TÉCNICO LISBOA

Aplicação da fatoração em matrizes não negativas à separação de fontes de áudio para reconhecimento automático de fala

Cristina Tong Ribeiro

Dissertação para obtenção do Grau de Mestre em Engenharia Eletrotécnica e de Computadores

Sistemas, Decisão e Controlo

Júri

Presidente: Isabel Trancoso, IST/INESC-ID
Orientadora: Isabel Trancoso IST/INESC-ID
Orientador: Miguel Arjona Ramírez, Escola Politécnica da USP
Vogais: António Teixeira, Universidade de Aveiro
Hugo Meinedo, INESC-ID
João Sequeira, IST, Coord. Área Especialização

Outubro de 2013

Dedicatória

Aos meus pais, um
pouquinho da minha gratidão.

Agradecimentos

Meus agradecimentos ao professor Miguel, pela disponibilidade, pelo interesse e pela prontidão.
À professora Isabel, pelo brilho nos olhos que despertou meu interesse.
Ao Hugo, pelas ferramentas, pelo suporte e pela paciência.
E ao amor da minha vida, que entendeu, aguentou e cuidou.

Resumo

Este trabalho tem por objetivo avaliar o potencial da Fatoração em Matrizes Não Negativas (NMF) quando aplicada à Separação de Fontes de Áudio (ASS), com objetivo de melhorar o desempenho de um Reconhecedor Automático de Fala (ASR). Incialmente apresentam-se os fundamentos teóricos da NMF, a motivação que a originou, os algoritmos para sua realização e como ela pode ser aplicada, de maneira supervisionada ou não, para a ASS. Em seguida, realizam-se experimentos com misturas de voz e música que visam avaliar quais dos parâmetros envolvidos influenciam os resultados da separação e como os influenciam. Os resultados apresentados nesse trabalho evidenciam que a NMF tem potencial significativo para melhorar sinais de voz corrompidos por música, em termos da relação de energia entre esses dois sinais e também em relação ao desempenho do ASR. Verificou-se que o tamanho das bases é bastante relevante para os resultados da separação, ao contrário do número de iterações realizadas pela NMF. Para a construção das bases, concluiu-se indubitavelmente que treiná-las pela NMF é muito mais eficaz que compô-las com exemplares. Para a fase de separação, os dois algoritmos experimentados geraram resultados semelhantes.

Palavras-Chave: Fatoração em matrizes não negativas. Separação de fontes de áudio. Reconhecimento automático de fala. Melhoria de sinais de voz. Misturas de voz e música.

Abstract

The goal of this work is to evaluate Non-negative Matrix Factorization's (NMF) potential to Audio Source Separation (AuSS), relative to improving the performance of an Automatic Speech Recognition System (ASR). First, we present the theoretical foundations of NMF, its origins, some algorithms for its implementation and methods for its application on ASS, in both supervised and unsupervised manners. Second, we experiment on mixtures of speech and music in order to evaluate how much various conditions and parameters affect the result of the ASS. The results here presented make it clear that the NMF can be applied to enhance speech signals corrupted by music, both in terms of signal-to-noise-ratios and of ASR's performance. The number of iterations taken by the NMF algorithms turned out not to be of great relevance. On the contrary, base sizes were very important on the separation results. For the construction of the base-matrices, it was found that training them with NMF yields results indubitably better than building them with exemplars. During the separation phase, results of both the algorithms experimented were equally good.

Keywords: Audio source separation. Automatic speech recognition. Non-negative matrix factorization. Speech and music mixtures. Speech enhancement.

Lista de Ilustrações

Lista de Abreviações

ASR	Reconhecimento/Reconhecedor Automático de Fala (*Automatic Speech Recognition*)
AuSS	Separação de Fontes de Áudio (*Audio Source Separation*)
BASS	Separação cega de fontes de áudio (*Blind Audio Source Separation*)
BSS	Separação cega de fontes (*Blind Source Separation*)
DFT	Transformada de Fourier Discreta (*Discrete Fourier Transform*)
EVAL	Avaliação (*Evaluation*)
ICA	Análise em Componentes Independentes (*Independent Component Analysis*)
INESC-ID	Instituto de Engenharia de Sistemas e Computadores – Investigação e Desenvolvimento
ISD	Divergência de Itakura-Saito (*Itakura-Saito Divergence*)
IS-NMF	Fatoração em Matrizes Não Negativas com Divergência de Itakura-Saito
ISTFT	Tranformada Inversa de Fourier de Curto Prazo (*Inverse STFT)*
KLD	Divergência de Kullback-Leibler (*Kullback-Leibler Divergence*)
KL-NMF	Fatoração em Matrizes Não Negativas com Divergência de Kullback-Leibler
LVA	Análise de Variáveis Latentes (*Latent Variable Analysis*)
MSR	Relação música-voz (Music-to-speech ratio*)*
NIST	National Institute of Standards and Technology
NMF	Fatoração em Matrizes Não Negativas (*Non-negative Matrix Factorization*)
PCA	Análise em Componentes Principais (*Principal Component Analysis*)
RGB	Modelo de cores "red, green, blue"
RTP	Rádio e Televisão de Portugal
SAR	Relação sinal-artefatos (*Signal-to-artifacts ratio*)
SCLITE	Score Lite scoring package
SDR	Relação sinal-distorção (*Signal-to-distortion ratio*)
SIR	Relação sinal-interferências (*Signal-to-interferences ratio*)
SMR	Relação voz-música (*Speech-to-music ratio*)
STFT	Transformada de Fourier de Curto Prazo (*Short Time Fourier Transform*)
SVD	Decomposição em Valores Singulares (*Sigular Value Decomposition*)

Conteúdo

1. Introdução

1.1. Motivação

O Laboratório de Sistemas de Língua Falada (L²F) do Instituto de Engenharia de Sistemas e Computadores – Investigação e Desenvolvimento em Lisboa (INESC-ID Lisboa) possui um sistema de reconhecimento automático de fala (ASR), o AUDIMUS [17]. Em observação aos resultados da versão AUDIMUS.MEDIA quando aplicada à transcrição, ao vivo, das falas dos noticiários da Rádio e Televisão de Portugal (RTP), percebeu-se uma considerável queda de desempenho em situações em que a voz estava misturada a algum outro sinal, nomeadamente música. Estendendo a aplicação desde noticiários até outros tipos de programas televisivos, como documentários, novelas, programas de perguntas e respostas e *reality shows*, constatou-se que era bastante comum o uso de música como plano de fundo para a voz, tanto na forma de vinhetas quanto para ambientação. Assim, um sistema capaz de separar os sinais de voz e música ou de reduzir os efeitos da melodia de fundo sobre a fala poderia ser bastante útil para melhorar o desempenho desse ASR.

De fato, o problema de separação de fontes de áudio (AuSS) é ainda um tópico de pesquisa sem solução satisfatória. Por estar há tanto tempo no foco de especialistas de diferentes áreas – processamento de sinais digitais, acústica e linguística, por exemplo –, a questão já foi abordada por inúmeros ângulos. Mais ainda, cada uma dessas abordagens gerou múltiplas técnicas para atacar o problema.

Dessa forma, um problema de caráter tão desafiador, com abundantes aplicações práticas e diversas linhas de desenvolvimento não poderia ser motivação maior para o início de uma pesquisa.

1.2. Revisão de Técnicas Existentes

Processamento de sinais de áudio é uma área de pesquisa extensa e com ainda muito espaço para desenvolvimento. Um dos tópicos que merecem atenção é o de distinção de múltiplos sinais dentro de uma mistura. Pode-se dividir essa questão entre dois problemas: a separação de fontes de áudio (AuSS) e a extração de um sinal mais proeminente (*signal enhancement*).

Somente é possível separar perfeitamente dois sinais se eles forem, de alguma maneira, ortogonais [1]. Assim sendo, antes de proceder à AuSS, há de se representar os sinais envolvidos de maneira tão ortogonal quanto possível. Para isso, aplicam-se transformadas – comumente, a transformada de Fourier – e extraem-se características dos sinais – como coeficientes cepstrais e de predição linear –, de acordo com a conveniência para a aplicação.

Uma vez que os sinais envolvidos estejam devidamente representados, resta executar a separação propriamente dita. Podem-se agrupar os métodos de separação existentes entre supervisionados e não supervisionados. Os primeiros são aqueles que recebem informações *a priori* sobre características dos sinais que compõem a mistura, quantidade de fontes emissoras de áudio e/ou processo de mixagem. Os métodos não supervisionados, por sua vez, são aqueles que possuem nenhuma (ou quase nenhuma) informação acerca dos parâmetros citados.

Métodos não supervisionados buscam estimar as fontes de áudio a partir de misturas delas. Para tal, algoritmos de aprendizagem automática sujeitos a restrições separam a mistura em diversos componentes. Outro algoritmo, então, é necessário para agrupar os componentes e aproximar os sinais das fontes.

Por outro lado, nos métodos supervisionados, modelam-se os sinais envolvidos (todos eles, para métodos completamente supervisionados, ou apenas os de interesse, para semi-supervisionados) e, então, filtra-se a mistura utilizando os modelos criados. O desafio aqui reside na criação de modelos precisos e na exploração mais vantajosa das características tanto dos próprios sinais como do processo de gravação deles.

Dentre as técnicas utilizadas para modelar sinais, destacam-se, no âmbito de processamento de sinais áudio, os Modelos Ocultos de Markov [3], Modelos Lineares de Sistemas Dinâmicos [20] e Modelos Gráficos [9].

Para proceder à AuSS, modelos de decomposição ou fatoração espectral são os mais comuns. Destacam-se a Análise em Variáveis Latentes (LVA) [21], a Fatoração em Matrizes Não Negativas (NMF) [25][13][14], a Análise em Componentes Independentes (ICA) [3] e [5] e a Análise em Componentes Principais (PCA) [10].

1.3. Proposta

Como apresentado na seção 1.1, nesse projeto, trabalhar-se-á com misturas de sinais de voz e de música. Tendo a fala sempre como sinal de interesse, muitas vezes se referirá à música como *ruído* ou dir-se-á que a voz é *corrompida* por música. Vale ressaltar ainda que se utilizará *voz* como sinônimo de *fala* (em oposição a um sinal de voz que tome apenas um fone, por exemplo), assim como o sinal de música deverá sempre ser interpretado como um trecho de uma canção (em vez de, por exemplo, um tom de um único instrumento).

Devido ao fato de tanto música quanto voz serem sinais altamente não estacionários, modelos simples raramente serão capazes de representá-los bem. Além disso, a variedade de palavras, de fonemas, de timbres de voz, de gêneros e de instrumentos musicais torna quase impossível descrever um modelo que abranja com generalidade sinais de música e de fala. Para completar, na maioria dos casos,

os espectros desses sinais se sobrepõem significativamente. Assim, a representação ortogonal deles é um patamar de pesquisa ainda não alcançado. Por todos esses motivos é que a separação de fontes de áudio em misturas de voz e música é um problema tão desafiador.

Uma das técnicas mais promissoras para a realização de AuSS com misturas de voz e música é a NMF. Com ela, a mistura, representada apenas com coeficientes não negativos, é decomposta em um conjunto de bases e outro de pesos, também de elementos não negativos. As bases serão responsáveis por descrever as características espectrais de cada um dos sinais envolvidos, enquanto que os pesos proverão a evolução temporal de tais características. O espectro de cada sinal é aproximado por uma combinação linear ponderada, puramente aditiva, dos elementos da base.

Se tomada sem modificações, a NMF é uma técnica de separação não supervisionada, como se explicará em mais detalhes adiante. Contudo, ao se elaborar o conjunto de bases consoante as características dos sinais envolvidos, insere-se conhecimento *a priori* no método e torna-se-o supervisionado.

Para um trabalho de pesquisa baseado em NMF e em sinais de voz e música, os desafios estarão principalmente em:

- Criar as bases de forma a que elas carreguem em si características de grande significado para cada sinal e que sejam capazes de diferenciá-los;
- Determinar as restrições apropriadas para a aplicação da NMF tendo em vista o objetivo de AuSS.

Para atacar ambos os pontos acima, estudar-se-ão algumas abordagens. No primeiro caso, dar-se-á ênfase a bases criadas com exemplos de sinais de voz e de música (como em [20]), investigando-se também os resultados provenientes de bases treinadas pela NMF (como em [7]). No segundo, averiguar-se-ão os efeitos de restrições que visam explorar todo o potencial da NMF, comparando-se divergências aplicadas como função de custo da etapa de separação (divergência de Kullback-Leibler ou de Itakura-Saito).

Por fim, para avaliar a qualidade das separações realizadas, dois tipos de medidas serão usados: comparação da energia do sinal separado e do original em termos do que se propõe em [24] e o desempenho do AUDIMUS.MEDIA [17], com auxílio do NIST SCLITE [29].

2. Fundamentação Teórica

2.1. A Origem da NMF

A decomposição ou fatoração de matrizes é há muito tempo explorada para diferentes fins: resolução de sistemas lineares de equações, análise multivariável, redução de dimensão de dados, revelação de estruturas ocultas e representação de transformadas, por exemplo. Por serem aplicáveis a tantas áreas, diferentes técnicas de fatoração de matrizes foram, e ainda são, desenvolvidas. Dentre elas, destaquemos a Análise em Componentes Principais (PCA).

Baseada na Decomposição em Valores Singulares (SVD), a PCA busca representar um conjunto de dados de grande dimensão pela combinação linear de uma série de vetores-base ortogonais entre si (as componentes principais, dadas por aqueles vetores singulares que estão associados aos maiores valores singulares). Em geral, aplica-se-a de forma a que a quantidade de vetores-base seja menor que a dimensão dos dados originais e, assim, transporta-se a análise a um espaço de menor dimensão. Contudo, é sempre possível tomar a totalidade dos vetores singulares como base da PCA e, assim, obtém-se reconstrução perfeita dos dados originais, a custo de não se reduzir sua dimensão. Note-se, ainda, que a decomposição gerada pela PCA não somente promove redução dimensional, como também pode ser usada para revelar estruturas latentes no conjunto de observações.

Abordando agora o ponto sob a ótica da revelação de estruturas nos dados, chama-se a atenção para o fato de que a PCA não restringe de nenhuma maneira o sinal dos fatores da decomposição. Podem ocorrer, portanto, elementos positivos e negativos tanto entre os vetores-base quanto nos coeficientes que os combinam linearmente. Em muitas aplicações, porém, elementos negativos não podem ser fisicamente interpretados. A intensidade dos pixels de uma imagem, por exemplo, assume valores dentro da faixa de 0 a 255. Um valor negativo entre os vetores-base, nesse caso, não possuiria interpretação intuitivamente razoável. Da mesma maneira, se tomarmos vetores-base como as cores vermelho, verde e azul do sistema aditivo de cores RGB, então elementos negativos também não devem aparecer entre os coeficientes da combinação.

