UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de Ilha Solteira

Thiago Garcia de Andrade

Melhoramento de voz baseado em representações esparsas usando dicionários treinados

Ilha Solteira
2020

UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de Ilha Solteira

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

Thiago Garcia de Andrade

Melhoramento de voz baseado em representações esparsas usando dicionários treinados

Texto submetido ao Programa de Pós Graduação em Engenharia Elétrica - UNESP - Campus de Ilha Solteira, Como requisito para obtenção do título de Mestre em Engenharia Elétrica.

Área de Conhecimento: Automação.

Orientador: Prof. Dr. Francisco Villarreal Alvarado

Ilha Solteira
2020

FICHA CATALOGRÁFICA
Desenvolvido pelo Serviço Técnico de Biblioteca e Documentação

A553m

Andrade, Thiago Garcia de.
Melhoramento de voz baseado em representações esparsas usando dicionários treinados / Thiago Garcia de Andrade. -- Ilha Solteira: [s.n.], 2020
55 f. : il.

Dissertação (mestrado) - Universidade Estadual Paulista. Faculdade de Engenharia de Ilha Solteira. Área de conhecimento: Automação, 2020

Orientador: Francisco Villarreal Alvarado
Inclui bibliografia

1. Melhoramento de voz. 2. Representações esparsas. 3. Treinamento de dicionários.

Raiane da Silva Santos
Supervisora Técnica de Seção
Seção Técnica de Referência, Atendimento ao usuário e Documentação
Diretoria Técnica de Biblioteca e Documentação
CRB/8 - 9999

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Ilha Solteira

CERTIFICADO DE APROVAÇÃO

TÍTULO DA DISSERTAÇÃO: Melhoramento de voz baseado em representações esparsas usando dicionários treinados

AUTOR: THIAGO GARCIA DE ANDRADE

ORIENTADOR: FRANCISCO VILLARREAL ALVARADO

Aprovado como parte das exigências para obtenção do Título de Mestre em ENGENHARIA ELÉTRICA, área: Automação pela Comissão Examinadora:

Prof. Dr. FRANCISCO VILLARREAL ALVARADO
Departamento de Matemática / Faculdade de Engenharia de Ilha Solteira - UNESP

Prof. Dr. JOZUE VIEIRA FILHO
Coordenadoria Executiva / Câmpus Experimental de São João da Boa Vista - UNESP

Prof. Dr. CAIO CESAR ENSIDE DE ABREU
Curso de Ciência da Computação / Universidade do Estado de Mato Grosso - UNEMAT

Ilha Solteira, 18 de agosto de 2020

Agradecimentos

Primeiramente, agradeço aos professores Francisco Villarreal Alvarado e Marco Aparecido Queiroz Duarte por me orientarem e ter desempenhado tal função com confiança, dedicação e amizade.

Aos meus colegas de curso, com quem convivi intensamente durante os últimos dois anos, pelo companheirismo e pela troca de experiências que me permitiram crescer não só como pessoa, mas também como pesquisador.

Aos meus pais e irmãos, que me incentivaram nos momentos difíceis e compreenderam a minha ausência enquanto eu me dedicava à realização deste trabalho.

Aos amigos, que sempre estiveram ao meu lado, pela amizade incondicional e pelo apoio demonstrado ao longo de todo o período em que me dediquei a este trabalho.

A todos que participaram, direta ou indiretamente do desenvolvimento deste trabalho de pesquisa, enriquecendo o meu processo de aprendizagem.

When dealing with people, remember you are not dealing with creatures of logic, but with creatures bristling with prejudice and motivated by pride and vanity.

Dale Carnegie

Resumo

Melhorar sinais de voz degradados por ruídos não-estacionários é uma tarefa importante e de interesse em diversas áreas de pesquisa. Os espectros variantes no tempo de ruídos não-estacionários comprometem o desempenho de métodos clássicos de melhoramento de voz. Este trabalho explora a utilização de representações esparsas utilizando dicionários treinados no melhoramento de voz. O sinal ruidoso no domínio tempo-frequência é codificado de maneira esparsa utilizando um dicionário formado pela concatenação de um dicionário de voz e um dicionário de ruído. A voz pura é estimada pela representação gerada pelo dicionário de voz enquanto a estimação do ruído é dada pela representação fornecida pelo dicionário de ruído. Uma codificação muito esparsa aumenta o erro de aproximação, denotado por distorção de fonte. Uma codificação muito densa causa confusão de fonte, onde a voz é parcialmente representada pelo dicionário de ruído, e o ruído é parcialmente codificado pelo dicionário de voz. A esparsidade da representação é regulada para melhorar o desempenho. Os resultados experimentais mostram que esta abordagem alcança resultados superiores à subtração espectral, filtro de Wiener e MMSE-STSA usando diferentes medidas objetivas de avaliação.

Palavras-chave: Melhoramento de voz. Representações esparsas. Treinamento de dicionários.

Abstract

Enhancing speech degraded by non-stationary noises is an important task and of great interest in several research areas. The time-varying spectra of non-stationary noises compromise the performance of classical speech enhancement methods. This work explores the use of sparse representations using trained dictionaries in speech enhancement. The mixture in time-frequency domain is sparsely encoded in a dictionary formed by the concatenation of a speech dictionary and a noise dictionary. The cleaned speech is estimated by the representation generated by the speech dictionary while the noise estimation is given by the representation provided by the noise dictionary. Very sparse coding increases the approximation error, denoted by source distortion. Very dense encoding causes source confusion, where the voice is partially represented by the noise dictionary, and the noise is partially encoded by the voice dictionary. The sparsity of the representation is regulated to improve performance. Experimental results shows that this approach achieves results superior to spectral subtraction, Wiener filter and MMSE-STSA using different objective evaluation measures.

Keywords: Speech enhancement. Sparse representations. Dictionary learning.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AK-SVD	*Approximate K-Singular Value Decomposition*
DL	*Dictionary Learning*
DTFT	*Discrete-Time Fourier Transform*
FIR	*Finite Impulse Response*
fwSSNR	SNR segmentada ponderada na frequência
GDL	*Generative Dictionary Learning*
IIR	*Infinite Impulse Response*
K-SVD	*K-Singular Value Decomposition*
LARC	*Batch Least-Angle Regression with Coherence Criterion*
LARS	*Least-Angle Regression*
LASSO	*Least Absolute Shrinkage and Selection Operator*
MMSE-STSA	*Minimum Mean-Square Error Short-Time Spectral Amplitude*
MP	*Matching Pursuit*
ms	milisegundos
OMP	*Orthogonal Matching Pursuit*
PESQ	*Perceptual Evaluation of Speech Quality*
SNR	*Signal-to-Noise Ratio* (Relação Sinal Ruído)
SSNR	SNR segmentada
STFT	*Short-Time Fourier Transform*
STOI	*Short-Time Objective Intelligibility*
VAD	*Voice Activity Detection*

Lista de símbolos

$X(\omega, n)$	Transformada de Fourier de tempo curto de $\boldsymbol{x}(n)$
$\lvert X(\omega, n)\rvert$	Magnitude da transformada de Fourer de tempo curto de $\boldsymbol{x}(n)$
$e^{j\phi_x(\omega,n)}$	Fase da transformada de Fourer de tempo curto de $\boldsymbol{x}(n)$
$X^*(\omega, n)$	Conjugado complexo da transformada de Fourier de tempo curto de $\boldsymbol{x}(n)$
$X(\omega)$	Transformada de Fourier de tempo discreto de $\boldsymbol{x}(n)$
$E\{\boldsymbol{x}(n)\}$	Esperança matemática de $\boldsymbol{x}(n)$
$\boldsymbol{r}_{xy}$	Vetor de autocorrelação cruzada entre $\boldsymbol{x}(n)$ e $\boldsymbol{y}(n)$
$\boldsymbol{R}_{xx}$	Matriz de autocorrelação cruzada de $\boldsymbol{x}(n)$
$\rho(\boldsymbol{X}, \boldsymbol{Y})$	Função densidade de probabilidade conjunta das matrizes $\boldsymbol{X}$ e $\boldsymbol{Y}$
$\rho(\boldsymbol{X} \mid \boldsymbol{Y})$	Função densidade de probabilidade condicional das matrizes $\boldsymbol{X}$ e $\boldsymbol{Y}$
$E_{\boldsymbol{X}\mid\boldsymbol{Y}}\{\boldsymbol{X} \mid \boldsymbol{Y}\}$	esperado entre as matrizes $\boldsymbol{X}$ e $\boldsymbol{Y}$
$\lVert\boldsymbol{x}\rVert_0$	Pseudo-norma ℓ_0 do vetor $\boldsymbol{x}$
$\lVert\boldsymbol{x}\rVert_1$	Norma ℓ_1 do vetor $\boldsymbol{x}$
$\lVert\boldsymbol{x}\rVert_2$	Norma ℓ_2 do vetor $\boldsymbol{x}$
$\lVert\boldsymbol{X}\rVert_F$	Norma Frobenius da matriz $\boldsymbol{X}$
$\langle\boldsymbol{x}, \boldsymbol{y}\rangle$	Produto interno entre os vetores $\boldsymbol{x}$ e $\boldsymbol{y}$
$\lvert\boldsymbol{x}\rvert$	Módulo do vetor $\boldsymbol{x}$
$sup\lvert\boldsymbol{x}\rvert$	Valor máximo do vetor $\boldsymbol{x}$
$argmin_{\boldsymbol{y}}\boldsymbol{x}$	Função que retorna o valor mínimo de $\boldsymbol{x}$ em relação ao argumento $\boldsymbol{y}$
$(\boldsymbol{X})^p$	Operador que eleva a potência p todos os elementos da matriz $\boldsymbol{X}$
$\boldsymbol{X} \otimes \boldsymbol{Y}$	Operador que multiplica a matriz $\boldsymbol{X}$ elemento por elemento com a matriz $\boldsymbol{Y}$
$\mu(\boldsymbol{X})$	Coerência mútua do dicionário $\boldsymbol{X}$
$\mu(\boldsymbol{X}, \boldsymbol{Y})$	Coerência mútua entre os dicionários $\boldsymbol{X}$ e $\boldsymbol{Y}$

Sumário

1 Introdução

O melhoramento de voz é um campo de pesquisa que tem gerado grande interesse da comunidade científica nas últimas décadas por sua importância em áreas como reconhecimento de voz, aparelhos auditivos e a comunicação de dispositivos portáteis (LOIZOU, 2013). Melhorar a voz significa melhorar sua qualidade e sua inteligibilidade atenuando o ruído e evitando a sua degradação de forma substancial. Um sinal de voz com alta qualidade pode ser ouvido de forma confortável por longos períodos de tempo. Enquanto que a alta inteligibilidade permite maior compreensão das palavras.

Métodos clássicos de melhoramento de voz são categorizados por: métodos de subtração espectral (BOLL, 1979; LU; LOIZOU, 2008), métodos estatísticos (EPHRAIM; MALAH, 1984; EPHRAIM; MALAH, 1985; EPHRAIM, 1992) e métodos baseados em subespaços vetoriais (HU; LOIZOU, 2003; JABLOUN; CHAMPAGNE, 2003). Tais métodos mostram-se ineficazes em atenuar ruídos não-estacionários, pois seus espectros variantes no tempo inviabilizam estimativas suficientemente adequadas. Recentemente, abordagens de melhoramento de voz baseadas em representações esparsas têm sido apresentadas com o objetivo de alcançar resultados mais satisfatórios em ambientes não-estacionários (SIGG; DIKK; BUHMANN, 2010; SIGG; DIKK; BUHMANN, 2012).

