{R}}$ 
10       $\mathbf{G}_t = \mathbf{D}_t \cdot \mathbf{S}_{-t}(\mathbf{P}_t \odot (\mathbf{x} \cdot \mathbf{1}_{1 \times R}))$ 
11    fim
12    Atualiza  $\mathbf{G}$  como a média de todos  $\mathbf{G}_t$ .
13    Calcula a reconstrução  $\mathbf{R}$ .
14    para cada  $t$  de 1 até  $T$  faça
15       $\mathbf{P}_t = \frac{\mathbf{V}_t \odot \mathbf{S}_t(\mathbf{G})}{\mathbf{R}}$ 
16       $\mathbf{V}_t = \mathbf{D}_f \cdot \mathbf{S}_{-t}(\mathbf{P}_t \odot (\mathbf{x} \cdot \mathbf{1}_{1 \times R}))$ 
17    fim
18  fim
19 fim
```

Figura 4: NRCNMF

4.1 Aplicação em espectro de áudio

No primeiro exemplo, o sinal de áudio utilizado é composto de dois sinais do tipo *chirp* exponencial de 0,5 segundo, um crescente e outro decrescente, entre as frequências 1,5kHz e 2,5kHz. O sinal possui taxa de amostragem de 8kHz. A CQT foi realizada com frequência mínima $f_{min} = 1\text{kHz}$, frequência máxima $f_{max} = 4\text{kHz}$ e resolução de 48 *bins* por oitava (ou 1/4 de tom). Foram utilizados os algoritmos NRNMF e NRCNMF a fim de separar o espectro em Figura 6(a) em duas partes ($R = 2$).

Utilizando NRNMF, é possível notar na Figura 6 que o algoritmo não consegue reconstruir o espectrograma original. Isso ocorre porque o modelo do algoritmo não considera nenhuma característica das amostras, como por exemplo, a

Entrada: $\mathbf{M}$ - matriz $K \times R$ a ser deslocada, $\mathbf{L}$ - vetor contendo o número de amostras em cada *bin* de frequência, t - tamanho do deslocamento.

Saída: $\mathbf{M}_s$ - matriz com as amostras deslocadas no tempo.

```

1 início
2    $t_d = |t_s| \cdot t / T$ 
3   Calcula  $\mathbf{L}_c = [1, \text{cumsum}(\mathbf{L})]$ 
4   para cada  $j$  de 1 até  $\text{tamanho}(\mathbf{L})$  faça
5     O deslocamento para o  $j$ -ésimo bin
     de frequência  $d(j) = 0$ 
6     para cada  $k$  de  $\mathbf{L}_c(j)$  até
        $\mathbf{L}_c(j) + \mathbf{L}(j)$  faça
7       se  $\mathbf{t}(k) < |t_d|$  então
8         Incrementa  $d(j)$ .
9       fim
10    fim
11    Desloca  $\mathbf{M}(\mathbf{L}_c(j) : \mathbf{L}_c(j) + \mathbf{L}(j), :)$ 
     nas linhas de  $d(j)$  posições na
     direção do sinal de  $t_d$ , preenchendo
     com zeros.
12  fim
13 fim
```

Figura 5: Deslocamento $\mathbf{S}_t(\cdot)$

suavidade no tempo, que é típico para sinais de áudio. No caso da NRNMF, para alcançar uma reconstrução mais fiel, é necessário fatorar o espectrograma em mais partes ($R \gg 2$). Com isso, a matriz $\mathbf{V}$, formada pelos vetores $\mathbf{v}_r$ em (2), aprenderá mais padrões espectrais, a fim de refinar a reconstrução.

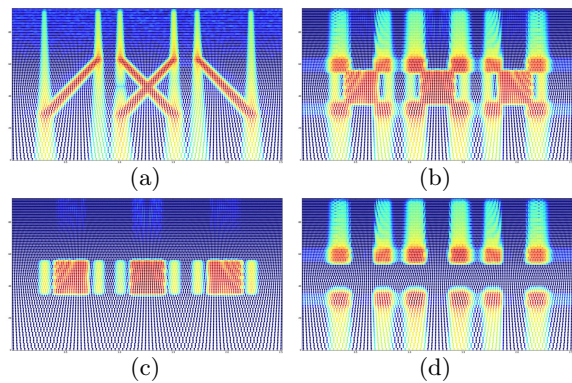


Figura 6: Exemplo de aplicação da NRNMF (Smaragdís and Kim, 2013). (a) Espectrograma do sinal. (b) Espectrograma do sinal reconstruído. (c) Parte 1. (d) Parte 2.

Para o algoritmo NRCNMF proposto, foi utilizado um deslocamento temporal de 0,55 segundo e o parâmetro $\gamma = 10^{-4}$ em (8). Nota-se na Figura 7 que o espectrograma foi reconstruído com fidelidade, e foi fatorado em duas partes, que correspondem aos sinais *chirp* decrescente e cres-

cente.