Visando sanar tal inconveniente, estudiosos passaram a forçar elementos exclusivamente não negativos nos fatores decompostos. Surge, assim, a Fatoração em Matrizes Não Negativas (NMF). Em [14], Lee e Seung relacionam a NMF a essa interpretação físico-intuitiva, mostrando que a restrição de não negatividade é compatível com a intuição de combinar aditivamente partes para formar um todo.

Com estudos publicados desde 1994 [18], a NMF é uma técnica relativamente recente e se tornou mais popular a partir de 1999, quando Lee e Seung publicaram algoritmos mais simples para a sua implementação. Ainda hoje, contudo, a NMF é uma técnica computacionalmente muito dispendiosa.

2.2. Descrição Matemática do Problema

O problema de decomposição que a NMF busca resolver pode ser descrito da maneira como se segue: dada uma matriz $\boldsymbol{V}$ de elementos não negativos, obter duas matrizes $\boldsymbol{B}$ e $\boldsymbol{W}$, também de elementos não negativos, tais que $\boldsymbol{V} \approx \boldsymbol{BW}$. A qualidade dessa aproximação pode ser medida com diferentes critérios, isto é, para diferentes funções de custo. Assim, o problema da NMF também pode ser descrito como o seguinte problema de otimização:

$$\min_{\boldsymbol{B},\boldsymbol{W} \geq 0} f(\boldsymbol{V},\boldsymbol{B},\boldsymbol{W}) \tag{1}$$

onde $f(\boldsymbol{V},\boldsymbol{B},\boldsymbol{W})$ é a função de custo que avalia a aproximação. Daqui para frente, a não ser que seja dito o contrário, as inequações $\boldsymbol{B} \geq 0$ e $\boldsymbol{W} \geq 0$ devem ser interpretadas como sendo operações elemento a elemento.

Nesse trabalho, adotar-se-á a seguinte notação: letras maiúsculas em negrito representarão matrizes, letras minúsculas em negrito representarão vetores e letras quaisquer sem negrito serão usadas para representar escalares.

Sejam $\boldsymbol{V}$ uma matriz $(m \times n)$ tal que $\boldsymbol{V} = [\boldsymbol{v_1} \ldots \boldsymbol{v_n}]$, $\boldsymbol{B}$ uma matriz $(m \times r)$ tal que $\boldsymbol{B} = [\boldsymbol{b_1} \ldots \boldsymbol{b_r}]$ e $\boldsymbol{W}$ uma matriz $(r \times n)$ tal que $\boldsymbol{W} = \{w_{ij}\}_{i=1,\ldots,r\ ,j=1,\ldots,m}$. Então, escreve-se:

$$\boldsymbol{v_j} = \sum_{i=1}^{r} \boldsymbol{b_i}\, w_{ij} \tag{2}$$

A equação (2) equivale a dizer que cada coluna $\boldsymbol{v_j}$ de $\boldsymbol{V}$ é uma combinação linear das colunas $\boldsymbol{b_i}$ de $\boldsymbol{B}$, ponderada pelos elementos $\{w_{ij}\}_{i=1,\ldots,r}$ da matriz $\boldsymbol{W}$. Assim, $\boldsymbol{B}$ pode ser interpretada como um conjunto de vetores-base, enquanto $\boldsymbol{W}$ é uma matriz de pesos representativa das ativações dos componentes de $\boldsymbol{B}$ que melhor aproximam $\boldsymbol{V}$.

Seja $y(t)$ um sinal de valores exclusivamente positivos e seja $\boldsymbol{y}$ um vetor que armazena $y(t)$. Divide-se $\boldsymbol{y}$ em n quadros e dispõem-se-os cada um em uma coluna de $\boldsymbol{V}$. Então, o índice j de (2) passa a representar uma variável de tempo, discretizada pela duração (e sobreposição) de cada quadro. Assim, a cada instante $j = t$, $\boldsymbol{v_t}$ é aproximado pela combinação linear de todos os vetores-base $\boldsymbol{b_i}$, ponderados pelos pesos do instante t, $\{w_{it}\}_{i=1,\ldots,r}$. Portanto, pode-se dizer que $\boldsymbol{B}$ revela um conjunto de características de $\boldsymbol{y}$, enquanto $\boldsymbol{W}$ provê a evolução temporal das mesmas.

A dimensão r, interna ao produto $\boldsymbol{BW}$, é arbitrária e representa a quantidade de vetores-base usados na fatoração. Convém, contudo, selecioná-la de forma a que $r < \min(m,n)$, para que se verifique de fato uma redução na dimensão dos dados.

Vale ressaltar, por fim, que a decomposição $\boldsymbol{V} = \boldsymbol{BW}$ não é única. Em outras palavras, há infinitas configurações, sem elementos negativos, de $\boldsymbol{B}$ e $\boldsymbol{W}$ que geram aproximações de mesma qualidade para $\boldsymbol{V}$.

2.3. Algoritmos para a NMF

Diferentes algoritmos foram publicados para a resolução do problema descrito em (1), dependendo da função de custo $f(\boldsymbol{V}, \boldsymbol{B}, \boldsymbol{W})$ e de outros objetivos, como velocidade de convergência, facilidade de implementação e complexidade temporal. Em realidade, nenhuma das abordagens apresentadas a seguir é capaz de garantir a convergência para o mínimo global de $f(\boldsymbol{V}, \boldsymbol{B}, \boldsymbol{W})$, mas apenas para um mínimo local. Assim, em geral, inicializam-se as matrizes $\boldsymbol{B}$ e $\boldsymbol{W}$ com valores não negativos aleatórios, na esperança de se atingir a solução global do problema.

Paatero e Tapper [18], em 1994, desejavam analisar os principais fatores responsáveis por um conjunto de observações. O problema que se buscava solucionar pode ser descrito como:

$$\min_{\boldsymbol{B},\boldsymbol{W} \geq 0} \|\boldsymbol{H} \cdot (\boldsymbol{V} - \boldsymbol{BW})\|_F^2 \tag{3}$$

onde $\boldsymbol{V}$ é o conjunto de observações/medições, $\boldsymbol{B}$ são os fatores responsáveis por $\boldsymbol{V}$, $\boldsymbol{W}$ são as influências de cada fator e $\boldsymbol{H}$ é a confiança de cada medição. $\|\boldsymbol{X}\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |x_{ij}|^2}$ representa a norma de Frobenius e $\cdot$ representa uma multiplicação elemento a elemento.

Para obter $\boldsymbol{B}$ e $\boldsymbol{W}$, Paatero e Tapper propuseram aplicar uma técnica de Mínimos Quadrados Alternados com algumas restrições. A técnica consiste em, iterativamente, fixar $\boldsymbol{B}$ e solucionar o problema de otimização para $\boldsymbol{W}$. Em seguida, fixar $\boldsymbol{W}$ e resolver para $\boldsymbol{B}$, repetindo até que a função de custo atinja certo valor de parada.

Lee e Seung, paralelamente, aplicaram um método de Gradiente Projetado Alternado [15] para a resolução de

$$\min_{\boldsymbol{B},\boldsymbol{W} \geq 0} \|\boldsymbol{V} - \boldsymbol{BW}\|_F^2 \tag{4}$$

A técnica, também iterativa e alternada, consiste em fixar $\boldsymbol{B}$, aplicar o método do gradiente (*gradient descent*) para $\boldsymbol{W}$ e zerar todos os elementos negativos em $\boldsymbol{W}$. Então, fixar os novos valores de $\boldsymbol{W}$ e atualizar $\boldsymbol{B}$ da mesma maneira como se fez com $\boldsymbol{W}$. Repetir até que a aproximação $\boldsymbol{V} \approx \boldsymbol{BW}$ tenha qualidade satisfatória.

Descontentes com o tempo de processamento exigido para se computar o método acima, Lee e Seung publicaram, no ano 2000, uma técnica baseada em regras multiplicativas para a atualização dos

valores de $\boldsymbol{B}$ e de $\boldsymbol{W}$ [13]. Nessa ocasião, duas funções de custo diferentes foram consideradas, o quadrado da distância Euclidiana e uma generalização da divergência de Kullback-Leibler (KLD), dadas respectivamente por:

$$\|\boldsymbol{A} - \boldsymbol{B}\|^2 = \sum_{ij} \left(A_{ij} - B_{ij}\right)^2 \tag{5}$$

$$D(\boldsymbol{A}||\boldsymbol{B}) = \sum_{ij} \left(A_{ij}\, log \frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij} \right) \tag{6}$$

Assim, os problemas de otimização que se busca solucionar podem ser enunciados como:

$$\min_{\boldsymbol{B},\boldsymbol{W} \geq 0} \|\boldsymbol{V} - \boldsymbol{BW}\|^2 \tag{7}$$

$$\min_{\boldsymbol{B},\boldsymbol{W} \geq 0} D(\boldsymbol{V}||\boldsymbol{BW}) \tag{8}$$

Uma vez que as funções $\|\boldsymbol{V} - \boldsymbol{BW}\|^2$ e $D(\boldsymbol{V}||\boldsymbol{BW})$ não são convexas para ambas as variáveis $\boldsymbol{B}$ e $\boldsymbol{W}$ juntas (apesar de serem para $\boldsymbol{B}$ e para $\boldsymbol{W}$ individualmente), não é possível solucionar (7) e (8) no sentido de se atingirem garantidamente seus mínimos globais. Técnicas numéricas de otimização, como algoritmos de gradiente, vêm à mente, então, para computar $\boldsymbol{B}$ e $\boldsymbol{W}$ que alcancem mínimos locais. A convergência de tais algoritmos, porém, pode ser demasiado lenta ou nunca ser atingida, a depender do tamanho do passo tomado.

As regras multiplicativas que se expõem a seguir foram apresentadas por Lee e Seung como um bom compromisso entre velocidade de computação e facilidade de implementação. Para introduzi-las, apresentar-se-á uma relação com o método do gradiente. Tome-se, então, a seguinte regra aditiva de atualização para W:

$$W_{a\mu} \leftarrow W_{a\mu} + \eta_{a\mu}\left[(\boldsymbol{B}^T\boldsymbol{V})_{a\mu} - (\boldsymbol{B}^T\boldsymbol{BW})_{a\mu}\right] \tag{9}$$

Se $\eta_{a\mu} > 0$ for um número pequeno, então (9) é equivalente ao método do gradiente. Assim, contanto que $\eta_{a\mu}$ seja suficientemente pequeno, a atualização (9) converge para um mínimo local da função de custo de (7), isto é, reduz $\|\boldsymbol{V} - \boldsymbol{BW}\|$.

Se dimensionarmos o passo de atualização η para

$$\eta_{a\mu} = \frac{W_{a\mu}}{(\boldsymbol{B}^T\boldsymbol{BW})_{a\mu}} \tag{10}$$

o que se obtém é:

$$W_{a\mu} \leftarrow W_{a\mu} + \frac{W_{a\mu}}{(\boldsymbol{B}^T\boldsymbol{B}\boldsymbol{W})_{a\mu}}\left[(\boldsymbol{B}^T\boldsymbol{V})_{a\mu} - (\boldsymbol{B}^T\boldsymbol{B}\boldsymbol{W})_{a\mu}\right]$$

$$W_{a\mu} \leftarrow W_{a\mu} + \frac{W_{a\mu}(\boldsymbol{B}^T\boldsymbol{V})_{a\mu}}{(\boldsymbol{B}^T\boldsymbol{B}\boldsymbol{W})_{a\mu}} - \frac{W_{a\mu}(\boldsymbol{B}^T\boldsymbol{B}\boldsymbol{W})_{a\mu}}{(\boldsymbol{B}^T\boldsymbol{B}\boldsymbol{W})_{a\mu}}$$

$$W_{a\mu} \leftarrow W_{a\mu}\frac{(\boldsymbol{B}^T\boldsymbol{V})_{a\mu}}{(\boldsymbol{B}^T\boldsymbol{B}\boldsymbol{W})_{a\mu}} \tag{11}$$

Analogamente, obtém-se para a atualização de $\boldsymbol{B}$:

$$B_{ia} \leftarrow B_{ia}\frac{(\boldsymbol{V}\boldsymbol{W}^T)_{ia}}{(\boldsymbol{B}\boldsymbol{W}\boldsymbol{W}^T)_{ia}} \tag{12}$$

Prova-se em [13] que, apesar de o η definido em (10) não ser pequeno, de forma a garantir a convergência de (9), a distância Euclidiana $\|\boldsymbol{V} - \boldsymbol{B}\boldsymbol{W}\|$ de fato é não crescente para as regras multiplicativas (11) e (12). Além disso, demonstra-se que $\|\boldsymbol{V} - \boldsymbol{B}\boldsymbol{W}\|$ torna-se invariante se e somente se $\boldsymbol{B}$ e $\boldsymbol{W}$ formarem um ponto fixo das atualizações (11) e (12). Portanto, as regras multiplicativas acima encontram um mínimo local da função de custo em (7).

Agora, para a Divergência de Kullback-Leibler generalizada, adote-se a regra aditiva

$$W_{a\mu} \leftarrow W_{a\mu} + \eta_{a\mu}\left[\sum_i B_{ia}\frac{V_{i\mu}}{(\boldsymbol{B}\boldsymbol{W})_{i\mu}} - \sum_i B_{ia}\right] \tag{13}$$

e o passo de atualização

$$\eta_{a\mu} = \frac{W_{a\mu}}{\sum_i B_{ia}} \tag{14}$$

Obtém-se, então:

$$W_{a\mu} \leftarrow W_{a\mu} + \frac{W_{a\mu}}{\sum_i B_{ia}}\left[\sum_i B_{ia}\frac{V_{i\mu}}{(\boldsymbol{B}\boldsymbol{W})_{i\mu}} - \sum_i B_{ia}\right]$$

$$W_{a\mu} \leftarrow W_{a\mu} + \frac{W_{a\mu}\sum_i B_{ia}\frac{V_{i\mu}}{(\boldsymbol{B}\boldsymbol{W})_{i\mu}}}{\sum_k B_{ka}} - \frac{W_{a\mu}\sum_i B_{ia}}{\sum_k B_{ka}}$$

$$W_{a\mu} \leftarrow W_{a\mu} \frac{\sum_i \frac{B_{ia} V_{i\mu}}{(\boldsymbol{BW})_{i\mu}}}{\sum_k B_{ka}} \tag{15}$$

Analogamente, atualiza-se $\boldsymbol{B}$ por

$$B_{ia} \leftarrow B_{ia} \frac{\sum_\mu \frac{W_{a\mu} V_{i\mu}}{(\boldsymbol{BW})_{i\mu}}}{\sum_\nu W_{a\nu}} \tag{16}$$

Para atualizações calculadas pelas regras multiplicativas (15) e (16), prova-se em [13] que a divergência $D(\boldsymbol{V}||\boldsymbol{BW})$ é não crescente e se torna invariante se e somente se $\boldsymbol{B}$ e $\boldsymbol{W}$ forem um ponto fixo das atualizações. Ou seja, (15) e (16) descobrem um mínimo local da função de custo em (8).