Em processamento de sinais, representar um sinal $\boldsymbol{x} \in \mathbb{R}^N$ significa expressá-lo como a combinação linear de elementos do dicionário $\boldsymbol{D} = [\boldsymbol{d}_1, ..., \boldsymbol{d}_K]$, com $\boldsymbol{D} \in \mathbb{R}^{N \times K}$ e cada $\boldsymbol{d}_i \in \mathbb{R}^N$, ou seja, $x = \sum_{i=1}^{K} c_i \boldsymbol{d}_i$, onde $\boldsymbol{c}$ é a representação de $\boldsymbol{x}$ em $\boldsymbol{D}$. Se $\boldsymbol{x}$ é representado de forma exata a partir de um pequeno número de elementos de $\boldsymbol{D}$, essa representação é dita esparsa. Em geral, opta-se por dicionários sobrecompletos com $K > N$, e o problema de encontrar a solução mais esparsa resume-se a solucionar o sistema de equações lineares indeterminado $x = \boldsymbol{D}\boldsymbol{c}$. Mallat e Zhang (1993) apresentaram o algoritmo de busca *Matching Pursuit* (MP), que decompõe sinais em uma combinação linear de átomos selecionando os átomos que melhor se ajustam a estrutura do sinal, e controlando a esparsidade de $\boldsymbol{c}$ pela sua pseudo-norma ℓ_0. Em seguida, algoritmos como o *Least Absolute Shrinkage and Selection Operator* (LASSO) foram propostos e, diferente de métodos como o MP, controlam a esparsidade de $\boldsymbol{c}$ minimizando sua norma ℓ_1, e, ao contrário do MP, possuem solução analítica em tempo polinomial (TIBSHIRANI, 1996). Aplicações de métodos de otimização com norma ℓ_1 têm sido amplamente utilizadas em áreas como *machine learning* e reconhecimento de padrões (PATEL; CHELLAPPA, 2011; LIU *et al.*, 2014; YUAN; LU; LI, 2014).

Buscar representações adequadas para uma determinada tarefa envolve a escolha de um dicionário que possua átomos coerentes com as diferentes estruturas presentes em uma determinada classe de sinais. Os primeiros dicionários foram criados a partir de formulações

analíticas e eram projetados para se adequarem a classes de sinais específicas. Dicionários de Fourier, *Wavelet* e Gabor são exemplos de dicionários analíticos. Contudo, o número fixo e limitado de átomos com estrutura simples mostra-se insuficientes para se representar fenômenos complexos. No final dos anos 1990, algoritmos de *dictionary learning* surgiram como alternativa aos dicionários analíticos. Tais métodos treinam dicionários a partir de amostras de sinais de uma determinada classe, gerando átomos que se ajustam melhor a estes sinais que átomos pré-definidos (RUBINSTEIN; BRUCKSTEIN; ELAD, 2010).

1.1 Contribuições do trabalho

Este trabalho tem como objetivo explorar uma nova abordagem de melhoramento de voz, conhecida como *Generative Dictionary Learning* (GDL), baseada na representação esparsa da voz degradada usando dicionários treinados e compará-la com métodos clássicos, expondo seu potencial na filtragem de voz em ambientes não estacionários de diferentes naturezas.

Diferente de métodos como a subtração espectral, o GDL não assume que o ruído é invariante no tempo e ao contrário de métodos baseados em dicionários analíticos, os dicionários treinados utilizados no GDL permitem uma melhor separação da voz degradada por ruídos com estruturas semelhantes a da voz, mesmo em condições com baixo SNR (SIGG; DIKK; BUHMANN, 2012).

O GDL utiliza o *Batch LARS with Coherence Criterion* (LARC) como algoritmo de codificação esparsa tanto na etapa de treinamento quanto na etapa de melhoramento de voz, em vez de algoritmos como o *Orthogonal Matching Pursuit* (OMP) e o LASSO, amplamente utilizados para esta tarefa. Desenvolvido como uma extensão do LARS (do inglês *Least Angle Regression*), o LARC utiliza um limiar de coerência em vez do erro residual ou cardinalidade como critério de parada. O LARC foi formulado de forma a tornar mais eficiente a codificação de um grande número de amostras por um mesmo dicionário. Além disso, o limiar de coerência permite que um mesmo critério de parada seja utilizado para sinais com diferentes distribuições de energia.

1.2 Organização do trabalho

O restante do trabalho é organizado da seguinte forma:

- O capitulo 2 apresenta os conceitos matemáticos básicos de métodos clássicos de melhoramento de voz;
- O capítulo 3 introduz a teoria das representações esparsas, os algoritmos existentes para encontrar tais representações e uma de suas principais aplicações: O treinamento de dicionários sobrecompletos;

- O capítulo 4 apresenta o GDL, explorando a utilização de dicionários treinados na filtragem de sinais de voz degradados por ruídos não-estacionários. Ao final do capítulo, O desempenho do GDL é comparado com métodos clássicos a partir dos resultados gerados;
- Ao final do trabalho, são feitas as considerações finais e possíveis futuros trabalhos são apresentados.

2 Métodos Clássicos de Melhoramento de Voz

Para sinais de voz captados por um único microfone, o sinal ruidoso $\boldsymbol{x}(n)$ é modelado como a soma do sinal de voz limpo e do ruído aditivo, isto é,

$$\boldsymbol{x}(n) = \boldsymbol{s}(n) + \boldsymbol{i}(n) \tag{1}$$

onde $\boldsymbol{s}(n)$ e $\boldsymbol{i}(n)$ são o sinal de voz limpo e o ruído no domínio do tempo, respectivamente. Melhorar a voz a partir da Equação (1) significa estimar um sinal $\hat{\boldsymbol{s}}(n)$, que de alguma forma se aproxime de $\boldsymbol{s}(n)$ usando apenas $\boldsymbol{x}(n)$.

2.1 Subtração Espectral

A subtração espectral, ilustrada na Figura 1, é cronologicamente um dos primeiros algoritmos de melhoramento de voz (BOLL, 1979). Sejam $X(\omega, n)$, $S(\omega, n)$ e $I(\omega, n)$ a Transformada de Fourier de Curto Prazo (STFT, do inglês *Short-Time Fourier transform*) de $\boldsymbol{x}(n)$, $\boldsymbol{s}(n)$ e $\boldsymbol{i}(n)$, respectivamente. Tomando a STFT de ambos os lados da Equação (1), $X(\omega, n)$ pode ser expressa como

$$X(\omega, n) = S(\omega, n) + I(\omega, n) \tag{2}$$

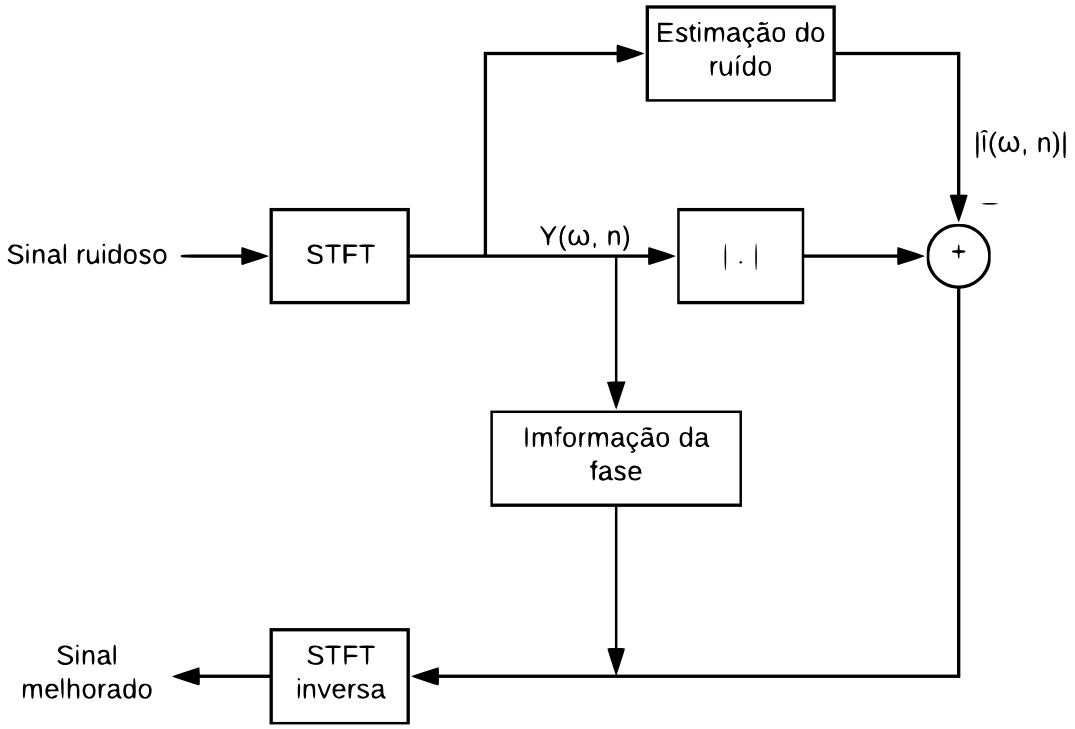
onde n é o índice que indica a posição da janela no domínio do tempo e ω é o índice de frequência, respectivamente. $X(\omega, n)$, $S(\omega, n)$ e $I(\omega, n)$ são sinais com valores complexos. Como a única informação disponível é o sinal $\boldsymbol{x}(n)$, tanto a magnitude $|I(\omega, n)|$ quanto a fase $e^{j\phi_i(\omega,n)}$ do ruído são desconhecidas. $|I(\omega, n)|$ pode ser estimada a partir dos trechos de pausa de $\boldsymbol{x}(n)$, enquanto que $e^{j\phi_i(\omega,n)}$ pode ser substituído pela fase do sinal ruidoso $e^{j\phi_x(\omega,n)}$. O método de subtração espectral mais simples estima o espectro de magnitude do sinal de voz puro subtraindo-se o espectro de magnitude do ruído do espectro de magnitude do sinal ruidoso:

$$\hat{S}(\omega, n) = (|X(\omega, n)| - |\hat{I}(\omega, n)|)e^{j\phi_x(\omega,n)} \tag{3}$$

onde $|\hat{S}(\omega, n)|$ é a estimação do espectro de magnitude da voz pura e $|\hat{I}(\omega, n)|$ é a estimação do espectro de magnitude do ruído fornecida pelos trechos de pausa. O sinal de voz melhorado é dado pela transformada de Fourier inversa de $\hat{S}(\omega, n)$.

O espectro de magnitude do sinal de voz melhorado $|\hat{S}(\omega, n)|$ pode assumir valores negativos caso o espectro de magnitude do ruído seja estimado de forma imprecisa. Contudo, $|\hat{S}(\omega, n)|$ não

Figura 1 – Ilustração do algoritmo de subtração espectral.