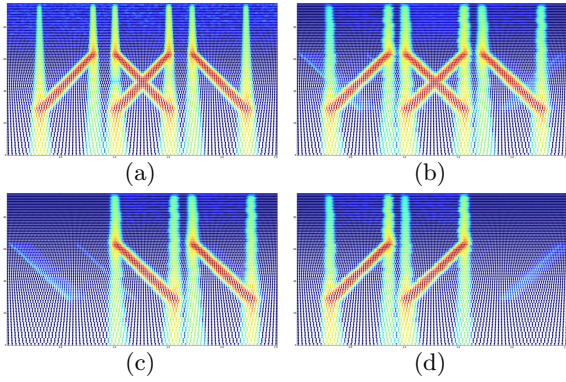


Figura 7: Exemplo de aplicação da NRCNMF. (a) Espectrograma do sinal. (b) Espectrograma do sinal reconstruído. (c) Parte 1. (d) Parte 2.

No caso da NRCNMF, as matrizes $\mathbf{V}_t$, formadas pelos vetores $\mathbf{v}_{tr}$ em (6), aprenderão sequências de T padrões espectrais ao longo do tempo. Como a CQT é construída com *bins* de frequências em posições bem definidas, é possível fazer um mapeamento das matrizes $\mathbf{V}_t$ para matrizes no formato de $\mathbf{W}$ em (1), e dessa forma, visualizar os padrões espectrais aprendidos. O mapeamento é

$$\mathbf{w}_r(m) = \mathbf{v}_r(i), \quad \forall i : \mathbf{f}(i) = m, \quad (9)$$

onde $\mathbf{w}_r$ são as colunas de $\mathbf{W}$, $m \in \{1, \dots, M\}$ é um índice inteiro uniforme, e M é o número de *bins* de frequências.

A Figura 8 mostra as duas sequências de padrões espectrais obtidas a partir de (9). Nota-se que cada sequência forma objetos referentes a cada uma das partes do espectrograma fatorado.

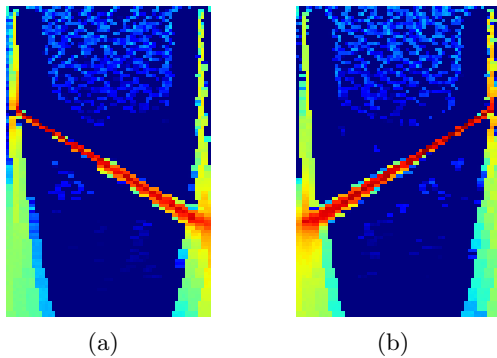


Figura 8: As sequências dos padrões espectrais relativas à separação do sinal Fig.7(a) utilizando a NRCNMF.

4.2 Aplicação em música

O segundo exemplo consiste de uma aplicação simples num contexto mais realista, a fim de ver o comportamento do algoritmo proposto. O áudio a ser fatorado é composto pela sequência de notas

tocadas por um piano sintetizado, como mostra a Figura 9, com taxa de amostragem de 16kHz. As frequências fundamentais das notas musicais utilizadas são 523,25Hz (C5), 659,255Hz (E5) e 783,99Hz (G5).



Figura 9: Partitura do áudio tocado pelo piano.

A CQT foi realizada com frequência mínima $f_{\min} = 220\text{Hz}$, frequência máxima $f_{\max} = 8\text{kHz}$ e resolução de 48 *bins* por oitava. Para o algoritmo proposto, foi utilizado um deslocamento temporal de 0.55 segundo, parâmetro $\gamma = 10^{-4}$ e $R = 3$. Para uma melhor percepção da fatoração, a Figura 10(a) mostra os sinais no tempo das partes encontradas pelo algoritmo, assim como a soma das três partes, que é o sinal reconstruído.

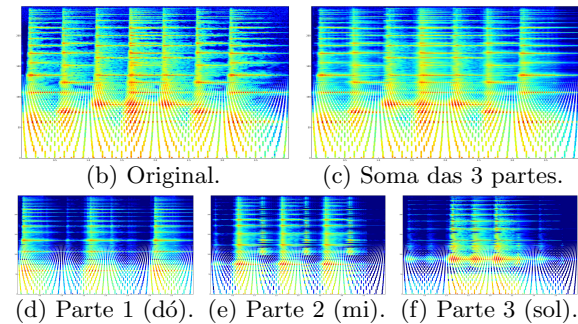
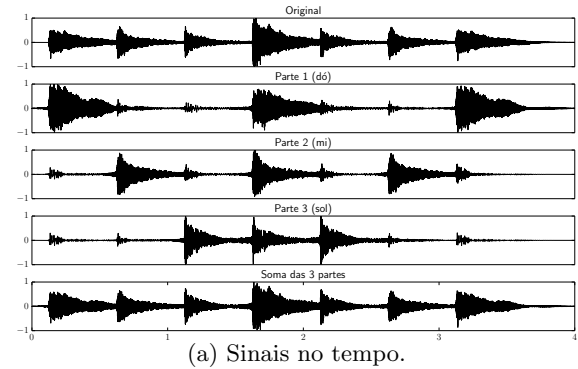


Figura 10: Sinais no tempo e espectrogramas relacionado à aplicação com música.