Lee e Seung afirmam que as regras multiplicativas (11), (12), (15) e (16) são computacionalmente mais velozes que o método de Gradiente Projetado Alternado proposto em [15]. O que de fato se pode dizer, por simples observação das expressões, é que elas são de implementação bastante simples.

Pela facilidade de realização, os algoritmos propostos em [13] popularizaram o uso da NMF em diversas áreas de aplicação. Virtanen, por exemplo, utilizou-se das regras multiplicativas de Lee e Seung para realizar separação cega de fontes de áudio (BASS) [25]. Nessa ocasião, Virtanen utiliza a KLD como uma das parcelas da sua função de custo, acrescentando ainda uma parcela referente a um critério de continuidade temporal e outra de esparsidade. A função de custo composta é escrita

$$f(\boldsymbol{V}, \boldsymbol{B}, \boldsymbol{W}) = f_r(\boldsymbol{V}, \boldsymbol{B}, \boldsymbol{W}) + \alpha f_t(\boldsymbol{W}) + \beta f_s(\boldsymbol{W}) \tag{17}$$

onde $f_r(\boldsymbol{V}, \boldsymbol{B}, \boldsymbol{W}) = D(\boldsymbol{V}||\boldsymbol{BW})$, $f_t(\boldsymbol{W})$ representa o critério de continuidade temporal e $f_s(\boldsymbol{W})$, o termo de esparsidade. As constantes $\alpha, \beta \geq 0$ ponderam as parcelas de $f(\boldsymbol{V}, \boldsymbol{B}, \boldsymbol{W})$.

Visto que o único termo que afeta B é a própria KLD, sua regra de atualização mantém-se idêntica a (16) e é apresentada a seguir numa forma compacta:

$$\boldsymbol{B} \leftarrow \boldsymbol{B} \cdot \frac{\frac{\boldsymbol{V}}{\boldsymbol{BW}} \boldsymbol{W}^T}{\boldsymbol{1W}^T} \tag{18}$$

onde $\boldsymbol{1}$ é uma matriz $(m \times n)$ preenchida de números 1. O produto representado pelo símbolo $\cdot$ e todas as divisões são efetuadas elemento a elemento.

A atualização de $\boldsymbol{W}$ é derivada a partir do gradiente $\nabla f(\boldsymbol{V}, \boldsymbol{B}, \boldsymbol{W})$ da função de custo (17):

$$\nabla f(\boldsymbol{V}, \boldsymbol{B}, \boldsymbol{W}) = \nabla f_r(\boldsymbol{V}, \boldsymbol{B}, \boldsymbol{W}) + \alpha \nabla f_t(\boldsymbol{W}) + \beta \nabla f_s(\boldsymbol{W}) \tag{19}$$

Pode-se reescrever o gradiente da função de custo como a soma

$$\nabla f(\boldsymbol{V},\boldsymbol{B},\boldsymbol{W}) = \nabla f^{+}(\boldsymbol{V},\boldsymbol{B},\boldsymbol{W}) + \nabla f^{-}(\boldsymbol{V},\boldsymbol{B},\boldsymbol{W}) \tag{20}$$

em que os sobrescritos + e – indicam, respectivamente, os elementos positivos e os negativos da expressão, $\nabla f^{+}(\boldsymbol{V},\boldsymbol{B},\boldsymbol{W}) = \nabla \mathrm{f}_r^{+}(\boldsymbol{V},\boldsymbol{B},\boldsymbol{W}) + \alpha\nabla \mathrm{f}_t^{+}(\boldsymbol{W}) + \beta\nabla \mathrm{f}_s^{+}(\boldsymbol{W})$ e $\nabla f^{-}(\boldsymbol{V},\boldsymbol{B},\boldsymbol{W}) = \nabla \mathrm{f}_r^{-}V,(\boldsymbol{B},\boldsymbol{W}) + \alpha\nabla \mathrm{f}_t^{-}(\boldsymbol{W}) + \beta\nabla \mathrm{f}_s^{-}(\boldsymbol{W})$.

A regra de atualização de $\mathbf{W}$ é, então, dada por:

$$\boldsymbol{W} \leftarrow \boldsymbol{W} \cdot \frac{-\nabla f^{-}(\boldsymbol{V},\boldsymbol{B},\boldsymbol{W})}{\nabla f^{+}(\boldsymbol{V},\boldsymbol{B},\boldsymbol{W})} \tag{21}$$

onde novamente o produto $\cdot$ e a divisão são calculados elemento a elemento.

Não há provas de que as regras multiplicativas (18) e (21) reduzam a função de custo (17) – porque, de fato, elas nem sempre a reduzem. Contudo, para a aplicação em [25], constatou-se empiricamente que a probabilidade de que haja aumento em (17) é extremamente baixa e só ocorre quando α assume um valor muito alto. Assim, considerou-se o método suficiente para a minimização de $f(\boldsymbol{V},\boldsymbol{B},\boldsymbol{W})$.

Févotte, por sua vez, adaptou as regras multiplicativas para proceder à análise de música, em [7]. Nesse contexto, considerou mais adequado o uso da divergência de Itakura-Saito (ISD) como função de custo:

$$f(\boldsymbol{V},\boldsymbol{B},\boldsymbol{W}) = D_{IS}(\boldsymbol{V}||\boldsymbol{BW}) = \sum_{ij}\left(\frac{V_{ij}}{(BW)_{ij}} - log\frac{V_{ij}}{(BW)_{ij}} + 1\right) \tag{22}$$

Assim, derivou, à semelhança do que fizeram Lee e Seung, novas atualizações para $\boldsymbol{B}$ e $\boldsymbol{W}$:

$$\boldsymbol{W} \leftarrow \boldsymbol{W} \cdot \frac{\boldsymbol{B}^{T}[(\boldsymbol{BW})^{\cdot -2} \cdot \boldsymbol{V}]}{\boldsymbol{B}^{T}(\boldsymbol{BW})^{\cdot -1}} \tag{23}$$

$$\boldsymbol{B} \leftarrow \boldsymbol{B} \cdot \frac{[(\boldsymbol{BW})^{\cdot -2} \cdot \boldsymbol{V}]\boldsymbol{W}^{T}}{(\boldsymbol{BW})^{\cdot -1}\boldsymbol{W}^{T}} \tag{24}$$

onde, o produto $\cdot$, as potenciações $(\,)^{\cdot -\alpha}$ e todas as divisões são calculados elemento a elemento.

Nenhuma prova matemática foi apresentada para garantir que as expressões (23) e (24) façam $\boldsymbol{B}$ e $\boldsymbol{W}$ convergirem a um mínimo local da função de custo (22), mas apenas se observou que seu valor decresceu continuamente durante as experimentações práticas.

Outros algoritmos, baseados em diversas técnicas, são apresentados em [8], [11], [12] e [16].

2.4. A NMF para AuSS

Destaca-se, agora, a aplicação da NMF para AuSS. Seja $v(t) = v^1(t) + \cdots + v^K(t)$ uma mistura instantânea aditiva de K sinais de áudio. Considere-se, ainda, que a mistura é monofônica e não é afetada por efeitos como reverberação, eco, espalhamento, distorções etc. Sabe-se que o espectro complexo dos sinais $v^1(t), \ldots, v^K(t)$ soma-se linearmente. Contudo, devido ao fato de o espectro de fase de sinais de áudio ser, em geral, muito difícil de estimar e de o sistema humano de percepção auditiva ser relativamente insensível à fase, pode-se aproximar a soma do espectro complexo pela soma do espectro de magnitude [25].

Considere-se, então, $\boldsymbol{V}$ como o espectro de magnitude da mistura $v(t)$, aproximado pela soma dos espectros de magnitude de cada fonte $v^k(t)$. Cada coluna de $\boldsymbol{V}$ abriga os coeficientes do espectro de magnitude $\boldsymbol{v_t}$ de um quadro do sinal $v(t)$. O problema de AuSS pode ser formulado como a decomposição de $\boldsymbol{v_t}$ em:

$$\boldsymbol{v_t} \approx \sum_{j=1}^{r} w_{jt} \boldsymbol{b_j} \tag{25}$$

onde r é a quantidade de vetores-base em $\boldsymbol{B}$ e w_{jt} é o peso da base $\boldsymbol{b_j}$ no instante t.

Denote-se por *componente* o par formado pelo vetor-base $\boldsymbol{b_j}$ e pelo seu respectivo conjunto de pesos $\{w_{jt}\}_{t=1,\ldots,T}$, em que T é o número de quadros em V. Cada fonte de áudio é modelada como a soma de uma ou mais componentes. Executa-se a AuSS decompondo-se $\boldsymbol{V} \approx \boldsymbol{BW}$ e, então, agrupando-se as componentes para formar as fontes.

Vale ressaltar que todos os elementos de $\boldsymbol{V}$, um espectro de magnitude, são não negativos. Se cada fonte de áudio é modelada pelo seu próprio espectro de magnitude, então seus elementos também serão todos maiores ou iguais a zero. Faz sentido, portanto, restringir que a decomposição de $\boldsymbol{V}$ no produto $\boldsymbol{BW}$ possua exclusivamente elementos não negativos e o problema que se quer solucionar é uma NMF.

A dificuldade que se encontra para solucionar esse problema, além da já discutida convergência dos algoritmos para NMF, reside em determinar um valor para r, a quantidade de componentes (ou de vetores-base) da decomposição. Por mais que se saiba a quantidade K de fontes de áudio presentes na mistura, nada garante que o algoritmo para a NMF decomporia $\boldsymbol{V}$ em $r = K$ componentes equivalentes às K fontes. Em geral, faz-se $r > K$.

Agora, com $r > K$, levanta-se um novo problema: como agrupar os componentes nas suas respectivas fontes. Uma vez que não há informação sobre as fontes individualmente, dever-se-ia aplicar algum método não supervisionado de agrupamento. Nesse caso, mesmo que a NMF seja capaz de separar a mistura em componentes que pertençam a uma única fonte de áudio, o método de agrupamento ainda poderia deteriorar os resultados.

2.5. A NMF Supervisionada para AuSS

Motivada pelas dificuldades descritas na seção anterior – definição da quantidade de componentes e agrupamento das componentes nas respectivas fontes –, surge uma versão modificada da NMF: a NMF supervisionada. Nesse caso, busca-se a solução para o seguinte problema: dadas duas matrizes $\boldsymbol{V}$ e $\boldsymbol{B}$ de elementos não negativos, obter uma matriz $\boldsymbol{W}$, também de elementos não negativos, tal que $\boldsymbol{V} \approx \boldsymbol{BW}$. Na forma de um problema de otimização, queremos solucionar:

$$\min_{\boldsymbol{W} \geq 0} f(\boldsymbol{V}, \boldsymbol{B}, \boldsymbol{W}) \tag{26}$$

onde $f(\boldsymbol{W})$ é a função de custo que mede a qualidade da decomposição $\boldsymbol{V} \approx \boldsymbol{BW}$. A princípio, o problema (26) poderia ser resolvido de forma exata:

$$\boldsymbol{V} = \boldsymbol{BW} \Rightarrow \boldsymbol{B}^{-1}\boldsymbol{V} = \boldsymbol{B}^{-1}\boldsymbol{BW} \Longrightarrow \boldsymbol{B}^{-1}\boldsymbol{V} = \boldsymbol{IW} \Longrightarrow \boldsymbol{B}^{-1}\boldsymbol{V} = \boldsymbol{W} \tag{27}$$

onde $\boldsymbol{B}^{-1}$ representa a pseudo-inversa da matriz $\boldsymbol{B}$, uma vez que $\boldsymbol{B}$ não é necessariamente quadrada.

Contudo, a solução $\boldsymbol{W} = \boldsymbol{B}^{-1}\boldsymbol{V}$ não garante que $\boldsymbol{W}$ tenha somente entradas não negativas. Assim, para solucionar (26), podem-se utilizar os mesmos algoritmos descritos na seção 2.3, bastando que apenas se itere sobre a matriz $\boldsymbol{W}$, mantendo-se sempre fixa a matriz $\boldsymbol{B}$.

A ideia por detrás da NMF supervisionada é a inserção de conhecimento *a priori* no algoritmo de fatoração, isto é, passa-se a direcionar ou a supervisionar a decomposição $\boldsymbol{V} \approx \boldsymbol{BW}$, anteriormente "livre". Tal conhecimento é inserido treinando-se a matriz de vetores-base $\boldsymbol{B}$ antes de proceder à fatoração, que agora consistirá somente no cálculo da matriz de pesos.

As colunas da matriz $\boldsymbol{B}$, portanto, podem ser interpretadas como modelos das fontes $\{v^1(t), \ldots, v^k(t), \ldots, v^K(t)\}$ que compõem a mistura. Escreve-se

$$\boldsymbol{B} = [\boldsymbol{B}^1 \ldots \boldsymbol{B}^k \ldots \boldsymbol{B}^K] \tag{28}$$

onde $\boldsymbol{B}^k$ é uma matriz $(m \times r^k)$, cujos vetores-base modelam a fonte $v^k(t)$. Assim, a matriz $\boldsymbol{B}$ tem dimensões $(m \times r)$, com $r = r^1 + \cdots + r^K$.

Em um dado instante t, o espectro da k-ésima fonte será aproximado por

$$\boldsymbol{v_t^k} \approx \sum_{j=R+1}^{R+1+r^k} w_{jt}\boldsymbol{b_j} \tag{29}$$

em que $R = \sum_{l=1}^{k-1} r^l$. Da expressão (29), percebe-se que a questão do agrupamento das componentes nas suas respectivas fontes é sanada, uma vez que se tem predefinido que os vetores-base posicionados em $\{\boldsymbol{b}_{R+1}, \dots, \boldsymbol{b}_{R+1+r^k}\}$ modelam o espectro da fonte $v^k(t)$. Por outro lado, persiste a dificuldade de se definir um valor para a dimensão r. Na NMF supervisionada, r representa não somente a quantidade de componentes a se decompor a mistura, mas a quantidade de vetores-base necessária para que se tenha um bom modelo para cada fonte.