Fonte: Elaborado pelo próprio autor.

pode assumir valores negativos. Uma alternativa a este problema é zerar os valores negativos de $|X(\omega,n)| - |\hat{S}(\omega,n)|$ como na equação (4).

$$|\hat{S}(\omega,n)| = \begin{cases} |X(\omega,n)| - |\hat{I}(\omega,n)| & \text{se} \quad |X(\omega,n)| > |\hat{I}(\omega,n)| \\ 0 & \text{caso contrário} \end{cases} \tag{4}$$

Esta é apenas uma maneira de garantir a não-negatividade de $|\hat{S}(\omega,n)|$.

A Equação (3) pode ser reescrita da seguinte maneira

$$\hat{S}(\omega,n) = H(\omega,n)|X(\omega,n)|e^{j\phi_x(\omega,n)} \tag{5}$$

onde

$$H(\omega,n) = 1 - \frac{|\hat{I}(\omega,n)|}{|X(\omega,n)|}. \tag{6}$$

Na teoria dos sistemas lineares, $H(\omega,n)$ é a função de transferência do sistema, enquanto que em melhoramento de voz é identificada como a função de ganho, também conhecida como função de supressão. $H(\omega,n)$ assume apenas valores positivos no intervalo $0 \leq H(\omega,n) \leq 1$, suprimindo os valores de $|X(\omega,n)|$ para então obter o espectro de voz melhorado $\hat{S}(\omega,n)$.

A subtração espectral também pode ser realizada usando os espectros de potência levando à uma relação semelhante à Equação (2):

$$|X(\omega,n)|^2 = |S(\omega,n)|^2 + |I(\omega,n)|^2 + S(\omega,n)I^*(\omega,n) + S^*(\omega,n)I(\omega,n). \tag{7}$$

$S^*(\omega,n)$ e $I^*(\omega,n)$ são os conjungados complexos de $S(\omega,n)$ e $I(\omega,n)$, respectivamente. O espectro de potência $|I(\omega,n)|^2$ é desconhecido, e pode ser aproximado pela esperança $E[|I(\omega,n)|^2]$ a partir de trechos de pausa e é denotado por $|\hat{I}(\omega,n)|^2$. Assumindo que $\boldsymbol{i}(n)$ tem média nula e não tem correlação com $\boldsymbol{s}(n)$, os termos cruzados $S(\omega,n)I^*(\omega,n)$ e $S^*(\omega,n)I(\omega,n)$ podem ser anulados. Assim, a estimação de $|S(\omega,n)|^2$ é dada por

$$|\hat{S}(\omega,n)|^2 = |X(\omega,n)|^2 - |\hat{I}(\omega,n)|^2. \tag{8}$$

A sinal de voz melhorado é obtido pela transformada de Fourier inversa da raiz quadrada de $|\hat{S}(\omega,n)|^2$, usando a fase de $X(\omega,n)$ para estimar a fase de $\hat{S}(\omega,n)$.

Caso a estimação do espectro do ruído não seja suficientemente precisa, a subtração pode gerar picos no espectro do sinal de voz melhorado. Estes artefatos gerados pela subtração espectral são conhecidos como ruído musical. Além disso, estimações de ruídos realizadas com ferramentas como VAD (do inglês, *Voice Activity Detection.*) não são uma tarefa simples, dada a natureza não estacionária dos ruídos encontrados na vida real (LOIZOU, 2013).

2.2 Métodos Estatísticos Ótimos

Abordagens envolvendo subtração espectral são baseadas em ideias intuitivas e heurísticas. Dessa forma, o espectro do sinal melhorado não é estimado de maneira ótima (LOIZOU, 2013; HENDRIKS; GERKMANN; JENSEN, 2013). Algoritmos de melhoramento de voz ótimos propõem que o problema de melhorar o sinal de voz seja abordado como um problema de estimação estatístico com critério de otimalidade e suposições estatísticas definidas de maneira restrita.

2.2.1 Filtros de Wiener Básicos

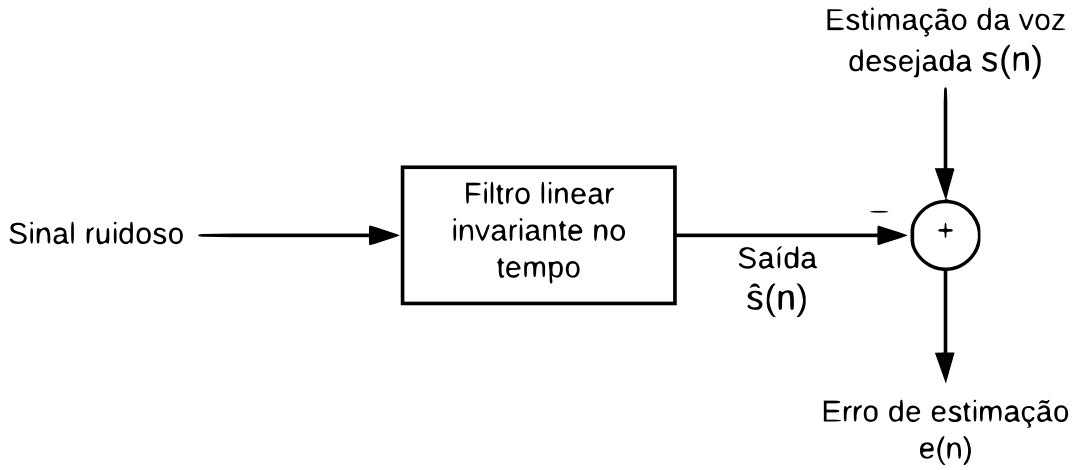
Filtros de Wiener são filtros ótimos de minimização de erro quadrático médio e são formulados a partir do modelo aditivo da Equação (1), assumindo que $\boldsymbol{s}(n)$ e $\boldsymbol{i}(n)$ são processos estacionários não correlacionados (HENDRIKS; GERKMANN; JENSEN, 2013). Filtros de Wiener são filtros lineares e invariantes no tempo, podendo ter comprimento finito (FIR, do inglês

Finite Impulse Response), ou infinito (IIR, do inglês *Infinite Impulse Response*). Para o filtro de Wiener do tipo FIR, a estimação da voz melhorada é dada por

$$\hat{\boldsymbol{s}}(n) = \sum_{k=0}^{M+1} h_k x(n-k) \tag{9}$$

onde h_k é o k-ésimo elemento do filtro FIR. A Figura 2 ilustra o funcionamento do método.

Figura 2 – Diagrama de bloco do projeto do filtro estatístico ótimo.



Fonte: Elaborado pelo próprio autor.

Os coeficientes do filtro devem ser projetados para minimizar o erro estimado

$$\begin{aligned} \boldsymbol{e}(n) &= \boldsymbol{s}(n) - \hat{\boldsymbol{s}}(n) \\ &= \boldsymbol{s}(n) - \boldsymbol{h}^T(n)\underline{\boldsymbol{x}}(n) \end{aligned} \tag{10}$$

sendo $\boldsymbol{h}^T(n) = [h_0, h_1, ..., h_M]$ os coeficientes do filtro ótimo e $\underline{\boldsymbol{x}}(n) = [x(n), x(n-1), ..., x(n-M+1)]^T$ é o vetor contendo os últimos M elementos de $\boldsymbol{x}(n)$. O erro médio quadrático médio é dado por

$$\begin{aligned} J = E\{e^2(n)\} &= E\{\boldsymbol{s}(n) - \boldsymbol{h}^T(n)\underline{\boldsymbol{x}}(n)\}^2 \\ &= E\{\boldsymbol{h}^2(n)\} - 2\boldsymbol{h}^T(n)E\{\boldsymbol{h}^T(n)\underline{\boldsymbol{x}}(n)\} + \boldsymbol{h}^T(n)E\{\underline{\boldsymbol{x}}(n)\underline{\boldsymbol{x}}(n)^T\}\boldsymbol{h}(n) \\ &= E\{\boldsymbol{h}^2(n)\} - 2\boldsymbol{h}^T(n)\boldsymbol{r}_{xh} + \boldsymbol{h}^T(n)\boldsymbol{R}_{xx}\boldsymbol{h}(n) \end{aligned} \tag{11}$$

onde $\boldsymbol{r}_{xh}$ é o vetor de autocorrelação cruzada entre o sinal ruidoso $\boldsymbol{x}(n)$ e o filtro $\boldsymbol{h}(n)$ e $\boldsymbol{R}_{xx}$ é a matriz de autocorrelação cruzada de $\boldsymbol{x}(n)$.

Minimizar a função de custo J equivale a zerar o vetor gradiente

$$\frac{\partial J}{\partial h_k} = 2E\left\{ \boldsymbol{e}(n)\frac{\partial \boldsymbol{e}(n)}{\partial h_k} \right\} = 0, k = 0, 1, ..., M-1. \tag{12}$$

A partir da Equação (10), a Equação (12) torna-se

$$\frac{\partial J}{\partial h_k} = -2E\{\boldsymbol{e}(n)\boldsymbol{x}(n-k)\} = 0, k = 0, 1, ..., M-1. \tag{13}$$

A Equação (13) precisa atender ao princípio de ortogonalidade para minimizar J, ou seja, o erro $\boldsymbol{e}(n)$ precisa ser ortogonal a $\boldsymbol{x}(n)$. O anulamento da derivada de J em relação ao filtro $\boldsymbol{e}(n)$ leva a seguinte condição de minimização

$$\frac{\partial J}{\partial \boldsymbol{h}(n)} = -2\boldsymbol{r}_{xh} + \boldsymbol{h}^T(n)\boldsymbol{R}_{xx} = 0. \tag{14}$$

Determinar $\boldsymbol{h}(n)$ pela Equação (14) leva ao filtro ótimo $\boldsymbol{h}^*(n)$:

$$\boldsymbol{R}_{xx}\boldsymbol{h}^*(n) = \boldsymbol{r}_{xh}$$

$$\boldsymbol{h}^*(n) = \boldsymbol{R}_{xx}^{-1}\boldsymbol{r}_{xh}. \tag{15}$$

A Equação acima é conhecida como solução de Wiener-Hopf (LOIZOU, 2013).

Um filtro de Wiener alternativo pode ser projetado no domínio da frequência (LIM; OPPENHEIM, 1979). Permitindo que o filtro $\boldsymbol{h}(n)$ seja projetado como um filtro IIR, sua determinação no domínio da Transformada de Fourier de Tempo Discreto (DTFT, do inglês *Discrete-Time Fourier Transform*) é obtida a partir da Equação (16).

$$\hat{S}(\omega) = H(\omega)X(\omega) \tag{16}$$

onde $\hat{S}(\omega)$, $H(\omega)$ e $X(\omega)$ são as transformadas de Fourier de tempo discreto do sinal de voz melhorado $\hat{\boldsymbol{s}}(n)$, do filtro IIR $\boldsymbol{h}(n)$ e do sinal ruidoso $\boldsymbol{x}(n)$, respectivamente. O filtro de Wiener no domínio da frequência é dado por

$$H(\omega) = \frac{|S(\omega)|^2}{|I(\omega)|^2 + |S(\omega)|^2}. \tag{17}$$

O filtro $H(\omega)$ é real, não-negativo e par, o que o torna um filtro de fase linear, ou seja, $H(\omega)$ não modifica a fase de $X(\omega)$, assim $\hat{S}(\omega)$ terá a mesma fase de $X(\omega)$, se assemelhando aos métodos de subtração espectral.