Como pode-se notar, os espectrogramas da Figura 10 se aproximam mais de um som do mundo real, em que há a presença de componentes harmônicas de cada nota e do timbre do piano. Cada nota foi isolada, como se vê na Figura 10, e as sequências de padrões espectrais de cada nota fatorada são mostradas na Figura 11. Importante notar que as sequências de padrões espectrais se repetem ao longo do espectrograma, e os vetores $\mathbf{g}_r$ em (6) possuem os coeficientes que regem as posições e intensidades que esses padrões se repetem.

Através desses experimentos, pôde-se mostrar as vantagens da introdução do modelo convolutivo

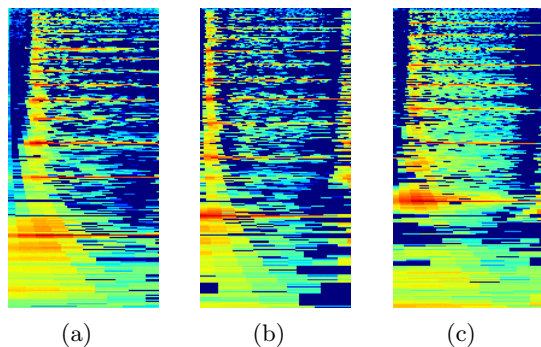


Figura 11: As sequências dos padrões espectrais relativas à separação ao espectrograma em Fig.10(b) utilizando a NRCNMF.

à NMF para amostras espaçadas irregularmente aqui apresentado. Para um R pequeno, a separação espectral com NRNMF não retorna o espectrograma fiel ao original, pois o modelo aprende padrões espectrais isolados temporalmente. Enquanto que o NRCNMF aprende uma sequência consecutiva de padrões espectrais no tempo, sendo possível usar um R pequeno, e dessa forma, explicar melhor os objetos a serem fatorados. Para fins de implementação, o tamanho das matrizes ficam menores, precisando de menos espaço em memória.

O uso da CQT baseada em *frames* de Gabor não-estacionários também é vantajoso pelo fato de ser uma representação que possui transformação inversa eficiente, baseada na percepção humana do som e mais apropriada para tratar música do que a STFT. No caso da CQT, ao escolher uma frequência mínima igual à frequência de uma nota musical (por exemplo, $f_{min} = 220\text{Hz}$, que corresponde à nota musical A3), os demais *bins* de frequência também corresponderão à frequências de notas musicais.

Os resultados em áudio presentes nesse artigo se encontram hospedados no website <https://sites.google.com/site/nrcnmfcba2016>.

5 Conclusão

Nesse artigo foi apresentada uma abordagem convolutiva para o algoritmo de fatoração de matriz não-negativa para dados amostrados irregularmente, como é o caso da transformada Q constante utilizando *frames* de Gabor não-estacionários, onde não é possível representação matricial dos dados. A abordagem convolutiva considera uma característica inerente aos sinais de áudio, que é a dependência temporal. Através de experimentos, foi possível mostrar que tal abordagem torna o processo de separação espectral mais apropriado para aplicações de separação espectral em áudio. Observando algumas características próprias da CQT, também foi possível fazer adaptações no algoritmo a fim de evitar alguns problemas na re-

construção espectral.

Referências

- Brown, J. (1991). Calculation of a constant q spectral transform, *Journal of the Acoustical Society of America* **89**(1): 425–434.
- Fevotte, C., Bertin, N. and Durrieu, J. (2009). Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis, *Neural Computation* **21**(3): 793–830.
- Lee, D. D. and Seung, H. S. (2001). Algorithms for non-negative matrix factorization, *In NIPS*, MIT Press, pp. 556–562.
- M. Dörfler, N. Holighaus, T. G. and Velasco, G. (2011). Constructing an invertible constant- q transform with nonstationary gabor frames, *In Proceedings of the 14th International Conference on Digital Audio Effects (DAFx 11)*, Paris, France, 2011.
- Smaragdis, P. (2004). *Non-negative Matrix Factor Deconvolution; Extraction of Multiple Sound Sources from Monophonic Inputs*, Springer Berlin Heidelberg, Berlin, Heidelberg, chapter Independent Component Analysis and Blind Signal Separation: Fifth International Conference, ICA 2004, Granada, Spain, September 22–24, 2004. Proceedings, pp. 494–499.
- Smaragdis, P. (2007). Convolutional speech bases and their application to supervised speech separation, *Trans. Audio, Speech and Lang. Proc.* **15**(1): 1–12.
- Smaragdis, P. and Brown, J. C. (2003). Non-negative matrix factorization for polyphonic music transcription, *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.*, pp. 177–180.
- Smaragdis, P. and Kim, M. (2013). Non-negative matrix factorization for irregularly-spaced transforms., *WASPAA*, IEEE, pp. 1–4.