Tem-se, agora, a questão de se modelar cada fonte de maneira eficaz. Idealmente, o modelo $\boldsymbol{B}^k$ deve ser capaz de generalizar o espectro da fonte $v^k(t)$ e nenhuma outra mais. Assim, a aproximação (29) conterá todo o conteúdo da fonte $v^k(t)$ e não carregará nenhum conteúdo das fontes $\{v^i(t)\}_{i\neq k}$. Ainda, é interessante que a modelagem seja também eficiente, isto é, que $\boldsymbol{B}^k$ utilize tantos menos vetores quanto seja possível.

Uma maneira de se construírem as matrizes $\boldsymbol{B}^k$ é treinando-as com a NMF não supervisionada [7]. Nesse caso, toma-se um conjunto de treino de sinais gerados pela fonte $v^k(t)$, computam-se seus espectros de magnitude e decompõem-se-os iterando sobre $\boldsymbol{B}$ e sobre $\boldsymbol{W}$, com dimensão interna igual a r^k. Então, toma-se $\boldsymbol{B}^k$ como os vetores-base resultantes dessa fatoração. Denominar-se-ão as bases geradas por essa técnica como bases *treinadas pela* NMF. Outra maneira é compor cada $\boldsymbol{B}^k$ diretamente com exemplos de sinais provenientes da fonte $v^k(t)$ [20]. Nesse caso, computa-se o espectro de magnitude do conjunto de treino e simplesmente se escolhem r^k vetores para compor $\boldsymbol{B}^k$. Repete-se o procedimento escolhido para cada uma das fontes e se compõe $\boldsymbol{B}$ como em (28). As bases geradas por essa técnicas serão chamadas bases *compostas por exemplares*.

Em ambas as técnicas, a dimensão r continua indefinida. Em geral, seu valor ótimo é obtido empiricamente, avaliando-se o compromisso entre a qualidade da modelagem das fontes e o tempo de processamento (ou a memória computacional necessária).

2.6. A Sintetização dos Sinais das Fontes

Uma vez aplicada a NMF supervisionada sobre uma mistura, o que se obtém é uma matriz $\boldsymbol{W}$ com pesos que ponderam os vetores-base da matriz $\boldsymbol{B}$, de maneira a aproximar a matriz $\boldsymbol{V}$. Perceba-se que a estrutura inerente a B, descrita em (28), implica que $\boldsymbol{W}$ será da seguinte forma:

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}^1 \\ \vdots \\ \boldsymbol{W}^k \\ \vdots \\ \boldsymbol{W}^K \end{bmatrix} \tag{30}$$

onde, se $\boldsymbol{V}$ é $(m \times n)$, $\mathbf{W}^{\boldsymbol{k}}$ é uma matriz $(r^k \times n)$ que representa os pesos indexados à k-ésima fonte.

Dessa forma, a maneira mais direta de se calcular a contribuição de cada fonte v^k é fazendo

$$\widehat{\boldsymbol{V}}^k = \boldsymbol{B}^k \boldsymbol{W}^k \tag{31}$$

em que $\widehat{\boldsymbol{V}}^k$ é a matriz $(m \times n)$ que aproxima o espectro de magnitude da fonte v^k.

Ressalta-se, então, que a fatoração de $\boldsymbol{V}$ em $\boldsymbol{BW}$ não é exata, mas uma aproximação. Assim, a somatória de todos os $\widehat{\boldsymbol{V}}^k$ não é exatamente igual a $\boldsymbol{V}$, como seria o ideal (postas juntas, as contribuições de todas as fontes deveriam reconstituir a mistura). Uma abordagem mais interessante para obter a contribuição individual de cada fonte seria fazer

$$\widehat{\boldsymbol{V}}^k = \boldsymbol{V} \cdot \frac{\boldsymbol{B}^k \boldsymbol{W}^k}{\boldsymbol{BW}} \tag{32}$$

com o produto $\cdot$ e a divisão calculados elemento a elemento [20].

A expressão (32) faz $\widehat{\boldsymbol{V}}^k$ ser diretamente derivado do espectro original $\boldsymbol{V}$, ponderando-se-o com o fator $\boldsymbol{B}^k \boldsymbol{W}^k$ referente à fonte k, normalizado pela aproximação $\boldsymbol{BW}$. Em outras palavras, filtra-se o espectro $\boldsymbol{V}$ com um filtro $\boldsymbol{B}^k \boldsymbol{W}^k / (\boldsymbol{BW})$.

Feita a separação das fontes, atenta-se ao fato de, durante a aplicação da NMF, ter-se trabalhado apenas com espectros de magnitude. Resta retornar os sinais obtidos ao domínio do tempo e, para tal, é necessário estimar também a fase que acompanha cada $\widehat{\boldsymbol{V}}^k$. Ephraim e Malah provaram, em 1984, que a melhor estimativa para o espectro de fase – em uma situação em que um sinal de voz é corrompido por ruído aditivo descorrelacionado e que somente está disponível o sinal ruidoso – é a fase do próprio sinal ruidoso [6]. Assim, tomar-se-á diretamente o espectro de fase da mistura $\boldsymbol{V}$, anterior à separação.

3. Metodologia

Neste trabalho, serão realizados experimentos visando estudar o potencial da NMF para a resolução do problema apresentado na seção 1.1. Para esse fim, optou-se por utilizar as regras multiplicativas de Lee e Seung [13]. Apesar de outras técnicas, mais recentes [2], [8], [11], [12], [16], já terem sido publicadas, o algoritmo de [13] ainda é muito atrativo, principalmente pela simplicidade de sua implementação e pela garantia de sua convergência para um mínimo local. Outras questões influenciaram a tomada de decisão de uso desse algoritmo: o método matemático em que se baseiam as regras multiplicativas (método do gradiente) é de fácil compreensão e amplamente difundido como técnica de aprendizagem automática; o algoritmo pode ser aplicado para qualquer matriz $\boldsymbol{V}$ de elementos não negativos, sem nenhuma exigência extra às suas características (como dimensão, simetria, determinante etc); qualquer restrição complementar que se queira impor sobre a NMF pode ser incorporada às regras multiplicativas, como fizeram Virtanen e Févotte em [25] e [7], respectivamente, bastando que se as escrevam como parte da função de custo do algoritmo; a adaptação do algoritmo para transformar a NMF em um método supervisionado é também incrivelmente simples e consiste em retirar a etapa de atualização da matriz de bases, iterando-se apenas sobre a matriz de pesos; por tamanha popularidade, o algoritmo de Lee e Seung já foi aplicado a sinais de voz e de música, assim como a AuSS [7], [20], [25], e demonstrou resultados positivos.

3.1. Sinais de Áudio

Todos os sinais de voz utilizados durante as experimentações deste trabalho foram extraídos do corpus TIMIT [27]. A base é formada por pequenas frases em inglês lidas por locutores de ambos os sexos e residentes de diferentes regiões dialetais dos Estados Unidos.

Todos os sinais de música utilizados ao longo deste projeto são canções do gênero Jazz, retiradas do álbum 25 Anos, da Traditional Jazz Band. Nenhuma das faixas é cantada, isto é, não há voz na música.

Tanto os sinais de música quanto os de voz foram amostrados a uma taxa de 16 kHz.

3.2. Espectro dos Sinais

Como descrito nas seções 2.4 a 2.6, utilizam-se espectros de magnitude para proceder à AuSS. Os espectros foram obtidos pela computação da Transformada de Fourier de Curto Prazo (STFT). Dividiu-se cada sinal em quadros de 20 ms de duração (o equivalente a 320 amostras) e aplicou-se a eles uma janela de Hanning. A STFT foi calculada com 50% de sobreposição, tomando-se a Transformada de Fourier Discreta (DFT) com todos os 320 pontos de cada quadro. Tanto as frequências positivas quanto as negativas foram utilizadas.

A Transformada Inversa de Fourier de Curto Prazo (ISTFT) é calculada com os mesmos parâmetros da STFT, de forma a se obter reconstrução perfeita de um sinal ao qual se tenha aplicado a STFT.

3.3. Critérios de Avaliação

Os resultados obtidos durante as experimentações deste trabalho foram avaliados por duas ferramentas: o AUDIMUS.MEDIA, com auxílio do NIST Score Lite (SCLITE) [29], e o BSS_EVAL Toolbox [27].

O AUDIMUS.MEDIA é o ASR que motivou este projeto. Assim, nada mais lógico do que avaliar os resultados das técnicas aqui empregadas pelo seu próprio desempenho. Os sinais que se desejava avaliar foram salvos em formato de onda (.wav) e transferidos para o AUDIMUS.MEDIA. As transcrições por ele feitas, foram comparadas a um texto de referência utilizando o método de alinhamento do SCLITE e, então, obtiveram-se quatro medidas de desempenho: acertos (palavras corretamente transcritas pelo ASR), substituições (palavras substituídas por outras), inserções (palavras adicionadas ao texto) e omissões (palavras não identificadas no áudio), todas em valores percentuais em relação à quantidade total de palavras no texto de referência. Para esse projeto, utilizar-se-á apenas o percentual de acertos.

O BSS_EVAL Toolbox é um conjunto de rotinas escritas em MATLAB destinadas, originalmente, a avaliar resultados de separações cegas de fontes. Em aplicações reais, não é possível utilizar essa ferramenta, uma vez que ela exige conhecimento completo dos sinais individuais que compuseram a mistura. Para este trabalho, em que as misturas são geradas artificialmente, os índices de desempenho do BSS_EVAL Toolbox serão bastante interessantes.

Seja $s(t) = s_1(t) + \cdots + s_n(t)$ uma mistura de n fontes de áudio. Seja também $\hat{s}_i(t)$ a estimativa da fonte $s_i(t)$, obtida a partir de um método de separação de fontes. Se se deseja avaliar o resultado da separação de uma fonte i, as entradas do sistema são os sinais $\{s_j(t)\}_{j=1,\dots,n}$ e $\hat{s}_i(t)$.

Inicialmente, decompõe-se o sinal estimado para a fonte i em uma soma de quatro fatores:

$$\hat{s}_i(t) = s_{target}(t) + e_{interf}(t) + e_{noise}(t) + e_{artif}(t) \tag{33}$$

em que $s_{target}(t)$ representa uma deformação permitida do sinal $s_i(t)$, $e_{interf}(t)$ refere-se aos sinais relativos às fontes $\{s_j(t)\}_{j\neq i}$, $e_{noise}(t)$ refere-se a ruídos (que não as outras fontes) e $e_{artif}(t)$ refere-se a "artefatos" introduzidos pelo algoritmo de separação. Nas experimentações desse projeto, não se introduz nenhum tipo de ruído além da música nas misturas. Então, daqui para frente, considere-se $e_{noise}(t) \equiv 0$.

A partir dos fatores de (33), computam-se as seguintes razões de energia:

$$SDR \coloneqq 10\,log_{10}\frac{\left\|s_{target}\right\|^2}{\left\|e_{interf}+e_{artif}\right\|^2} \tag{34}$$

$$SIR \coloneqq 10\,log_{10}\frac{\left\|s_{target}\right\|^2}{\left\|e_{interf}\right\|^2} \tag{35}$$

$$SAR \coloneqq 10\,log_{10}\frac{\left\|s_{target}+e_{interf}\right\|^2}{\left\|e_{artif}\right\|^2} \tag{36}$$

O SDR mede a distorção do sinal estimado em relação às interferências e aos “artefatos”; o SIR avalia quão livre das outras fontes está o sinal $\hat{s}_i(t)$; e o SAR quantiza os “artefatos” introduzidos pela separação.

As razões (34 – 36) são particularmente interessantes para esse trabalho porque permitem avaliar que tipos de distorção, “artefato” ou interferência são mais relevantes para a degradação do desempenho do AUDIMUS.MEDIA.

3.4. Algoritmos Avaliados

Nesse trabalho, serão avaliados dois métodos de construção das bases de voz e de música (modelagem dos sinais envolvidos na mistura) e duas funções de custo para o problema de otimização (1) e (26). Comparam-se os resultados obtidos a partir de bases compostas por exemplares e de bases treinadas pela NMF. Quanto às funções de custo, serão avaliados os resultados oriundos de NMFs baseadas na Divergência de Kullback-Leibler (6), chamada KL-NMF, e na Divergência de Itakura-Saito (22), denominada IS-NMF.

Em todos os casos, utilizam-se as regras multiplicativas para a decomposição em matrizes não negativas. Para a KL-NMF, aplicam-se as atualizações (15) e (16) e, para a IS-NMF, aplicam-se (23) e (24). Ressalta-se que a função de custo utilizada na etapa de separação é a mesma utilizada na fase de treinamento, para as bases treinadas pela NMF. No caso das bases compostas por exemplares, a função de custo somente é considerada na etapa de separação.

Ainda, serão comparados dois critérios de parada para o algoritmo da NMF. Visando controlar mais facilmente o tempo computacional exigido para cada teste e/ou treino e tendo que lidar com máquinas não projetadas para esse tipo de processamento pesado, o critério de parada da NMF é determinado em quantidade de iterações – em vez de ser baseado diretamente no valor da função de

custo ou no quão estático é o ponto $(\boldsymbol{B}, \boldsymbol{W})$. Compara-se, então, a aplicação dos algoritmos para 100 e para 500 atualizações.

3.5. Fase de Treino

A NMF supervisionada exige uma fase de treino, dedicada à construção das matrizes de vetores-base de voz ($\boldsymbol{B_s}$) e de música ($\boldsymbol{B_m}$), que serão fixadas durante o processo de separação. Os processos de construção dessas matrizes foram descritos na seção 2.5 deste trabalho.

O CONJUNTO DE TREINO

O corpus TIMIT está dividido entre um conjunto de treino e outro de testes. Dentro do conjunto de treino, há oito regiões dialetais, cada uma delas com certo número de locutores de cada sexo. Cada locutor possui gravadas dez frase de duração entre 2 e 5 segundos. O conjunto de treino do corpus de fala utilizado nas experimentações deste projeto foi composto, então, tomando-se as frases gravadas por dois locutores do sexo masculino e um do sexo feminino, de cada uma das oito regiões dialetais da divisão de treino da base do TIMIT. A escolha dos locutores foi feita de maneira aleatória e a proporção de 2:1 entre os sexos foi motivada pela composição do conjunto reduzido de testes sugerido na documentação da base de dados.

Reuniram-se, assim, 240 frases, equivalentes a cerca 12 minutos de treino. Computou-se o espectro de magnitude do conjunto de treino e agrupou-se o resultado em uma única matriz de treino, com 72 282 vetores.