Para voz e ruído aditivos estatisticamente independentes a minimização de $E\{\hat{S}(\omega) - S(\omega)\}$ por $H(\omega)$ é definida como

$$H(\omega) = \frac{E\{|S(\omega)|^2\}}{E\{|I(\omega)|^2\} + E\{|S(\omega)|^2\}} = \frac{\psi}{1+\psi}. \tag{18}$$

A razão entre os espectros de potência $E\{|S(\omega)|^2\}$ e $E\{|I(\omega)|^2\}$ é denominada relação sinal ruído (SNR, do inglês *Signal to Noise Ration*) *a priori* como segue abaixo

$$\psi = \frac{E\{|S(\omega)|^2\}}{E\{|I(\omega)|^2\}}. \tag{19}$$

A aplicação direta do filtro de Wiener depende dos espectros de potência de $\boldsymbol{x}(n)$ e $\boldsymbol{s}(n)$. Além disso, o filtro não é causal, logo não é realizável.

Ephraim e Malah (1984) apresentaram um método de decisão direta para estimar ψ e assim projetar o filtro de Wiener na Equação (18). A SNR *a priori* é estimada como a combinação ponderada de estimações passadas e presentes de ψ. Para o segmento de voz referente ao índice de tempo n_k, a SNR *a priori* ψ_{n_k} é estimada como

$$\hat{\psi}_{n_k} = \alpha \frac{|\hat{S}(\omega, n_k - 1)|^2}{E\{|I(\omega, n_k - 1)|^2\}} + (1 - \alpha)\max\left(\frac{|X(\omega, n_k)|^2}{E\{|I(\omega, n_k)|^2\}}, 0\right) \tag{20}$$

onde $|\hat{S}(\omega, n_k - 1)|^2$ é o espectro de potência melhorado no segmento $n-1$ e α é um coeficiente de suavização, com $\alpha = 0.98$. O quociente à direita é a SNR *a posteriori* definida como

$$\gamma_{n_k} = \frac{|X(\omega, n_k)|^2}{E\{|I(\omega, n_k)|^2\}}. \tag{21}$$

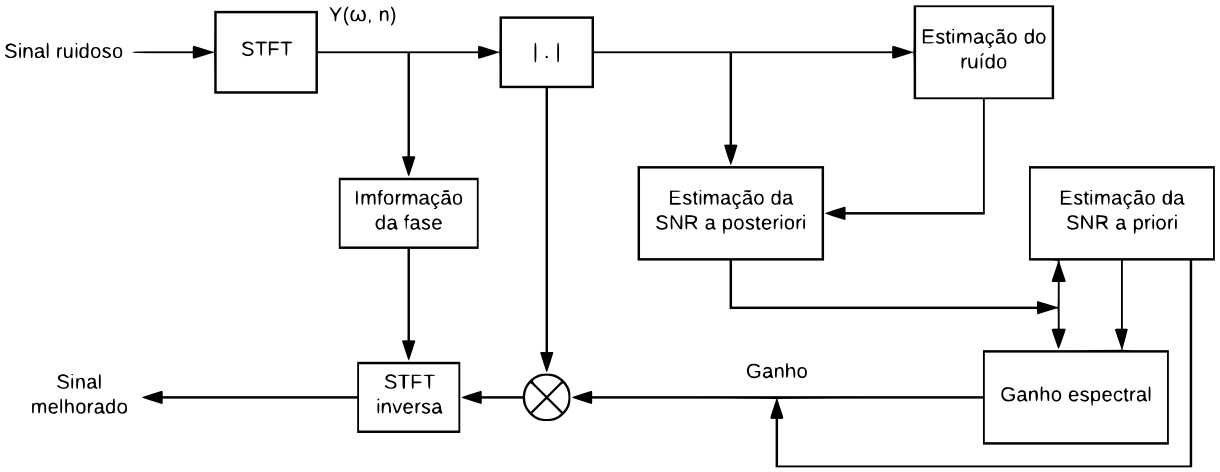
Quando α assume valores próximos de 1, $\hat{\psi}_{n_k}$ na Equação (20) é uma combinação ponderada da estimação passada e o segmento atual a ser melhorado. Esta relação suaviza a estimativa de ψ_{n_k} e pode eliminar o ruído musical (CAPPÉ, 1994).

2.2.2 Estimadores MMSE de Espectro de Magnitude

Os filtros de Wiener são filtros ótimos para a estimação do espectro complexo da voz. Porém, se a fase do espectro for desconsiderada (GERKMANN; KRAWCZYK; REHR, 2012), as abordagens de melhoramentos de voz são conduzidas aos estimadores ótimos de espectro de magnitude, também conhecidos como estimadores MMSE de espectro de magnitude (MMSE-STSA, do inglês *non-linear Minimum Mean Squared Error Short-Time Spectral Amplitude)*) (EPHRAIM; MALAH, 1984).

Diferentes dos filtros de Wiener que consideram uma relação linear entre o sinal ruidoso e o sinal melhorado, os estimadores MMSE-STSA utilizam um modelo estatístico Bayesiano, onde são feitas suposições explícitas sobre a distribuição de probabilidade dos coeficientes da DTFT da voz e do ruído. Na Figura 3 é apresentado o diagrama de blocos desse método. Para estimar a voz a partir de um sinal ruidoso, este método aplica um ganho sobre o seu espectro de magnitude. O filtro de ganho é calculado ao minimizar o erro quadrático médio entre o espectro de magnitude da voz pura e da voz ruidosa. O erro do espectro de voz pode ser avaliado utilizando-se a SNR *a priori* e a SNR *a posteriori*.

Figura 3 – Diagrama de blocos do métodos MMSE-STSA.



Fonte: Elaborado pelo próprio autor.

Sejam $X(\omega, n)$, $S(\omega, n)$, $\hat{S}(\omega, n)$ e $I(\omega, n)$ as variáveis aleatórias dadas pelas STFT do sinal ruidoso $\boldsymbol{x}(n)$, da voz pura $\boldsymbol{s}(n)$, do sinal de voz melhorado $\hat{\boldsymbol{s}}(n)$ e do ruído $\boldsymbol{i}(n)$, respectivamente. S_{ω_k} e X_{ω_k} representam o k-ésimo componente espectral do espectro de magnitude do segmento de índice temporal n_k de $S(\omega, n)$ e de $X(\omega, n)$, respectivamente. O índice de tempo será ocultado por simplicidade de notação. O erro quadrático médio Bayesiano entre S_{ω_k} e $\hat{S}_{\omega_k}$ é definido como

$$\mathcal{J}_{MSE} = E\left\{(S_{\omega_k} - \hat{S}_{\omega_k})^2\right\}. \tag{22}$$

O estimador MMSE-STSA ótimo que minimiza o erro quadrático médio bayesiano com respeito a $\hat{S}_{\omega_k}$ é dado por

$$\hat{S}_{\omega_k} = E\left\{S_{\omega_k} | X(\omega_k)\right\} \tag{23}$$

que é o valor esperado de S_{ω_k} dado $X(\omega_k)$.

Um grande número de estimadores pode ser projetado a partir da Equação (23). Ephraim e Malah (1984) assumiram que $S(\omega, n)$ e $I(\omega, n)$ são variáveis aleatórias independentes com distribuições gaussianas de média zero. A partir destas suposições, o estimador é obtido como mostrado na Equação (24).

$$\hat{S}_{\omega_k} = \frac{\int_0^{+\infty}\int_0^{2\pi} s_k \rho(X(\omega_k)|s_k, \phi_s)\rho(s_k, \phi_s) d\phi_s ds_k}{\int_0^{+\infty}\int_0^{2\pi} \rho(X(\omega_k)|s_k, \phi_s)\rho(s_k, \phi_s) d\phi_s ds_k} \tag{24}$$

onde s_k é a realização da variável aleatória S_{ω_k} e ϕ_s é a realização da fase da variável aleatória de $S(\omega_k)$. A partir das suposições feitas para as variáveis aleatórias, $\rho(X(\omega_k)|s_k, \phi_s)$ e $\rho(s_k, \phi_s)$ são dadas por

$$\rho(X(\omega_k)|s_k, \phi_s) = \frac{1}{\pi\lambda_i(k)} \exp\left\{-\frac{1}{\lambda_i(k)}|X(\omega_k) - S(\omega_k)|^2\right\} \tag{25}$$

e

$$\rho(s_k, \phi_s) = \frac{s_k}{\pi\lambda_s(k)} \exp\left\{-\frac{s_k^2}{\lambda_s(k)}\right\}. \tag{26}$$

onde $\lambda_s(k)$ e $\lambda_i(k)$ são as variâncias da k-ésima componente espectral do segmento de voz e ruído, respectivamente. Substituindo as equações (25) e (26) na Equação (24), chega-se ao estimador ótimo:

$$\hat{S}_{\omega_k} = \sqrt{\lambda_k}\Gamma(1.5)\Phi(-0.5, 1; -v_k). \tag{27}$$

Na equação acima, $\Gamma(\cdot)$ é a função gama, $\Phi(-0.5, 1; -v_k)$ é a função hipergeométrica confluente, λ_k é definida como

$$\lambda_k = \frac{\lambda_s(k)\lambda_i(k)}{\lambda_s(k) + \lambda_i(k)} = \frac{\lambda_s(k)}{1 - \psi_k} \tag{28}$$

e v_k é dado por

$$v_k = \frac{\psi_k}{1 + \psi_k}\gamma_k \tag{29}$$

onde a SNR *a priori* ψ_k e a SNR *a posteriori* γ_k são expressas como

$$\psi_k = \frac{\lambda_s(k)}{\lambda_i(k)} \tag{30}$$

e

$$\gamma_k = \frac{X_{\omega_k}}{\lambda_i(k)}. \tag{31}$$

Com as equações (29) e (30), $\sqrt{\lambda_k}$ pode ser simplificado como mostrado abaixo

$$\begin{aligned}
\sqrt{\lambda_k} &= \sqrt{\frac{\lambda_i(k)}{1+\psi_k}} \\
&= \sqrt{\frac{\psi_k \lambda_i(k)}{1+\psi_k}} \\
&= \sqrt{\frac{\psi_k X_{\omega_k}^2}{(1+\psi_k)\gamma_k}\frac{\gamma_k}{\gamma_k}} \\
&= \sqrt{\frac{\psi_k \gamma_k}{(1+\psi_k)}\frac{1}{(\gamma_k)^2}X_{\omega_k}} \\
&= \frac{\sqrt{v_k}}{\gamma_k}X_{\omega_k}. \qquad (32)
\end{aligned}$$

Com o uso de funções de Bessel para expressar a função hipergeométrica confluente na Equação (27), o estimador ótimo pode ser redefinido como

$$\hat{S}_{\omega_k} = \frac{\sqrt{\pi}}{2}\frac{\sqrt{v_k}}{\gamma_k}\exp\left(\frac{-v_k}{2}\right)\left[(1+v_k)I_0\left(\frac{v_k}{2}\right)+v_k I_1\left(\frac{v_k}{2}\right)\right] \qquad (33)$$

onde $I_0(\cdot)$ e $I_1(\cdot)$ representam as funções de Bessel de ordem zero e primeira ordem, respectivamente.

A Equação (33) expressa as estimativas dos espectros de magnitude pela ação de um filtro de ganho, isto é, $\hat{S}_{\omega_k} = G(\psi_k, \gamma_k)X_{\omega_k}$. Sendo assim, $G(\psi_k, \gamma_k)$ é uma função dependente da SNR *a priori* e da SNR *a posteriori*.

Comparado com o filtro de Wiener, o estimador MMSE-STSA se comporta de forma similar para altas SNRs de entrada. Além disso, o ganho de $G(\psi_k, \gamma_k)$ introduz menos distorções em baixas SNRs de entrada (LOIZOU, 2013).

3 Representações Esparsas

Encontrar representações compactas de sinais a partir de dicionários sobrecompletos é uma tarefa que tem recebido muita atenção nos últimos anos. As representações esparsas têm se mostrado ferramentas poderosas com várias áreas de aplicação, como processamento de sinais, processamento de imagens e aprendizado de máquinas (BRUCKSTEIN; DONOHO; ELAD, 2009; ELAD; FIGUEIREDO; MA, 2010; DONOHO *et al.*, 2006).

As representações esparsas têm sua origem associada à amostragem compressiva (CS, do inglês *Compressed Sensing*). Donoho *et al.* (2006) foi o primeiro a propor o conceito de amostragem compressiva, mostrando que se um sinal é compressível ou esparso, pode ser reconstruído por um pequeno número de valores medidos, diferente de teorias anteriores como o teorema de amostragem de Nyquist-Shannon. A publicação de estudos teóricos sobre a amostragem compressiva permitiu a elaboração de um grande número de algoritmos abordando diferentes problemas em várias areas. A amostragem compressiva baseia-se em três etapas: representação esparsa, codificação de medição e algoritmo de reconstrução. A representação esparsa é a parte de maior destaque na amostragem compressiva, pois integra amostragem e compressão de codificação, superando as limitações do teorema de amostragem de Nyquist-Shannon e permitindo a amostragem de grandes quantidades de dados com economia de memória (ELAD; FIGUEIREDO; MA, 2010).

Uma observação $\boldsymbol{x} \in \mathbb{R}^N$ é representada por uma combinação linear de vetores, conhecidos como átomos. O conjunto $\boldsymbol{D} = [\boldsymbol{d}_1, ..., \boldsymbol{d}_K]$ dos átomos em forma matricial é chamado de dicionário, com $\boldsymbol{D} \in \mathbb{R}^{N \times K}$. Se o número de átomos de $\boldsymbol{D}$ é maior que a dimensão de $\boldsymbol{x}$, o dicionário é considerado sobrecompleto e $N < K$. A decomposição de $\boldsymbol{x}$ em termos dos átomos de $\boldsymbol{D}$ é escrita na forma

$$\boldsymbol{x} = \sum_{i=1}^{K} c_i \boldsymbol{d}_i. \tag{34}$$

A Equação (34) pode ser reescrita em forma matricial:

$$\boldsymbol{x} = \boldsymbol{D}\boldsymbol{c} \tag{35}$$

sendo $\boldsymbol{c} = [c_1, ..., c_K]^T$ o vetor da representação esparsa de $\boldsymbol{x}$ dada por $\boldsymbol{D}$.

Contudo, a Equação (35) é um sistema de equações lineares indeterminado, o que leva a infinitas soluções. A sua formulação não apresenta nenhuma condição ou restrição imposta sobre o vetor-solução $\boldsymbol{c}$. Para a tarefa de se obter uma representação esparsa é necessário impor uma condição que deve limitar, ou minimizar, o número de elementos não nulos de $\boldsymbol{c}$. A

solução mais esparsa é obtida ao se impor uma restrição de minimização da pseudo-norma ℓ_0 de $\boldsymbol{c}$, levando a Equação (35) a assumir a forma

$$\hat{\boldsymbol{c}} = argmin\|\boldsymbol{c}\|_0 \text{ sujeito a } \boldsymbol{x} = \boldsymbol{D}\boldsymbol{c} \tag{36}$$

onde $\|\cdot\|_0$ é a norma ℓ_0 e indica o número de elementos não nulos do vetor, sendo considerada uma medida de esparsidade. Se apenas $m(m \ll K)$ átomos do dicionário $\boldsymbol{D}$ forem escolhidos para representar $\boldsymbol{x}$, chega-se ao seguinte problema de otimização, equivalente à Equação (36):

$$\boldsymbol{x} = \boldsymbol{D}\boldsymbol{c} \text{ sujeito a } \|\boldsymbol{c}\|_0 \leq m. \tag{37}$$

Observações de fenômenos na vida real contêm ruído. Neste caso, o modelo da Equação (35) pode ser modificado para considerar a presença de ruído

$$\boldsymbol{x} = \boldsymbol{D}\boldsymbol{c} + \boldsymbol{i} \tag{38}$$

onde $\boldsymbol{i} \in \mathbb{R}^N$ é o ruído aditivo com energia limitada, isto é, $\|\boldsymbol{i}\|_2 \leq \varepsilon$. Os problemas de otimização apresentados nas equações (36) e (37) podem ser reformulados para sinais sob a presença de ruído:

$$\hat{\boldsymbol{c}} = argmin\|\boldsymbol{c}\|_0 \text{ sujeito a } \|\boldsymbol{x} - \boldsymbol{D}\boldsymbol{c}\|_2^2 \leq \varepsilon \tag{39}$$

e

$$\hat{\boldsymbol{c}} = argmin\|\boldsymbol{x} - \boldsymbol{D}\boldsymbol{c}\|_2^2 \text{ sujeito a } \|\boldsymbol{c}\|_0 \leq m. \tag{40}$$

De acordo com Zhang *et al.* (2015), com o teorema dos multiplicadores de Lagrange, as equações (39) e (40) são equivalentes ao seguinte problema de minimização sem restrições com um λ adequado:

$$\hat{\boldsymbol{c}} = L(c, \lambda) = argmin\|\boldsymbol{x} - \boldsymbol{D}\boldsymbol{c}\|_2^2 + \lambda\|\boldsymbol{c}\|_0 \tag{41}$$

onde λ é o multiplicador de Lagrange associado à pseudo-norma ℓ_0.

O modelo de representação com minimização da norma ℓ_0 fornece um vetor-solução esparso a partir do dicionário $\boldsymbol{D}$, contudo este problema de otimização é um problema do tipo *NP-hard*, tornando difícil se aproximar da solução (AMALDI; KANN, 1998). Modelos de representação com minimização da norma ℓ_1 de $\boldsymbol{c}$ levam à soluções suficientemente esparsas e podem ser equivalentes às soluções obtidas com a minimização da norma ℓ_0 com probabilidade total (DONOHO, 2006; CANDES; ROMBERG; TAO, 2006). Além disso, os problemas de otimização com norma ℓ_1 possuem soluções analíticas e são resolvidos em tempo polinomial, ou seja, possuem tempo de execução que é limitado superiormente por uma expressão polinomial

no tamanho da entrada do algoritmo (PAPADIMITRIOU,). Os modelos de otimização com minimização da norma ℓ_1 têm formulações semelhantes aos modelos com norma ℓ_0, como é mostrado abaixo.

$$\hat{\boldsymbol{c}} = argmin\|\boldsymbol{c}\|_1 \text{ sujeito a } \boldsymbol{x} = \boldsymbol{D}\boldsymbol{c} \tag{42}$$

$$\hat{\boldsymbol{c}} = argmin\|\boldsymbol{c}\|_1 \text{ sujeito a } \|\boldsymbol{x} - \boldsymbol{D}\boldsymbol{c}\|_2^2 \leq \varepsilon \tag{43}$$

ou

$$\hat{\boldsymbol{c}} = argmin\|\boldsymbol{x} - \boldsymbol{D}\boldsymbol{c}\|_2^2 \text{ sujeito a } \|\boldsymbol{c}\|_1 \leq \tau \tag{44}$$

$$\hat{\boldsymbol{c}} = L(c, \lambda) = argmin\|\boldsymbol{x} - \boldsymbol{D}\boldsymbol{c}\|_2^2 + \lambda\|\boldsymbol{c}\|_1 \tag{45}$$

para λ e τ suficientemente pequenos.

3.1 Estratégias de Aproximação para Representações Esparsas com Minimização da Norma ℓ_0

Representações esparsas geradas por problemas de otimização com a minimização da norma ℓ_0 são problemas *NP-hard*. Estratégias de aproximação com algoritmos gulosos são uma forma de se aproximar da solução mais esparsa nos problemas de otimização apresentados nas equações (39) e (40).

O algoritmo guloso é uma técnica de projeto de algoritmos que constrói uma solução pedaço a pedaço, escolhendo a cada iteração o elemento que oferece o maior benefício imediato. São conceitualmente muito simples, e são geralmente utilizados para resolver problemas de otimização. Um algoritmo guloso inicia com um conjunto vazio de candidatos escolhidos. A cada iteração, tenta-se adicionar o melhor candidato entre os restantes, guiado por uma função de seleção. Se a viabilidade do conjunto de candidatos escolhidos for comprometida, o candidato é removido e nunca mais selecionado. Porém se o conjunto permanecer viável, o novo candidato permanece no conjunto de candidatos escolhidos. A cada iteração é verificado se o conjunto formado é uma solução para o problema, e se o algoritmo opera corretamente, a solução obtida nem sempre é um ótimo global. O termo "guloso" refere-se ao fato de que, a cada iteração, o maior pedaço é devorado, e esta decisão não é questionada no futuro: se um candidato é adicionado ao conjunto, ele nunca mais será retirado, e se um candidato for retirado do conjunto, ele não será mais considerado (BRASSARD; BRATLEY, 1988).

Mallat e Zhang (1993) foram os primeiros a propor um algoritmo guloso, chamado de *Matching Pursuit* (MP), para aproximar os problemas de otimização restritos com norma ℓ_0. O

conceito central do MP é selecionar os melhores átomos do dicionário $\boldsymbol{D}$ através de uma medida de similaridade para assim obter de forma aproximada a solução esparsa de $\boldsymbol{x}$. A codificação de $\boldsymbol{x}$ em relação a $\boldsymbol{D} \in \mathbb{R}^{N \times K}$ pelo MP inicia-se com o resíduo de representação $\boldsymbol{R}_0 = \boldsymbol{x}$, com cada átomo de $\boldsymbol{D}$ possuindo energia unitária. Então é selecionado o átomo com maior produto interno em módulo em relação a $\boldsymbol{x}$, denotado por $\boldsymbol{d}_{l_0}$:

$$|\langle \boldsymbol{R}_0, \boldsymbol{d}_{l_0} \rangle| = sup|\langle \boldsymbol{R}_0, \boldsymbol{d}_i \rangle| \tag{46}$$

onde l_0 é um índice de classificação do dicionário $\boldsymbol{D}$. Então $\boldsymbol{x}$ é decomposto de acordo com a Equação a seguir:

$$\boldsymbol{x} = |\langle \boldsymbol{R}_0, \boldsymbol{d}_{l_0} \rangle| \boldsymbol{d}_{l_0} + \boldsymbol{R}_1 \tag{47}$$

com $|\langle \boldsymbol{R}_0, \boldsymbol{d}_{l_0} \rangle| \boldsymbol{d}_{l_0}$ sendo a projeção ortogonal de $\boldsymbol{R}_0$ sobre $\boldsymbol{d}_{l_0}$ e $\boldsymbol{R}_1$ é o resíduo necessário para representar $\boldsymbol{x}$.