Construiu-se ainda um segundo conjunto de treino, formado à semelhança do primeiro, mas com cerca de 50% mais de informação. Para tal, tomou-se o primeiro conjunto de treino e concatenou-se a ele as frases de mais 8 locutores do sexo masculino e 4 locutores do sexo feminino, tomados aleatoriamente dentre aqueles ainda não utilizados na divisão de treinos do TIMIT. Esse segundo conjunto de treino reuniu 18 minutos de treinamento para as bases de fala.

Relativamente ao corpus de música, selecionaram-se 15 das 16 faixas do álbum de Jazz, reservando-se apenas uma para formar o conjunto de testes. De cada uma dessas 15 faixas, extraíram-se, aleatoriamente, 50 segundos, contabilizando-se 12,5 minutos de treinamento.

BASES COMPOSTAS POR EXEMPLARES

O tamanho das bases compostas por exemplares é dado simplesmente pela quantidade de vetores tomados da matriz de treino. Para as bases de voz, compuseram-se bases com $\{80, 160, 500, 1000, 2000, 3000, 5000\}$ vetores. As bases de música, por sua vez, foram construídas com $\{160, 1000, 2000, 3000\}$ vetores. A escolha dos vetores-base foi feita de maneira aleatória, mas garantindo-se sempre que uma base maior contivesse as bases menores, isto é: $B_{80} \subset B_{160} \subset \cdots \subset$

B_{5000}, onde B_n representa uma base com n vetores. Note-se que as bases compostas por exemplares representam um subconjunto dos vetores da matriz de treino.

O procedimento de composição das bases é o mesmo para as bases de música e para as bases de voz.

BASES TREINADAS PELA NMF

O tamanho das bases treinadas pela NMF é dado pela dimensão interna r do produto $\boldsymbol{BW}$. Novamente, construíram-se bases de voz com $r_s = \{80, 160, 500, 1000, 2000, 3000, 5000\}$ e bases de música com $r_m = \{160, 1000, 2000, 3000\}$. As bases foram treinadas tanto com a KL-NMF como com a IS-NMF. Também foi variado o critério de parada, entre 100 ou 500 iterações. Ressalta-se que, durante a fase de treino, é utilizada a NMF não supervisionada, iterando-se sobre ambas as matrizes $\boldsymbol{B}$ e $\boldsymbol{W}$ e tomando-se todos os vetores de $\boldsymbol{B}$.

O procedimento de composição das bases é o mesmo para as bases de música e para as bases de voz.

3.6. Fase de Testes

Durante a fase de testes, aplica-se a NMF supervisionada sobre um conjunto de teste (uma mistura de fala e música). Fixa-se a matriz $\boldsymbol{B} = [\boldsymbol{B_s}\, \boldsymbol{B_m}]$, onde $\boldsymbol{B_s}$ é uma das bases de voz e $\boldsymbol{B_m}$ é uma das bases de música. $\boldsymbol{B}$ é sempre formada por bases exclusivamente compostas por exemplares ou exclusivamente treinadas pela NMF. O critério de parada utilizado para criar a base $\boldsymbol{B_s}$ deve ter sido o mesmo utilizado para $\boldsymbol{B_m}$.

Em todos os testes executados para este projeto, fez-se a síntese dos sinais separados filtrando-se o espectro da mistura com a expressão (32) e estimando a fase da contribuição de cada fonte pela fase da mistura.

CONJUNTO DE TESTES

O conjunto de testes é formado por uma longa mistura de voz e de música. Para compô-lo tomou-se o conjunto reduzido de testes sugerido na documentação do corpus TIMIT, formado por 240 frases extraídas das oito regiões dialetais da divisão de testes do TIMIT, em proporção de dois locutores para uma locutora.

Para corromper a voz, utilizou-se a única faixa do álbum de Jazz não utilizada para compor as bases de música. Procedeu-se da seguinte maneira: tomou-se uma das frases selecionadas para formar o conjunto de testes; selecionou-se um trecho aleatório de música de igual comprimento; multiplicou-se o sinal de música por uma constante que fizesse a relação de energia entre os sinais (SMR) atingir um valor definido; somaram-se os dois sinais. Repetiu-se esse procedimento para todas as frases do conjunto e,

enfim, concatenaram-se todas as misturas. Foram construídos conjuntos de testes para $SMR = \{0, 5, 10, 15, 20\}\ dB$.

Em alguns dos testes, utilizaram-se como conjunto de testes apenas os sinais de voz ou apenas os sinais de música, em vez de uma mistura de ambos.

TESTE A: AVALIANDO O MODELO DE VOZ

Em um primeiro momento, realizaram-se testes para selecionar a configuração dos parâmetros *tamanho da base* e *quantidade de iterações* que melhor modelassem os sinais de voz. O conjunto de teste, nesse caso, é formado apenas pelos sinais de voz, livres de música. Aplica-se a ele a NMF supervisionada para $\boldsymbol{B} = \boldsymbol{B_s}$ e, então, avalia-se quão bem a base $\boldsymbol{B}_s$ é capaz de explicar um sinal de fala, em termos de SDR. Note-se que, como não há outras fontes (música) corrompendo o sinal de interesse (voz), a SIR será sempre infinita e a SAR será idêntica à SDR.

TESTE B: AVALIANDO O MODELO DE MÚSICA

O Teste B é análogo ao Teste A, mas aplicado apenas aos sinais de música, antes de serem misturados à fala. Define-se $\boldsymbol{B} = \boldsymbol{B_m}$ e analisa-se o conjunto de parâmetros *tamanho da base* e *quantidade de iterações* que melhor expliquem a música, em termos de SDR.

TESTE C: SEPARAÇÃO DE FONTES

Utilizando-se os melhores parâmetros obtidos nos testes preliminares A e B, finalmente, no Teste C, avalia-se a capacidade dos algoritmos e das bases de separarem a voz da música que a corrompe. Aqui, o conjunto de teste é de fato a mistura dos dois sinais e fixa-se $\boldsymbol{B} = [\boldsymbol{B_s}\ \boldsymbol{B_m}]$.

Visa-se avaliar não somente o algoritmo da NMF que executa de fato a separação das fontes, mas também o método escolhido para modelar cada sinal, isto é, para compor as bases. Assim, aplica-se o Teste C para as seguintes combinações de métodos de criação das bases e algoritmos de separação:

		Algoritmo de Separação	
		KL-NMF	IS-NMF
Método de Criação das Bases	Composta por Exemplares	x	x
	Treinada pela KL-NMF	x	
	Treinada pela IS-NMF		x

Tabela 1: Combinações de métodos de criações de bases e algoritmos de separação

Os resultados do Teste C são avaliados em termos de SIR, SAR, SDR e também do percentual de acertos da transcrição gerada pelo ASR.

4. Resultados

São oito fatores, entre parâmetros (tamanhos de $\boldsymbol{B_s}$ e de $\boldsymbol{B_m}$ e critério de parada da NMF) e métodos (bases compostas por exemplares, treinadas pela KL-NMF ou treinadas pela IS-NMF; separação realizada com a KL-NMF ou com a IS-NMF), cujas influências sobre os resultados da AuSS estão sob análise. Realizar todo o conjunto combinatório de testes possíveis seria excessivamente extenso e, por isso, inicialmente, voltaram-se os testes para a otimização dos parâmetros *tamanho das bases* e *número de iterações*. Entenda-se, aqui, *otimização* como a seleção do melhor valor para um parâmetro, dentre os valores testados. Nesses testes preliminares, utilizaram-se as misturas a $SMR = 0\ dB$ e os sinais de fala e de música não corrompidos um pelo outro ($SMR = \pm\infty$). Obtidos tais parâmetros – ou, ao menos, reduzida a gama de possibilidades –, realizaram-se novos testes, restritos aos melhores valores experimentados, para os restantes valores de SMR.

4.1. Bases de Exemplares e Separação por KL-NMF

Os primeiros testes experimentados utilizaram bases compostas por exemplares e, para realizar a separação, aplicou-se a KL-NMF. Iniciou-se pelo Teste A, avaliando-se os resultados da KL-NMF não supervisionada com todas as dimensões disponíveis de $\boldsymbol{B_s}$. Não foi possível, contudo, perceber padrão ou tendência que levasse a alguma ideia de valores ótimos nem para o tamanho da base de voz nem para o critério de parada da NMF.

Seguiu-se, então, para o Teste B, aplicado a todas as dimensões disponíveis de $\boldsymbol{B_m}$. Os resultados podem ser vistos nos gráficos a seguir:

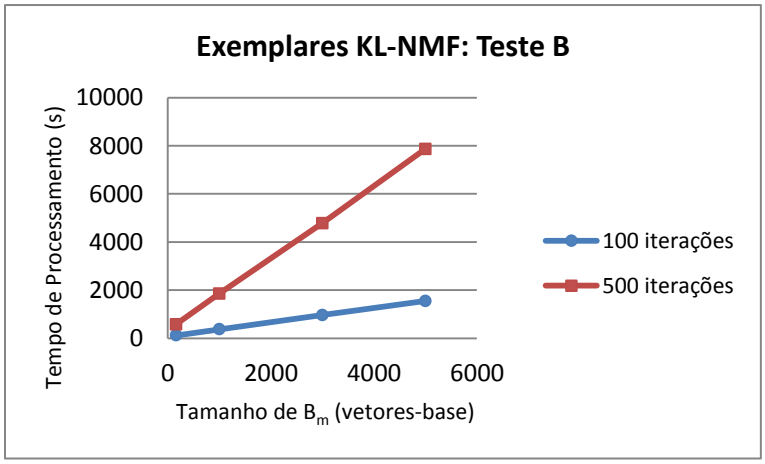

Figura 1: B_m composta por exemplares e separação realizada com KL-NMF. Tempo de Processamento do Teste B em função do tamanho de B_m.

Percebe-se pelos gráficos das Figuras Figura 1 e Figura 2 que o número de iterações utilizado como critério de parada da KL-NMF influi pouco em SDR, mas altera consideravelmente o tempo de

processamento do teste. A Figura 2 também indica $r = 1000$ como o melhor tamanho para a base de música nesse caso.

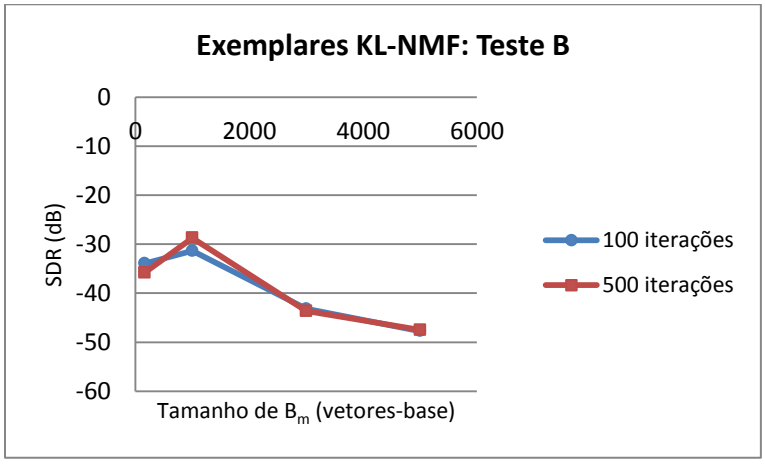

Figura 2: B_m composta por exemplares e separação realizada com KL-NMF. *Signal-to-Distortion Ratio* no Teste B em função do tamanho de B_m.

Executando-se, então, o Teste C para $\boldsymbol{B_m}$ com 1000 vetores-base, obtiveram-se os resultados das Figuras 3 a 6.

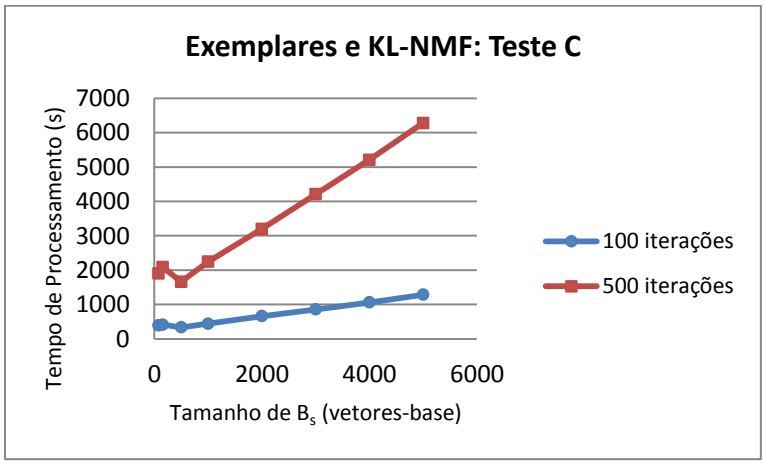

Figura 3: B_s composta por exemplares e separação realizada com KL-NMF. Tempo de Processamento do Teste C em função do tamanho de B_s.

Novamente, vê-se que o aumento no número de iterações pouco altera os resultados em termos de SIR, SDR e SAR, mas aumenta substancialmente o tempo de processamento. Assim, daqui em diante, os testes serão realizados apenas com 100 iterações como critério de parada.

Das Figuras 5 e 6, percebe-se uma tendência de que bases maiores gerem melhores resultados em termos de SDR e SAR. O gráfico $SIR\ x\ Tamanho\ da\ \boldsymbol{B_s}$, na Figura 4, porém, tem uma forma

curiosamente diferente dos gráficos de Figura 5 e Figura 6. Como os conjuntos de teste e de treino são iguais para todos os pontos plotados, desconfia-se que a anormalidade da Figura 4 tenha sido causada por uma inicialização “sortuda” da matriz de pesos. De qualquer maneira, buscando verificar tal hipótese, realizar-se-ão testes tanto para $\boldsymbol{B_s}$ com 160 quanto para 5000 vetores-base.

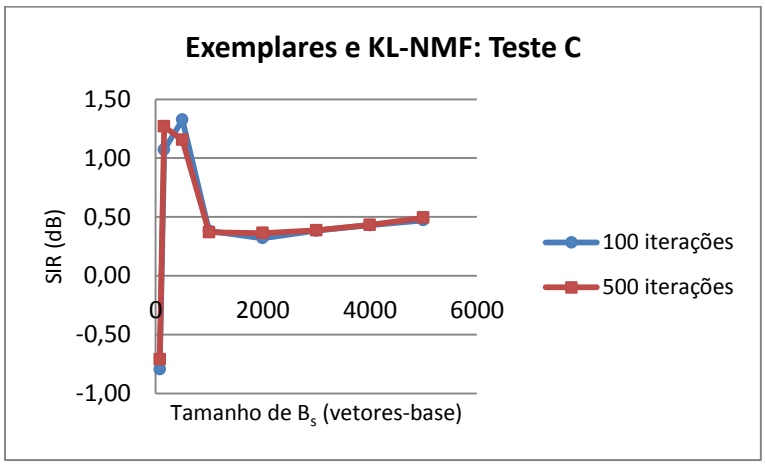

Figura 4: B_S composta por exemplares e separação realizada com KL-NMF. *Signal-to-Interferences Ratio* no Teste C em função do tamanho de B_S.