O átomo encontrado a partir da Equação (46) minimiza $\boldsymbol{R}_1$ na Equação (47). Para minimizar o resíduo de representação, o MP escolhe os melhores átomos a partir do dicionário sobrecompleto e utiliza o resíduo na iteração seguinte até atingir um erro de aproximação satisfatório ou o número de átomos desejado. Para a t-ésima iteração, o melhor átomo correspondente é encontrado pela Equação (48).

$$|\langle \boldsymbol{R}_t, \boldsymbol{d}_{l_t} \rangle| = sup|\langle \boldsymbol{R}_t, \boldsymbol{d}_i \rangle| \tag{48}$$

e a representação do t-ésimo resíduo é dada por

$$\boldsymbol{R}_t = |\langle \boldsymbol{R}_t, \boldsymbol{d}_{l_t} \rangle| \boldsymbol{d}_{l_t} + \boldsymbol{R}_{t+1}. \tag{49}$$

Se a energia do resíduo na n-ésima iteração for suficientemente pequena, $\boldsymbol{x}$ pode ser aproximado pela Equação (50).

$$\boldsymbol{x} \approx \sum_{j=1}^{n-1} \langle \boldsymbol{R}_j, \boldsymbol{d}_{l_j} \rangle \boldsymbol{d}_{l_j} \tag{50}$$

onde $n \ll N$.

Uma variação do MP é o *Orthogonal Matching Pursuit* (OMP), detalhado no Algoritmo 1. Como ocorre no MP, cada iteração do OMP consiste na seleção do átomo mais coerente seguido da atualização do vetor de representação. Na t-ésima iteração, o átomo mais coerente com o resíduo $\boldsymbol{R}_t$ é selecionado e seu índice k^* é adicionado ao conjunto de átomos escolhidos $\mathcal{A}$. O vetor de representação $\boldsymbol{c}_t$ é atualizado para as coordenadas da projeção ortogonal de $\boldsymbol{x}$ sobre o subespaço gerado pela matriz $\boldsymbol{D}_{(:,\mathcal{A})}$ e então o novo resíduo $\boldsymbol{R}_{t+1}$ é calculado. Atualizar o vetor de representação dessa maneira assegura que $\boldsymbol{R}_t$ é sempre ortogonal à $\boldsymbol{D}_{(:,\mathcal{A})}$ e que seu

conjunto de átomos é linearmente independente. Tal método pode levar a resultados melhores que o MP, porém implica em maior custo computacional.

Algorithm 1 Orthogonal Matching Pursuit

Entrada: $\boldsymbol{x} \in \mathbb{R}^N, \boldsymbol{D} \in \mathbb{R}^{N \times K}, m$ ou ϵ
Saída: $\boldsymbol{c} \in \mathbb{R}^K$
  $\mathcal{A} \leftarrow \{\}, \boldsymbol{c} \leftarrow \boldsymbol{0}, \boldsymbol{R} \leftarrow \boldsymbol{x}$
  enquanto $\|\boldsymbol{c}\|_0 \leq m$ ou $\|\boldsymbol{R}\|_2 \leq \epsilon$ **faça**
    $\boldsymbol{\mu} \leftarrow \boldsymbol{D}^T\boldsymbol{R}$
    $k^* \leftarrow argmax_k|\mu_k|, k \in \mathcal{A}^c$
    $\mathcal{A} \leftarrow \mathcal{A} \cup \{k^*\}$
    $\boldsymbol{c}_\mathcal{A} \leftarrow (\boldsymbol{D}_\mathcal{A}^T\boldsymbol{D}_\mathcal{A})^{-1}\boldsymbol{D}_\mathcal{A}^T\boldsymbol{x}$
    $\boldsymbol{R} \leftarrow \boldsymbol{x} - \boldsymbol{D}\boldsymbol{c}$
  fim enquanto

3.2 Estratégias de Aproximação para Representações Esparsas com Minimização da Norma ℓ_1

Uma alternativa à minimização da norma ℓ_0 do vetor-solução, é a minimização de sua norma ℓ_1. O modelo de otimização com penalização da norma ℓ_1 é conhecido como LASSO, descrito na Equação (45). O LASSO é uma ferramenta muito popular para estimar os coeficientes de uma representação, principalmente para o cenário envolvendo dicionários sobrecompletos. A esparsidade da representação é indiretamente controlada por λ. A escolha de valores elevados de λ leva a soluções mais esparsas com o LASSO. No geral o LASSO não possui solução em forma fechada, contudo ele pode ser resolvido de maneira eficiente a partir de algoritmos iterativos (TIBSHIRANI, 2011).

O algoritmo LARS (do inglês, *Least Angle Regression*) é um algoritmo guloso que permite, com pequenas modificações em sua estrutura, a estimação simultânea de todas as estimativas do modelo LASSO (EFRON *et al.*, 2004). Para estimar os coeficientes da representação de $\boldsymbol{x}$, inicia-se com o resíduo $\boldsymbol{R}_0 = \boldsymbol{x}$, o dicionário sobrecompleto $\boldsymbol{D}$ e o vetor de representação nulo $\boldsymbol{c}$. É escolhido o átomo com maior correlação com $\boldsymbol{R}_0$, denotado por $\boldsymbol{d}_{l_0}$. O coeficiente c_{l_0} é movido na direção do produto interno $\langle \boldsymbol{R}_0, \boldsymbol{d}_{l_0} \rangle$ até que algum outro átomo $\boldsymbol{d}_{l_1}$ tenha correlação com $\boldsymbol{R}_1 = \boldsymbol{R}_0 - c_{l_0}\boldsymbol{d}_{l_0}$ tão grande quanto $\boldsymbol{d}_{l_0}$. Os coeficientes c_{l_0} e c_{l_1}, são então movidos na direção dos seus mínimos quadrados conjuntos até outro átomo apresentar correlação com $\boldsymbol{R}_2 = \boldsymbol{R}_1 - c_{l_0}\boldsymbol{d}_{l_0} - c_{l_1}\boldsymbol{d}_{l_1}$ tão grande quanto $\boldsymbol{d}_{l_0}$ e $\boldsymbol{d}_{l_1}$. O algoritmo finaliza quando todos os átomos são selecionados. Para que os resultados do LARS sejam idênticos ao LASSO, é necessário modificar o algoritmo para que átomos com coeficientes nulos sejam eliminados e a direção de busca seja recalculada.

Sigg, Dikk e Buhmann (2012) apresentaram uma extenção do LARS, intitulada LARC (do inglês, *Least Angle Regression with Coherence Criterion*). No LARC, a matriz grammiana $\boldsymbol{G} = \boldsymbol{D}^T\boldsymbol{D}$ é utilizada para evitar os repetidos cálculos de produtos matriz-vetor envolvendo $\boldsymbol{D}$, diminuindo assim o custo computacional. Ao representar um sinal, tanto o OMP quanto o LARS codificam inicialmente seus componentes mais coerentes com o dicionário, e depois seus componentes incoerentes, isto é, a coerência do resíduo diminui a cada iteração. O LARC usa um limiar de coerência μ_{dl} como critério de parada, ao invés da norma ℓ_2 do resíduo ou a norma ℓ_1 do vetor de representação. Tal escolha não depende da magnitude do sinal a ser representado.

Três partes constituem o LARC, detalhado no Algoritmo 2: seleção do átomo mais coerente, cálculo da direção equiangular $\boldsymbol{u}$ e atualização de $\boldsymbol{c}$ a partir de γ. O produto $\boldsymbol{D}^T\boldsymbol{R}_t$ é calculado pelo termo constante $\boldsymbol{D}^T\boldsymbol{x}$ e pelo termo $\boldsymbol{D}^T\boldsymbol{y}_t$ que é computado eficientemente a cada iteração.

Algorithm 2 LARC

Entrada: $\boldsymbol{x} \in \mathbb{R}^N, \boldsymbol{D} \in \mathbb{R}^{NxK}, \boldsymbol{G} = \boldsymbol{D}^T\boldsymbol{D}, \mu_{dl}$
Saída: $\boldsymbol{c} \in \mathbb{R}^K$

$\mathcal{A} \leftarrow \{\}, \boldsymbol{c} \leftarrow \boldsymbol{0}, \boldsymbol{y} \leftarrow \boldsymbol{0}$
$\boldsymbol{\mu}^{(x)} \leftarrow \boldsymbol{D}^T\boldsymbol{x}; \boldsymbol{\mu}^{(y)} \leftarrow \boldsymbol{0}$
enquanto $\|\mathcal{A}| < K$ **faça**
  $\boldsymbol{\mu} \leftarrow \boldsymbol{\mu}^{(x)} - \boldsymbol{\mu}^{(y)}$
  $k^* \leftarrow argmax_k|\mu_k|, k \in \mathcal{A}^c$
  $\mathcal{A} \leftarrow \mathcal{A} \cup \{k^*\}$
  se $|\mu_{k^*}|/\|\boldsymbol{x} - \boldsymbol{y}\| < \mu_{dl}$ **então**
    break
  fim se
  $\boldsymbol{s} \leftarrow \text{sign}(\boldsymbol{\mu}_{\mathcal{A}})$
  $\boldsymbol{g} \leftarrow \boldsymbol{G}^{-1}_{(\mathcal{A},\mathcal{A})}\boldsymbol{s}$
  $b \leftarrow (\boldsymbol{g}^T\boldsymbol{s})^{-1/2}$
  $\boldsymbol{w} \leftarrow b\boldsymbol{g}$
  $\boldsymbol{u} \leftarrow \boldsymbol{D}_{(:,\mathcal{A})}\boldsymbol{w}$
  $\boldsymbol{a} \leftarrow \boldsymbol{G}_{(:,\mathcal{A})}\boldsymbol{w}$
  $\gamma \leftarrow \min^+_{k\in\mathcal{A}^c}[(|\mu_{k^*}| - \mu_k)/(b - a_k), (|\mu_{k^*}| + \mu_k)/(b + a_k)]$
  $\boldsymbol{y} \leftarrow \boldsymbol{y} + \gamma\boldsymbol{u}$
  $\boldsymbol{c}_{\mathcal{A}} \leftarrow \boldsymbol{c}_{\mathcal{A}} + \gamma\boldsymbol{w}$
  $\boldsymbol{\mu}_{(y)} \leftarrow \boldsymbol{\mu}_{(y)} + \gamma\boldsymbol{a}$
fim enquanto

3.3 Representações Esparsas Utilizadas no Treinamento de Dicionários

A escolha do dicionário é uma etapa importante para a representação esparsa de sinais em uma determinada aplicação. Dicionários podem ser compostos por formas de onda predefinidas ou podem ser obtidos com base em treinamento, isto é, métodos de treinamento de dicionários. Dicionários formados por grupos de sinais bem definidos, chamados de dicionários analíticos, podem ser ortogonais, biortogonais ou sobrecompletos, e possuem formulações simples e implementações rápidas. Os primeiros dicionários analíticos a serem utilizados foram as transformadas existentes, como Fourier, composta por exponenciais complexas harmonicamente relacionadas; Wavelet, contituída por bancos de filtros com diferentes configurações; e Gabor, formada por formas de ondas discretas derivadas das funções de Gabor, contituídas por versões transladadas e moduladas de uma janela. Dicionários fixos também podem ser formados por átomos gerados aleatóriamente, alcançando propriedades desejadas, como a incoerência entre os átomos. Dicionários estruturados, como por exemplo aqueles formados pela união de bases ortonormais, também possuem estrutura fixa, contudo apresentam flexibilidade relativamente restrita na prática (RUBINSTEIN; BRUCKSTEIN; ELAD, 2010).

Já dicionários treinados são sobrecompletos e são otimizados para representar sinais de uma classe específica de maneira esparsa com erro de aproximação menor que dicionários analíticos. Eles não possuem estrutura interna, portanto não existem algoritmos para implementação rápida.

Um algoritmo de treinamento de dicionários (DL, do inglês *Dictionary Learning*) modifica um dicionário inicial para representar uma determinada classe de sinais, objetivando alcançar o melhor, ou ao menos um muito adequado, dicionário para o problema exposto na Equação (51), considerando a presença de ruído de energia limitada no conjunto de sinais de treinamento. O treinamento de dicionários é formalmente um problema de fatoração matricial formulado como um problema de otimização:

$$argmin_{\boldsymbol{D},\boldsymbol{C}} \|\boldsymbol{X} - \boldsymbol{D}\,\boldsymbol{C}\|_F^2 \quad \text{sujeito a} \quad \begin{array}{ll} \|\boldsymbol{c}_i\|_0 \leq K & \forall i \\ \|d_j\| = 1 & \forall j \end{array} \tag{51}$$

onde $\| \bullet \|_F$ é a norma Frobenius, também conhecida como norma matricial, $\boldsymbol{X} \in \mathbb{R}^{N \times L}$ é a matriz composta pelo conjunto de amostras de treinamento, $\boldsymbol{D} \in \mathbb{R}^{N \times K}$ é o dicionário sobrecompleto com átomos de norma unitária e $\boldsymbol{c}_i$ são as colunas de $\boldsymbol{C} \in \mathbb{R}^{K \times L}$, a matriz de representação de $\boldsymbol{X}$. A primeira restrição controla a esparsidade de cada coluna de $\boldsymbol{C}$, limitando para no máximo K elementos não nulos. A segunda restrição impõe que todos os átomos sejam normalizados, característica herdada de dicionários analíticos. O problema acima é um problema de otimização em relação à $\boldsymbol{D}$ e $\boldsymbol{C}$ e é altamente não-convexo. Então no melhor

cenário é possível se buscar um mínimo local como solução (RUBINSTEIN; BRUCKSTEIN; ELAD, 2010).

O problema de minimização da Equação (51) pode ser dividido em 2 subproblemas: representação esparsa e atualização do dicionário. Solucionar os dois subproblemas de forma alternada é a abordagem feita pelos principais algoritmos de treinamento de dicionários. Quando $\boldsymbol{D}$ é fixado, a matriz de representação esparsa $\boldsymbol{C}$ é calculada por algum algoritmo de codificação esparsa, como o OMP, como mostrado pelo subproblema da Equação (52).

$$argmin_{\boldsymbol{C}} \|\boldsymbol{X} - \boldsymbol{D}\boldsymbol{C}\|_F^2 \quad \text{sujeito a} \quad \|\boldsymbol{c}_i\|_0 \leq K \quad \forall i \tag{52}$$

Quando $\boldsymbol{C}$ é fixado, o dicionário $\boldsymbol{D}$ é atualizado, o que se configura como um problema de otimização não linear descrito na Equação (53). Alguns algoritmos de DL também atualizam os coeficientes de representação, isto é, os elementos não nulos de $\boldsymbol{C}$. A normalização dos átomos é feita durante a atualização do dicionário ou ao final do DL.

$$\hat{D} = argmin_{\boldsymbol{D}} \|\boldsymbol{X} - \boldsymbol{D}\boldsymbol{C}\|_F^2 \tag{53}$$

Um dos algoritmos de DL mais utilizados é o K-SVD (Do inglês *K-Singular Value Decomposition*). Ele é um algoritmo do tipo *Block Coordinate Descent* (BCD). Algoritmos de descida coordenada são métodos iterativos que, a cada iteração, fixam muitas das coordenadas do vetor a ser otimizado, e minimizam a função objetivo em relação às coordenadas restantes. Se um bloco de coordenadas for atualizado a cada iteração, o algoritmo passa a ser classificado como um método BCD (YUAN; LU; LI, 2014).

A partir da Equação (51), o produto matricial $\boldsymbol{D}\boldsymbol{C}$ pode ser expresso pela soma dos produtos entre os átomos de $\boldsymbol{D}$ e as linhas de $\boldsymbol{C}$:

$$\boldsymbol{D}\boldsymbol{C} = \sum_{i=1}^{K} \boldsymbol{d}_i \boldsymbol{c}_T^i \tag{54}$$

onde em $\boldsymbol{c}_T^j$, a j-ésima linha $\boldsymbol{C}$, estão os coeficientes referentes ao átomo $\boldsymbol{d}_j$. Como $\boldsymbol{C}$ é uma matriz esparsa, $\boldsymbol{d}_j$ participa da representação de apenas algumas amostras de treinamento. O conjunto $\mathcal{N}_j = \{l | 1 \leq l \leq L, \boldsymbol{c}_T^j(l) \neq 0\}$ indica os índices das amostras que usam $\boldsymbol{d}_j$, isto é, os índices dos elementos não nulos de $\boldsymbol{c}_T^j$.

Como ocorre com boa parte dos algoritmos de DL, O K-SVD é dividido nas etapas de codificação esparsa e atualização do dicionário. A codificação esparsa busca minimizar o problema de otimização da Equação (52) fixando $\boldsymbol{D}$ e utilizando algum algoritmo de busca com minimização da norma ℓ_0, como o OMP. Na etapa de atualização dos átomos do dicionário, considerando $\boldsymbol{D}$ e $\boldsymbol{C}$ fixos, o K-SVD seleciona e atualiza cada átomo e seu vetor de representação isoladamente, até todos os átomos de $\boldsymbol{D}$ serem otimizados. A função objetivo da Equação (53) pode ser reescrita como segue

$$\begin{aligned} \|\boldsymbol{X} - \boldsymbol{D}\boldsymbol{C}\|_F^2 &= \|\boldsymbol{X} - \sum_{i=1}^{K} \boldsymbol{d}_i \boldsymbol{c}_T^i\|_F^2 \\ &= \|(\boldsymbol{X} - \sum_{i\neq k}^{K} \boldsymbol{d}_i \boldsymbol{c}_T^i) - \boldsymbol{d}_k \boldsymbol{c}_T^k\|_F^2 \\ &= \|\boldsymbol{E}_k - \boldsymbol{d}_k \boldsymbol{c}_T^k\|_F^2 \end{aligned} \tag{55}$$

Como mostrado na Equação (54), o produto $\boldsymbol{D}\boldsymbol{C}$ é decomposto pela soma de K produtos de matrizes de posto 1. O k-ésimo produto é isolado, enquanto os $K-1$ restantes são fixados. A matriz $\boldsymbol{E}_k$ é o erro de aproximação entre as amostras de treinamento e a representação esparsa dada pelos $K-1$ átomos de $\boldsymbol{D}$. Minimizar a norma residual na Equação (55) atualizando o k-ésimo átomo e seus respectivos coeficientes pode levar a um novo $\boldsymbol{c}_T^k$ que não seja esparso, já que a otimização de $\boldsymbol{d}_k$ não induz uma restrição de esparsidade.

A esparsidade de $\boldsymbol{c}_T^k$ pode ser mantida se a otimização envolver apenas os coeficientes não-nulos e suas respectivas amostras. A matriz $\boldsymbol{\Omega}_j \in \mathbb{R}^{L\times|\mathcal{N}_j|}$ é definida assumindo 1 nos índices $(\mathcal{N}_j(l), l)$ e 0 nos demais índices. O produto $\boldsymbol{c}_R^k = \boldsymbol{c}_T^k\boldsymbol{\Omega}_j$ elimina os elementos nulos de $\boldsymbol{c}_T^k$ gerando um vetor de comprimento $|\mathcal{N}_j|$. Já a multiplicação $\boldsymbol{E}_k^R = \boldsymbol{E}_k\boldsymbol{\Omega}_j$ seleciona as colunas de $\boldsymbol{E}_k$ referentes as amostras de treinamento que usam $\boldsymbol{d}_k$ para representá-las. A função objetivo da equação (55) pode ser reformulada para forçar que a atualização de $\boldsymbol{c}_T^k$ mantenha a esparsidade. Isto leva a minimização de

$$\|\boldsymbol{E}_k\boldsymbol{\Omega}_j - \boldsymbol{d}_k\boldsymbol{c}_T^k\boldsymbol{\Omega}_j\|_F^2 = \|\boldsymbol{E}_k^R - \boldsymbol{d}_k\boldsymbol{c}_R^k\|_F^2. \tag{56}$$

A solução para o problema é obtido via SVD. A matriz $\boldsymbol{E}_k^R$ é decomposta via SVD resultando no produto $\boldsymbol{E}_k^R = \boldsymbol{U}\boldsymbol{\Delta}\boldsymbol{V}^T$. O átomo $\boldsymbol{d}_k$ é substituído pela primeira coluna de $\boldsymbol{U}$ e $\boldsymbol{c}_T^k$ é substituído pelo produto entre $\boldsymbol{\Delta}(1,1)$ e a primeira coluna de $\boldsymbol{V}$. Esta solução garante a normalização de $\boldsymbol{D}$ e o mesmo número de elementos não nulos de $\boldsymbol{C}$.

O cálculo da decomposição SVD é a etapa com maior custo computacional do K-SVD. Para evitar este problema mantendo a ideia na etapa de atualização opta-se por otimizar primeiramente o átomo e depois otimizar os coeficientes de representação. Rubinstein, Zibulevsky e Elad (2008) propuseram o AK-SVD (Do inglês *Approximate K-Singular Value Decomposition*) com uma etapa de atualização que aproxima o SVD de $\boldsymbol{E}_k^R$ usando o método das potências, como mostrado no Algoritmo 3.

Tanto o K-SVD quanto o AK-SVD garantem a diminuição do erro na etapa de atualização dos átomos do dicionário. No K-SVD o decrescimento é máximo, enquanto no AK-SVD progride em direção a um mínimo. Porém, nota-se na prática que com um número um pouco maior de iterações o AK-SVD atinge o mesmo erro que o K-SVD.