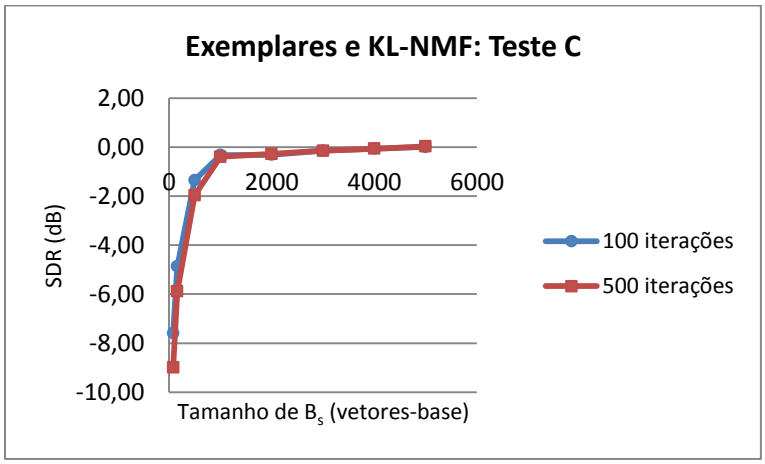

Figura 5: B_S composta por exemplares e separação realizada com KL-NMF. *Signal-to-Distortion Ratio* no Teste C em função do tamanho de B_S.

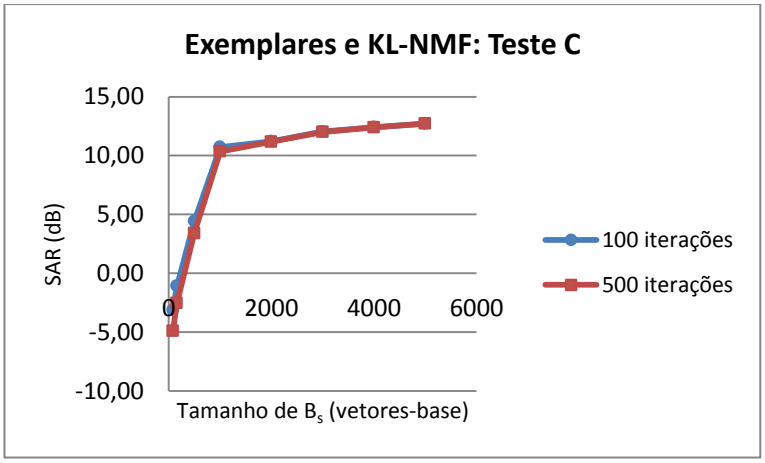

Figura 6: B_s composta por exemplares e separação realizada com KL-NMF. *Signal-to-Artifacts Ratio* no Teste C em função do tamanho de B_s.

4.2. Bases Treinadas e Separação com KL-NMF

Nesta segunda etapa de testes, treinaram-se as bases de voz e de música com o algoritmo não supervisionado da KL-NMF. Para a separação, utilizou-se a versão supervisionada do mesmo algoritmo. Como em 4.1, iniciou-se a bateria com o Teste A, mas já com o critério de parada fixado em 100 iterações. A Figura 7, abaixo, revela clara tendência de que, quanto mais vetores-base houver, melhor a base $\boldsymbol{B_s}$ explica a parcela de voz do conjunto de testes. Daqui para frente, portanto, fixar-se-á o tamanho de $\boldsymbol{B_s}$ a 5000 vetores-base.

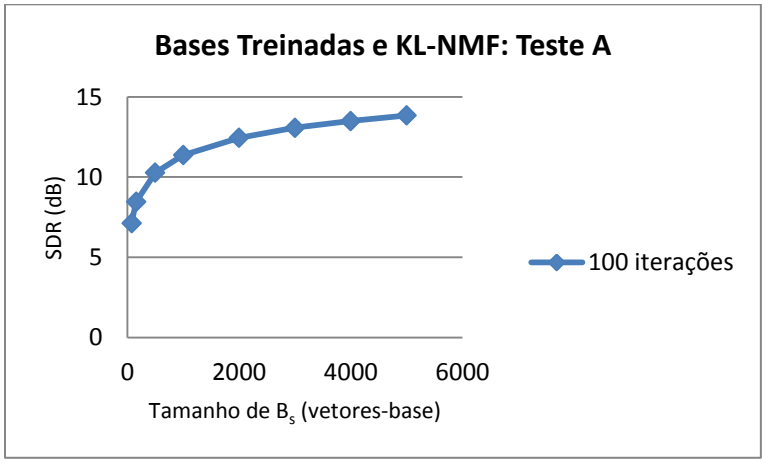

Figura 7: B_s treinada pela KL-NMF e separação realizada com KL-NMF. *Signal-to-Distortion Ratio* no Teste A em função do tamanho de B_s.

Partindo para o Teste B, observou-se novamente um gráfico sem tendências nítidas, representado na Figura 8. A melhor dimensão observada para compor $\boldsymbol{B_m}$ é 160. Outros testes serão realizados,

então, para $\boldsymbol{B_m}$ com 160 e também, para poder comparar com os resultados de 4.1, com 1000 vetores-base.

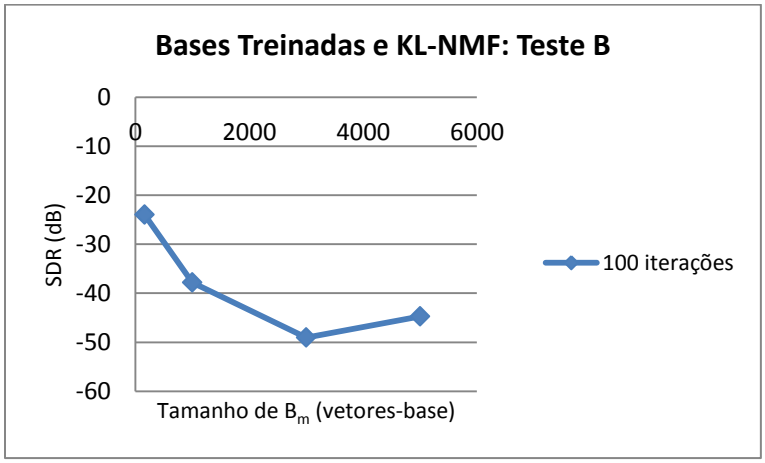

Figura 8: B_m treinada pela KL-NMF e separação realizada com KL-NMF. *Signal-to-Distortion Ratio* no Teste B em função do tamanho de B_{m}.

Os resultados obtidos pelo Teste C para as bases treinadas com a KL-NMF e para a separação das fontes na mistura também realizada pela KL-NMF serão apresentados na seção 4.3 a seguir.

4.3. Resultados Comparativos

Nesta seção do trabalho, realizaram-se teste que buscavam comparar os métodos de composição das bases e também os algoritmos aplicados para realizar a separação das fontes, utilizando os valores definidos para os parâmetros nas baterias de testes preliminares (número de iterações fixado em 100, tamanho de $\boldsymbol{B_s}$ fixado em 5000 vetores-base e tamanho de $\boldsymbol{B_m}$ igual a 160 ou a 1000 vetores-base). Aqui, foram processados os conjuntos de testes construídos a $SMR = \{0, 5, 10, 15, 20\}\, dB$ e os resultados foram avaliados também pelo desempenho do ASR.

Para facilitar a visualização dos gráficos a seguir, estabeleceu-se a seguinte nomenclatura para as séries: $[método\ de\ composição\ das\ bases] - [método\ de\ separação] - |B_m| = tamanho\ de\ Bm$, onde o método de composição das bases pode ser E, para bases compostas por exemplares, ou T, para bases treinadas pela NMF; e os métodos de separação podem ser KL-NMF ou IS-NMF. Ressalta-se que as bases treinadas pela NMF utilizam o mesmo algoritmo que aquele aplicado para a separação (isto é, não se treinam bases com a IS-NMF para, posteriormente, aplicá-las a uma separação com KL-NMF). A série *E-KL-NMF* $|B_m|$*=1000*, por exemplo, apresenta os resultados de uma separação feita com a KL-NMF e com bases compostas por exemplares.

A Figura 9, abaixo, estabelece graficamente uma comparação entre os resultados obtidos quando as bases foram compostas por exemplares e quando foram treinadas pela KL-NMF. Note-se que, nesse

quadro, a separação é sempre realizada pela KL-NMF. A linha preta contínua representa um valor de referência para os resultados. Visto que a música é a interferência que atinge a voz, a SIR pode ser interpretada como equivalente à SMR. Assim, a linha de referência é, em realidade, a reta $SIR = SMR$.

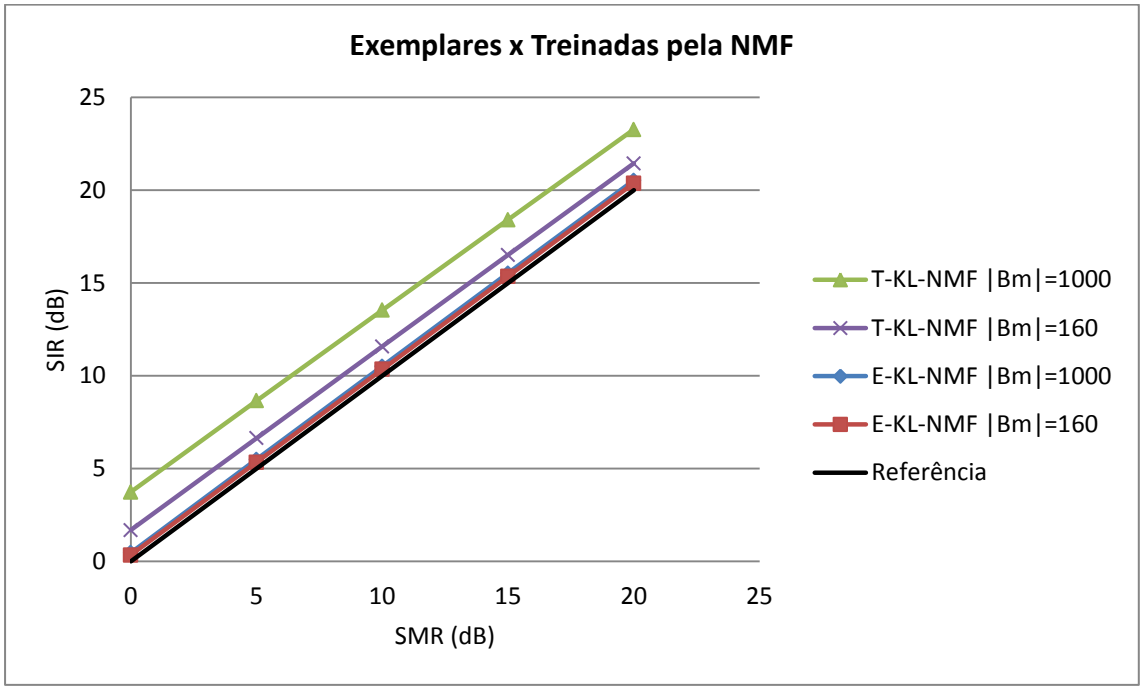

Figura 9: Comparação de resultados gerados por bases compostas por exemplares e bases treinadas pela KL-NMF. *Signal-to-Interferences Ratio* em função da *Speech-to-Music Ratio*.

Observa-se dessa comparação que todos os métodos melhoram, em termos da SIR, a composição da mistura. Em outras palavras, todos os métodos aplicados retiram pelo menos um pouco da música que corrompe a voz. Além disso, vê-se que o aumento no tamanho de $\boldsymbol{B_m}$ melhora o desempenho da separação no caso das bases treinadas pela NMF, mas não altera significativamente os resultados das bases compostas por exemplares. Fica evidente que as bases treinadas pela KL-NMF geram resultados superiores àqueles obtidos com as bases compostas por exemplares, mesmo quando a $\boldsymbol{B_m}$ treinada é muito menor que a $\boldsymbol{B_m}$ composta por exemplares.

A Figura 10 estabelece, em seguida, uma comparação entre os algoritmos KL-NMF e IS-NMF, tanto aplicados no treinamento das bases como na execução da separação. Novamente, verifica-se melhora na composição da mistura, em termos de SIR, com todos os métodos aplicados. Desta vez, contudo, não se percebe grandes diferenças de desempenho entre as funções de custo aplicadas. Assim, utilizar a Divergência de Kullback-Leibler ou a Divergência de Itakura-Saito gera resultados semelhantes. A maior diferença que se percebe no gráfico da Figura 10 é a mesma constatada com a comparação da Figura 9: bases treinadas pela NMF acarretam resultados melhores que as bases compostas por exemplares, qualquer que seja a função de custo utilizada.

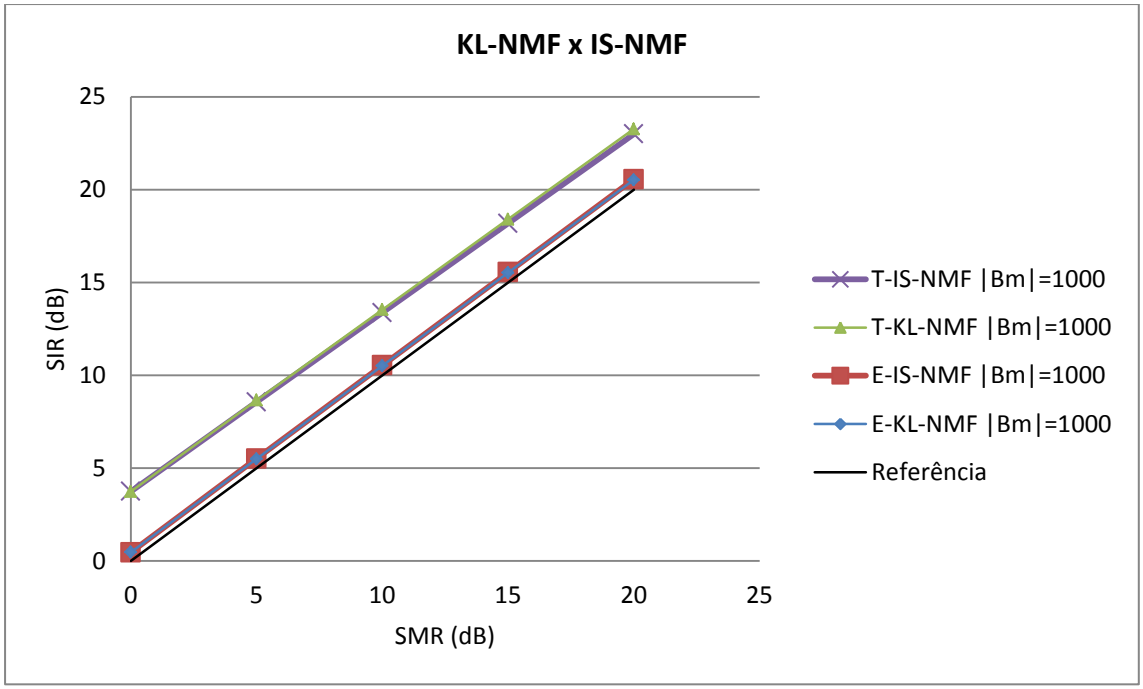

Figura 10: Comparação de resultados gerados pela aplicação dos algoritmos KL-NMF e IS-NMF, tanto para bases compostas por exemplares como para bases treinadas pela mesma NMF. *Signal-to-Interferences Ratio* em função da *Speech-to-Music Ratio*.

A seguir, na Figura 11, representa-se a mesma comparação feita na Figura 10, mas tomando-se como critério de desempenho o percentual de palavras corretamente transcritas pelo ASR. Note-se que a relação entre o desempenho do ASR e a SMR não é linear, como se verificava com a SIR. Portanto, conclui-se também que o percentual de acertos do ASR não se relaciona linearmente com a SIR e infere-se que há mais fatores (como a SAR e a SDR) que influenciem o desempenho do ASR. No gráfico da Figura 11, há duas linhas contínuas de referência: a primeira delas, superior, nomeada “Referência Apenas Voz”, indica o percentual de acertos do ASR para o caso em que não havia música corrompendo a fala, isto é, o máximo desempenho do ASR para esse conjunto de testes; a segunda, chamada “Referência”, representa o percentual de acertos do ASR para a mistura feita a uma determinada SMR, ou seja, o desempenho do ASR antes de se proceder à AuSS.

Atente-se, então, ao fato de que, apesar de todos os métodos de separação terem elevado a razão SIR, nem todos eles melhoraram o desempenho do ASR. Mais ainda, nenhum método é capaz de elevar o desempenho do ASR ao seu máximo, mesmo quando a mistura foi feita a uma SMR elevada (20 dB).

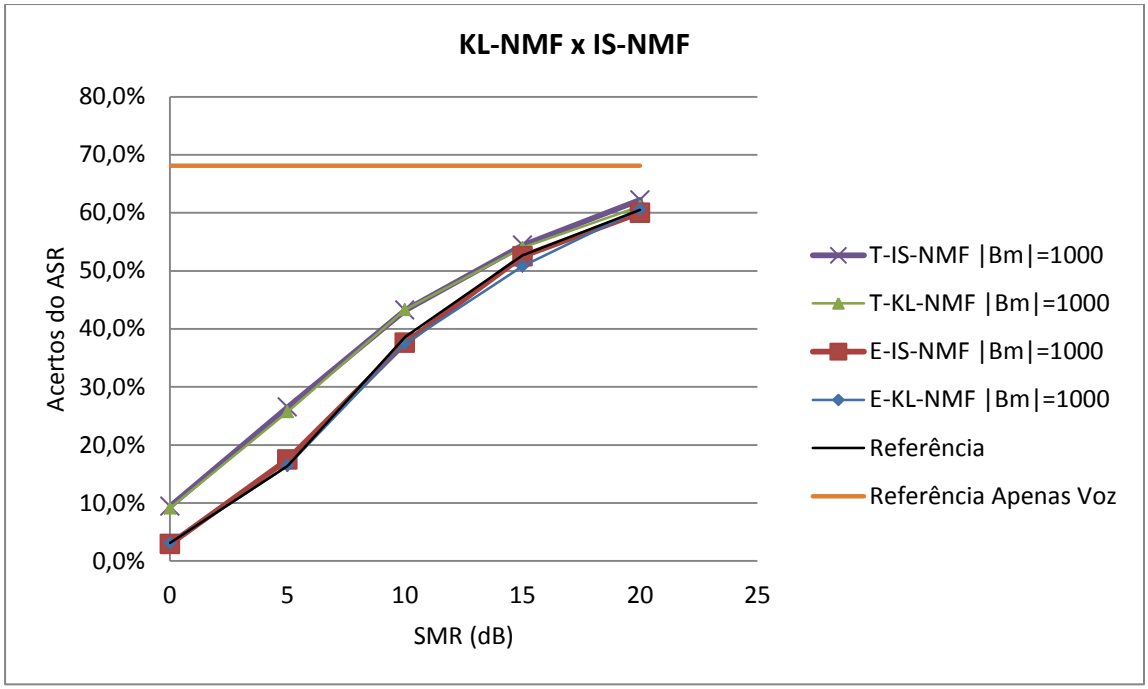

Figura 11: Comparação de resultados gerados pela aplicação dos algoritmos KL-NMF e IS-NMF, tanto para bases compostas por exemplares como para bases treinadas pela mesma NMF. Percentual de palavras corretamente transcritas pelo ASR em função da *Speech-to-Music Ratio*.

Por fim, realizou-se um último teste comparativo, com o intuito de se avaliar a influência de um aumento no conjunto de treino. Para isso, tomou-se o segundo conjunto de treinamento, 50% maior aquele utilizado nos testes até agora apresentados. Então, treinou-se $\boldsymbol{B_s}$ com a KL-NMF a partir desse novo conjunto. Manteve-se a mesma $\boldsymbol{B_m}$ utilizada anteriormente. Os resultados obtidos foram virtualmente idênticos àqueles alcançados com o conjunto menor de treinamento. O aumento no tamanho do conjunto de treino, portanto, não adicionou à base de voz informações que fossem capazes de melhorar seu desempenho no sentido de rejeitar mais da música corruptiva.

5. Discussão

Carência de recursos computacionais, de tempo – e possivelmente de paciência – fizeram com que as experimentações realizadas para este trabalho utilizassem conjuntos reduzidos de dados e fossem privadas de repetições. Os conjuntos de treinamento e de testes eram pequenos, se comparados aos aplicados por outros autores, e não houve repetição de nenhum dos experimentos (para tentar, por exemplo, contabilizar a aleatoriedade da inicialização da matriz de pesos nos algoritmos da NMF). Contudo, apesar de estatisticamente contestáveis, os resultados apresentados na seção anterior geraram inferências e conclusões interessantes e permitiram o levantamento de muitos pontos de discussão. A seguir, comenta-se sobre alguns deles.

5.1. Critério de Parada dos Algoritmos

Nesse trabalho, não se utilizou o critério de saída dos algoritmos da NMF mais intuitivo. Uma vez que o par ótimo para $(\boldsymbol{B},\boldsymbol{W})$ é um ponto de equilíbrio da função de custo $f(\boldsymbol{V},\boldsymbol{B},\boldsymbol{W})$, seria lógico aplicar um critério de parada que levasse tal aspecto em consideração – como a diferença entre os valores de $(\boldsymbol{B},\boldsymbol{W})$ ou de $f(\boldsymbol{V},\boldsymbol{B},\boldsymbol{W})$ de uma iteração para outra. O número de iterações, utilizado como critério de saída da NMF, porém, mostrou-se suficiente. Os valores testados foram muito mais críticos no tempo de computação dos algoritmos do que no desempenho dos mesmos.

5.2. Método de Composição das Bases

Compararam-se neste trabalho dois métodos de composição das bases, fixadas durante a etapa de separação (NMF supervisionada): composição por exemplares ou por treinamento com a NMF. Discutem-se a seguir algumas características de cada um dos dois métodos.

BASES COMPOSTAS POR EXEMPLARES

A fase de treinamento pelo método de composição de bases por exemplares é extremamente rápida de simples. Pode ser interpretada como a formulação de uma espécie de dicionário, que será utilizado para identificar as parcelas da mistura referentes a cada fonte treinada. Os resultados gerados por esse método, porém, foram pouco interessantes. Uma das possíveis razões para tal é que o método aplicado durante a fase de treinamento pode não ser compatível com aquele utilizado na fase de separação. Isto é, visto que a composição por exemplares não está relacionada de maneira alguma à NMF, a base por ela gerada pode ser completamente incompatível com a base esperada pelo algoritmo da NMF.

Outro aspecto das bases compostas por exemplares é o fato de que elas não contêm toda a informação disponibilizada durante a fase de treino. Como já antes comentado, a base composta por exemplares é um subconjunto do conjunto de treinamento. Portanto, para aumentar a quantidade de

informação dentro de uma base composta por exemplares, inevitavelmente, deve-se aumentar a quantidade de vetores-base nela. O aumento do tamanho das bases compostas por exemplares, porém, se mostrou ineficaz na tentativa de melhorar o desempenho da separação (vide Figura 9), o que pode indicar redundância de informação.

BASES TREINADAS PELA NMF

A primeira vantagem que as bases treinadas pela NMF têm sobre aquelas compostas por exemplares é a evidente compatibilidade com o método aplicado na fase de separação. Além disso, as bases treinadas pela NMF têm a propriedade de revelar características ocultas no conjunto de treinamento, agrupando toda a informação disponível em uma única base, independentemente da sua dimensão. Por conta dessa característica, as bases treinadas pela NMF são capazes de reunir mais informação (a partir de um conjunto de treino maior) utilizando uma mesma quantidade de vetores-base. Apesar disso, o aumento do conjunto de treino durante as experimentações não alterou o desempenho da separação, o que novamente leva a crer que há redundância de informações no conjunto de treinamento.

Calcularam-se os valores singulares gerados pela SVD tanto de uma base composta por exemplares como de uma treinada pela NMF. Seus valores estão representados em ordem crescente no eixo principal da Figura 12. No eixo secundário, plotou-se a porcentagem acumulada dos valores singulares para as duas bases. Percebe-se por esse gráfico que 80% da informação contida na base composta por exemplares poderiam ser representados por apenas 46 vetores singulares. A base treinada pela NMF, por outro lado, precisaria de 117 vetores singulares para que 80% de sua informação estivessem representados.

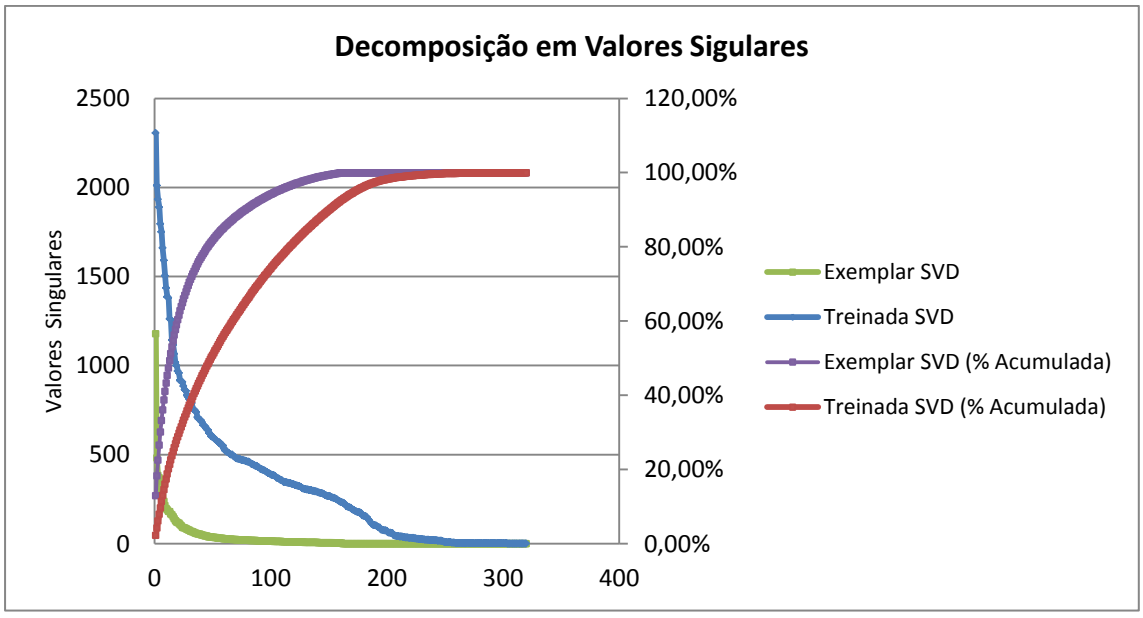

Figura 12: Valores singulares obtidos pela SVD para uma base composta por exemplares e para outra treinada pela NMF.

Dessa análise, extrai-se que as bases compostas por exemplares carregam menos informação que aquelas treinadas pela NMF e, além disso, são compostas com muita redundância. Em outras palavras, as bases treinadas pela NMF aproveitam com mais eficiência os vetores-bases que lhe estão disponíveis.

Outro aspecto interessante sobre as bases treinadas pela NMF refere-se à escolha da sua dimensão, isto é, à quantidade de vetores-base que a compõem. Apesar de a NMF ter surgido, a princípio, como uma maneira de se realizar redução dimensional, verificou-se que a separação de fontes foi mais bem sucedida com bases que, de fato, aumentavam a dimensão dos dados. Uma hipótese para explicar esse fenômeno é a de que a disponibilidade de mais vetores-base permita que a NMF não supervisionada, como técnica de revelação de estruturas ocultas, consiga individualizar melhor características dos dados. Tome-se como exemplo um conjunto de imagens de faces humanas. Lee e Seung, em [44], mostraram que a decomposição por matrizes não negativas produz uma base representativa de partes dos rostos, que se somam para formar um rosto completo. Isto é, cada um dos vetores-base abriga traços de olhos, nariz, sobrancelhas etc. Imagine-se, agora, que se permitam apenas três vetores na base. Nesse caso, o algoritmo teria que agrupar informações sobre os olhos e as sobrancelhas em único vetor, sobre a boca e o nariz em outro e sobre as orelhas e o cabelo em outro, por exemplo. Por outro lado, se houver disponíveis seis vetores para a base, então a NMF poderia dissociar os seis elementos e, assim, quando aplicada de maneira supervisionada, seria capaz de ponderar o vetor dos olhos sem afetar a aproximação das sobrancelhas.

COMPOSIÇÃO DAS BASES DE VOZ

Nos experimentos executados durante esse trabalho, treinou-se sempre a base de voz com uma mistura de locutores, de ambos os sexos e de diversas origens dialetais. Procurou-se com isso compor uma base de fala que generalizasse bem um sinal qualquer de voz e que fosse capaz de distinguir entre outros sinais quaisquer (música, nesse caso). É fato que as bases de fala foram capazes de explicar os sinais de voz, uma vez que os resultados foram pouco afetados por distorções e que continham, perceptualmente, toda a parcela da voz presente na mistura. Contudo, a base de fala também era ainda muito capaz de explicar parte dos sinais de música e, assim, a separação de fontes ficou longe de ser perfeita.

Não se pode, porém, atribuir toda a responsabilidade de bom desempenho da separação às bases de voz. Os sinais de música e de fala têm demasiadas semelhanças e, por isso, são tão difíceis de separar. Talvez uma representação diferente da STFT, que explore outras características espectrais dos sinais, obtenha resultados melhores.

Ressalta-se, ainda, a possibilidade de se especializar a base de voz para um determinado locutor ou, por exemplo, diferenciá-las para homens e mulheres. Uma mistura composta pela locução de um

homem, de voz tipicamente grave, e o som de uma flauta, tipicamente aguda, provavelmente poderia ser mais bem separada se a base de voz estiver modelada para identificar vozes masculinas.

COMPOSIÇÃO DAS BASES DE MÚSICA

Toma-se aqui a música como sinal de interesse e apresenta-se na Figura 13 o resultado de uma separação. Nesse caso, a relação sinal-ruído é entre música/voz e não mais voz/música. Chamamos essa relação MSR. Da mesma maneira, a SIR é calculada com a música no papel de sinal e a voz no papel de interferência.

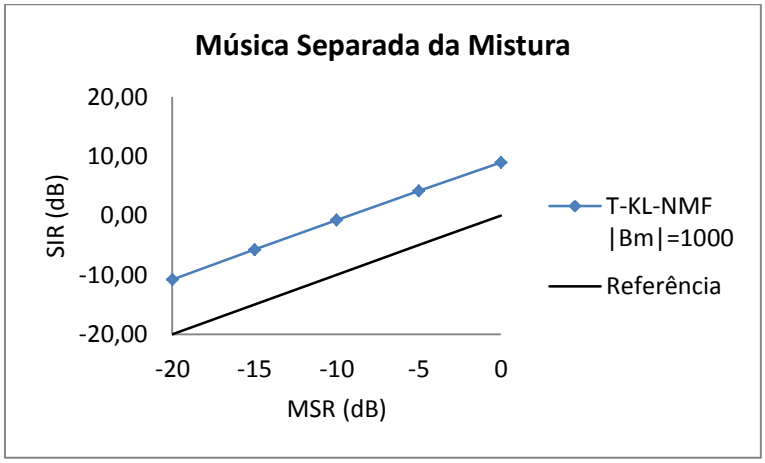

Figura 13: Bases treinadas pela KL-NMF e separação com o mesmo método. *Signal-to-Interferences Ratio* para o sinal de música em função da *Music-to-Speech Ratio.*

Vê-se no gráfico acima que a separação retira parte da voz que corrompe a música, mas que, ainda, a voz se mantém muito presente (até porque a proporção MSR das misturas privilegia o sinal de fala). Novamente, constata-se que a base de música é capaz de explicar o sinal para o qual foi desenvolvida, mas também explica o sinal de voz.

A música, com sua diversidade de sons, é um sinal extremamente complexo e, portanto, muito difícil de representar genérica e abrangentemente. Mais complicado ainda é realizar essa tarefa ao mesmo tempo em que se rejeitam os sinais de voz, em muitos aspectos semelhantes à música. Talvez seja necessária uma base de música muito maior e/ou com um conjunto de treinamento muito mais diverso.

Levanta-se também a questão de elaborar uma base de música que seja capaz de generalizar outros gêneros musicais. Trabalhou-se aqui exclusivamente com música Jazz não cantada. Eventualmente, seria interessante ser capaz de realizar a AuSS com qualquer tipo de música.

5.3. Algoritmos de Separação

Os algoritmos aqui experimentados, a KL-NMF e a IS-NMF, são bastante semelhantes e, de fato, não geraram resultados significativamente diferentes. Outras implementações da NMF poderiam acelerar o processamento das rotinas e também melhorar a aproximação $\boldsymbol{V} \approx \boldsymbol{BW}$, mas provavelmente não melhorariam tão significativamente os resultados apresentados na seção 4. Vale ressaltar que a divergência de Itakura-Saito foi desenvolvida como estimadora de máxima verossimilhança para espectros de potência e não de magnitude, como se aplica aqui. Assim, é possível que se tenha subutilizado a capacidade de IS-NMF. De fato, Févotte, em [7], obtém seus melhores resultados a partir da IS-NMF.

Outros autores, como Smaragdis e Virtanen, em [22] e [25] respectivamente, buscaram explorar características de *continuidade temporal* dos sinais acústicos. Virtanen, alterando apenas a função de custo utilizada pela NMF, mas mantendo o uso das regras multiplicativas de Lee e Seung, penalizou grandes diferenças entre os coeficientes da matriz de pesos de dois instantes adjacentes. Smaragdis, por outro lado, desenvolveu uma versão convolutiva da NMF, de forma que instantes anteriores e instantes futuros fossem levados em conta durante a decomposição de $\boldsymbol{V}$ no produto $\boldsymbol{BH}$.

Outra abordagem interessante foi a apresentada por Weninger *et al*, em [26], em que se propôs a NMF semi-supervisionada. Nesse método, a fase de treinamento é dedicada a formular apenas as bases de voz, enquanto a música é estimada e modelada simultaneamente durante a fase de separação. Os resultados apresentados em [26], com a NMF semi-supervisionada e baseada no método convolutivo de [22], são superiores aos melhores resultados da seção 4.3

Portanto, a ideia geral para continuar a evolução nos resultados de AuSS com NMF é explorar as principais características dos sinais acústicos e das transformadas utilizadas para representá-los, por meio de restrições direcionadas aos métodos de treinamento e de separação.

5.4. Desempenho de ASR

A forma dos gráficos das Figuras 9, 10 e 11 revelou que o desempenho do ASR não está somente relacionado à presença ou à ausência de música corrompendo a voz. Apesar de todos os testes realizados terem aumentado a SIR, nem todos elevaram o desempenho do ASR e, para os que melhoraram, não necessariamente o fizeram na mesma proporção. Ressaltam-se, ainda, casos em que a aplicação da NMF supervisionada deteriorou os resultados obtidos pelos ASR.

As observações acima indicam que os métodos que estão sendo desenvolvidos com objetivo de melhorar um sinal à entrada de um reconhecedor automático de fala devem também levar em conta a maneira como ASR foi projetado. Uma alternativa a isso seria re-treinar o ASR com sinais provenientes de separações com NMF e, assim, tê-lo mais bem preparado para esses casos.

Por fim, apesar de não ser o objetivo principal desse trabalho, chama-se a atenção para o tempo de processamento exigido pelas rotinas de NMF aqui utilizadas. Visto que a motivação apresentada em 1.1 envolvia a aplicação do ASR a um sistema de tempo real, verifica-se que a NMF, da maneira como foi implementada aqui, é um método demasiado lento.

6. Conclusões

Apresentou-se neste trabalho a aplicação da Fatoração em Matrizes Não Negativas (NMF) para realizar Separação de Fontes de Áudio (AuSS) com objetivo de elevar o desempenho de um Reconhecedor Automático de Fala (ASR). Os resultados obtidos foram bastante positivos em termos da relação voz-ruído e, em sua maioria, também em termos do desempenho do ASR.

Buscou-se otimizar o tamanho das bases de voz e de música, além do critério de parada da rotina NMF, em que se verificou grande relevância por parte do tamanho das bases. Compararam-se resultados oriundos de bases compostas por exemplares e de bases treinadas pela NMF, donde se conclui que as últimas são indubitavelmente superiores às primeiras. Por fim, compararam-se duas funções de custo utilizadas pela NMF, uma generalização da Divergência de Kullback-Leibler e a Divergência de Itakura-Saito. Nesse caso, não foi perceptível diferença entre os dois métodos.

Ressalta-se ainda, que o desempenho do ASR está intimamente vinculado à SIR, mas não depende exclusivamente disso. Assim, os métodos desenvolvidos com objetivo de aumentar o desempenho de um reconhecedor automático de fala devem levar em conta a maneira como este opera.

O seguimento desse projeto deve se orientar à experimentação de outras implementações da NMF; à exploração das características espectrais dos sinais de voz e de música por meio de estabelecimento de restrições à NMF (na forma de funções de custo e/ou de novos métodos); à busca de outras representações para os sinais envolvidos, que ressaltem as diferenças entre eles; ao direcionamento dos métodos para que se adaptem ao treinamento do ASR.

7. Referências

[1] M. Armstrong. (2011, Apr.). “Audio Processing and Speech Intelligibility: a literature review,” in *BBC Research White Paper WHP 190*. Available: *http://downloads.bbc.co.uk/rd/pubs/whp/whp-pdf-files/WHP190.pdf* [Sep. 27, 2013].

[2] S. Arora, R. Ge, R. Kannan and A. Moitra (2012). "Computing a Nonnegative Matrix Factorization – Provably". Available: *http://arxiv.org/abs/1111.0952* [Sep. 29, 2013].

[3] J. C. Brown and P. Smaragdis. “Independent component analysis for automatic note extraction from musical trills,” in *J. Acoust. Soc. Amer*., vol. 115, pp. 2295–2306, May 2004.

[4] J. Droppo and A. Acero, “Noise Robust Speech Recognition with a Switching Linear Dynamic Model,” in *Proc. ICASSP*, 2004.

[5] S. Dubnov. “Extracting sound objects by independent subspace analysis,” in *Proc. 22nd Int. Audio Eng. Soc. Conf.*, Espoo, Finland, un. 2002.

[6] Y. Ephraim and D. Malah. “Speech enhancement using a minimum mean square error short-time spectral amplitude estimator”, in *IEEE Transaction on Acoustic, Speech, Signal Processing*, vol. ASSP-32, pp. 1109-1121, Dec. 1984. Available : http://teal.gmu.edu/~yephraim/ephraim.html [Oct. 2013].

[7] C. Févotte, N. Bertin and J.L. Durrieu. “Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis.” In *Neural Computation*, vol. 21 (3), pp. 793–830, Mar. 2009.

[8] R. Gemulla, E. Nijkamp, P. J. Haas, Y. Sismanis (2011). "Large-scale matrix factorization with distributed stochastic gradient descent". Available: *http://www.mpi-inf.mpg.de/~rgemulla/publications/rj10481rev.pdf* [Sep. 29, 2013].

[9] J. R. Hershey, S. J. Rennie, O. P. A., and T. T. Kristjansson, “Super-human multi-talker speech recognition: A graphical modeling approach,” in *Computer Speech and Language*, 2010.

[10] J. E. Jackson. *A User’s Guide to Principal Components*. New York, NY: John Wiley & Sons, 2003.

[11] H. Kim and H. Park (2008). "Nonnegative Matrix Factorization Based on Alternating Nonnegativity Constrained Least Squares and Active Set Method". Available: *http://www.cc.gatech.edu/~hpark/papers/simax-nmf.pdf* [Sep. 29, 2013].

[12] J. Kim and H. Park (2011). "Fast Nonnegative Matrix Factorization: An Active-set-like Method and Comparisons". Available: *http://www.cc.gatech.edu/~hpark/papers/SISC_082117RR_Kim_Park.pdf* [Sep. 29, 2013].

[13] D. D. Lee and H. S. Seung. "Algorithms for Non-Negative Matrix Factorization," in Advances in *Neural Information Processing*, 2000.

[14] D. D. Lee and H. S. Seung. "Learning the parts of objects by non-negative matrix factorization," in *Nature*, vol. 401, pp. 788–791, Oct. 1999.

[15] D. D. Lee and H. S. Seung, "Unsupervised learning by convex and conic coding," in *Advances in Neural Information Processing Systems*, M. C. Mozer, M. I. Jordan, and T. Petsche, Eds., vol. 9. The MIT Press, 1997, pp. 515–521.

[16] C. J. Lin (Oct, 2007). "Projected Gradient Methods for Non-negative Matrix Factorization". Available: *http://www.csie.ntu.edu.tw/~cjlin/papers/pgradnmf.pdf* [Sep. 29, 2013].

[17] H. Meinedo, D. Caseiro, J. Neto, and I. Trancoso. "AUDIMUS.media: A broadcast news speech recognition system for the European Portuguese language," in *Computational Processing of the Portuguese Language: 6th International Workshop, PROPOR 2003 Proceedings*, N. J. Mamede, I. Trancoso, J. Baptista, and M. G. V. Nunes, Eds., pp. 9–17. Springer, Berlin Heidelberg, 2003, LNAI 2721.

[18] P. Paatero and U. Tapper, "Positive Matrix Factorization: A non-negative factor model with optimal utilization of error estimates of data values," in *Environmetrics*, vol. 5, pp. 111–126, 1994.

[19] B. Raj, R. Singh, and R. M. Stern, "On Tracking Noise with Linear Dynamical System Models," in *Proc. ICASSP*, 2004.

[20] B. Raj, T. Virtanen, S. Chaudhuri and R. Singh. "Non-negative matrix factorization based compensation of music for automatic speech recognition," presented at *Interspeech*, Makuhari, Japan, 2010.

[21] M. V. Shashanka, B. Raj, and P. Smaragdis, “Probabilistic Latent Variable Models as Non-Negative Factorizations,” in *Computational Intelligence and Neuroscience*, May 2008.

[22] P. Smaragdis. “Convolutive Speech Bases and their Application to Supervised Speech Separation,”in *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15 (1), pp; 1–12, Dec. 2006.

[23] A. P. Varga and R. K. Moore, “Hidden Markov Model decomposition of speech and noise,” in *Proc. ICASSP*, 1990.

[24] E. Vincent, R. Gribonval and C. Févotte. “Performance measurement in blind audio source separation,” in *IEEE Transactions on Speech and Audio Processing*, Jun. 2004.

[25] T. Virtanen. “Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria,” in *IEEE Transactions on Audio, Speech and Language Process*ing, vol. 15, pp. 1066–1074, Mar. 2007.

[26] F. Weninger, J. Feliu and B. Schuller. “Supervised and Semi-supervised Suppression of Background Music in Monaural Speech Recordings” in *ICASSP*, Mar. 2012.

Softwares e Bases de Dados

[27] J.S. Garofolo, et al. TIMIT Acoustic-Phonetic Continuous Speech Continuous. Linguistic Data Consortium, Philadelphia, 1993.

[28] E. Vincent, R. Gribonval and C. Févotte. *BSS_EVAL Toolbox*. Apr. 2005. Available: http://www.irisa.fr/metiss/bss_eval/ [Jul, 2013].

[29] NIST. *NIST SCLITE scoring package version 2.4.8.* Mar. 29, 2013. Available: http://www.itl.nist.gov/iad/mig/tools/ [Jul, 2013].