Algorithm 3 Atualização do dicionário do AK-SVD

Entrada: $\boldsymbol{X} \in \mathbb{R}^{N \times L}$, $\boldsymbol{D} \in \mathbb{R}^{N \times K}$, $\boldsymbol{C} \in \mathbb{R}^{K \times L}$
Saída: Dicionário atualizado $\boldsymbol{D}$
  para $l = 0 : L$ **faça**
    $\boldsymbol{d}_{(:,l)} \leftarrow 0$
    $\mathcal{N}_l \leftarrow \{n | 1 \leq n \leq N, \boldsymbol{c}_{(l,n)} \neq 0\}$
    $\boldsymbol{E} \leftarrow \boldsymbol{X}_{(:,\mathcal{N}_l)} - \boldsymbol{D}\boldsymbol{C}_{(:,\mathcal{N}_l)}$
    $\boldsymbol{g} \leftarrow \boldsymbol{c}^T_{(l,\mathcal{N}_l)}$
    $\boldsymbol{h} \leftarrow \boldsymbol{E}\boldsymbol{g}/\|\boldsymbol{E}\boldsymbol{g}\|_2$
    $\boldsymbol{g} \leftarrow \boldsymbol{E}^T\boldsymbol{h}$
    $\boldsymbol{d}_{(:,l)} \leftarrow \boldsymbol{h}$
    $\boldsymbol{c}_{(l,\mathcal{N}_l)} \leftarrow \boldsymbol{g}^T$
  fim para

4 Melhoramento de Voz usando Dicionários Treinados

Dicionários treinados têm sido amplamente utilizados no melhoramento de sinais de voz, baseando-se na capacidade da voz ser decomposta por um pequeno número de átomos pertencentes a um dicionário devidamente treinado. Sigg, Dikk e Buhmann (2012) propuseram o GDL (do inglês *Generative Dictionary Learning*), baseado na transformação do sinal ruidoso em um domínio adequado e na codificação esparsa do sinal transformado utilizando dicionários conjuntos.

O método é aplicado usando o modelo aditivo da Equação (1). Como mostrado na equação (2) a propriedade de linearidade da transformada de Fourier de tempo curto preserva a relação aditiva entre voz e ruído no domínio tempo-frequência. Tal relação é reescrita em forma matricial na Equação (57).

$$\boldsymbol{X} = \boldsymbol{S} + \boldsymbol{I} \tag{57}$$

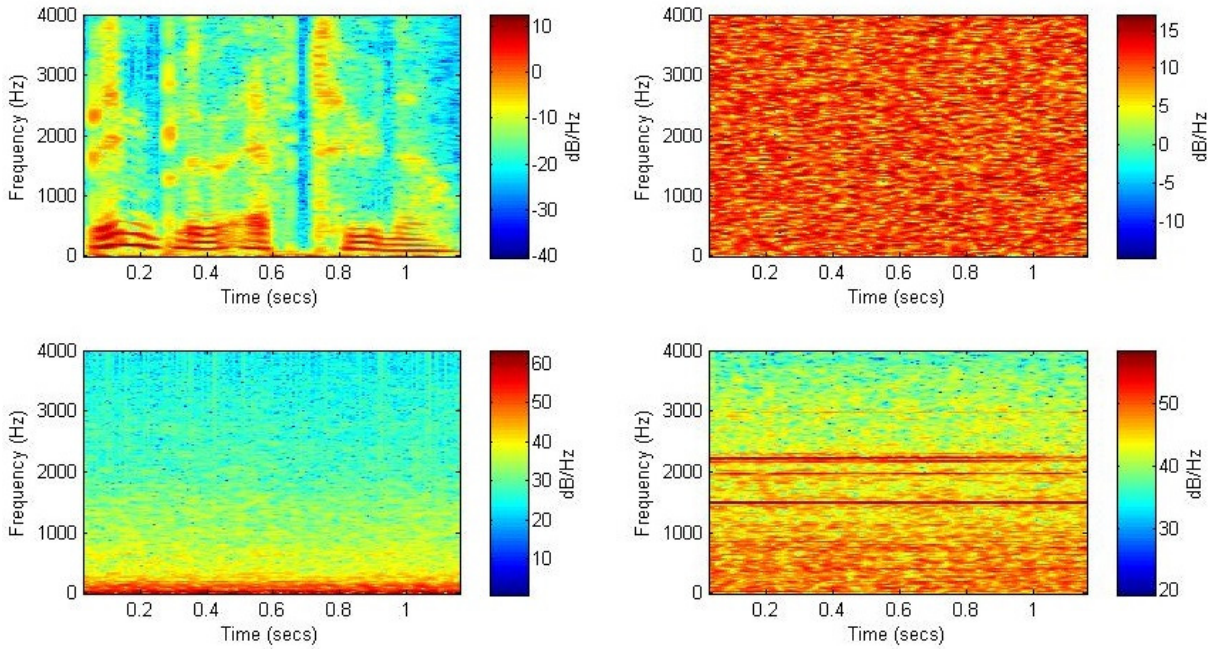
onde $\boldsymbol{X} \in \mathbb{R}^{M \times K}$, $\boldsymbol{S} \in \mathbb{R}^{M \times K}$ e $\boldsymbol{I} \in \mathbb{R}^{M \times K}$, são as matrizes de espectro de magnitude do sinal ruidoso, da voz pura e do ruído, respectivamente, e M e K são os números de amostras na frequência e no tempo. Comumente o melhoramento de voz no domínio da STFT procura modificar apenas a magnitude da STFT porque considera-se que a estrutura da voz encontra-se em grande parte contida na magnitude, ou seja, pouca informação sobre a voz é disponibilizada pela fase (GERKMANN; KRAWCZYK-BECKER; ROUX, 2015).

A voz possui uma distribuição de energia bem-definida no domínio tempo-frequência. Tal estrutura permite que a voz, assim como outras classes de sinais estruturados, seja representada de maneira esparsa com pequena margem de erro por dicionários coerentes. A coerência entre dois sinais de mesma dimensão é dada pelo módulo do produto interno entre eles. Assim, um átomo é dito coerente com um sinal se o valor absoluto do produto interno entre os dois é grande. Diferente da voz, ruídos não estruturados como o ruído gaussiano branco, Figura 4b, não podem ser codificados por uma pequena coleção de átomos de nenhum dicionário. Quando o dicionario escolhido é ortogonal, a energia do ruído não estruturado se distribui uniformemente pelos coeficientes. Mallat e Zhang (1993) mostraram que a voz contaminada por ruído gaussinano branco pode ser recuperada pelo dicionário de voz pelo fato do dicionário ser coerente com a voz, mas incoerente com o ruído.

Ruídos não estruturados não possuem padrão, já ruídos estruturados possuem padrões no domínio tempo-frequência. Tais padrões são coerentes com o dicionário de voz. Com isso, a codificação do sinal de voz ruidoso utilizando apenas o dicionário de voz pode levar a reconstru-

ção não apenas da voz pura, mas também de partes do ruído estruturado, diminuindo de forma significativa a eficiência do método. O problema pode ser diminuído codificando o sinal ruidoso em um dicionário conjunto constituído pela união de um dicionário de voz e um dicionário de ruído.

Figura 4 – Espectros de magnitude de diferentes tipo de sinais: voz pura (superior esquerdo); ruído gaussiano branco (superior direito); ruído de veículo (inferior esquerdo); ruído de contratorpedeiro (inferior direito).

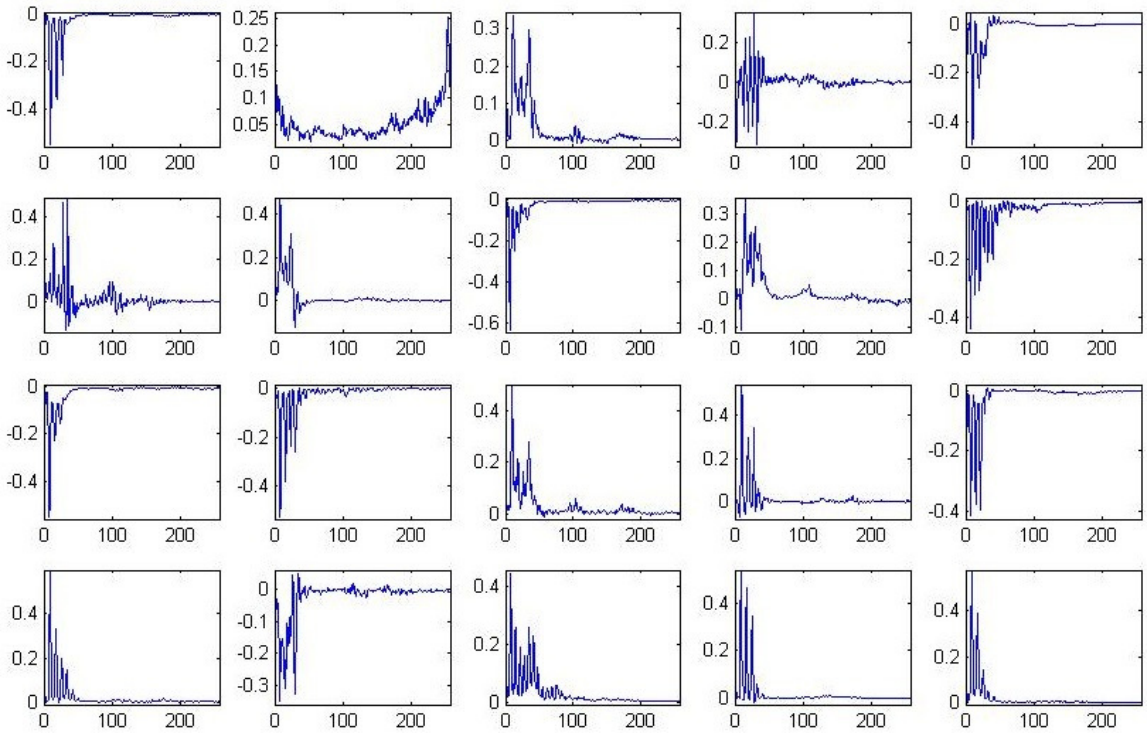


Fonte: Elaborado pelo próprio autor.

Dicionários analíticos, como bases de Fourier e Wavelets, apresentam alta coerência tanto com a voz quanto com ruídos estruturados, impossibilitando o melhoramento da voz via codificação esparsa em ambientes não-estacionários. Em contrapartida, um dicionário de voz treinado pode extrair partes da voz dada sua alta coerência com a voz pura. Da mesma forma componentes do ruído são capturados pelo dicionário de ruído. Na Figura 5, são mostrados exemplos de átomos de um dicionário treinado com amostras de sinais de voz e na Figura 6 estão alguns átomos pertencentes a um dicionário treinado com amostras de ruído de fábrica. Pode-se observar que os átomos mostrados não possuem estrutura simples. Além disso as formas apresentadas pelos átomos são capazes de representar diferentes estruturas complexas da sua respectiva classe de sinais.

Como a voz pura nunca esta disponível quando o melhoramento de voz é realizado, o treinamento do dicionário de voz é feito utilizando-se amostras de voz presentes em um banco de dados. A estrutura da voz é bem definida, o que viabiliza o uso de um dicionário treinado a partir de um modelo pré-estabelecido na etapa de melhoramento, mesmo no cenário inde-

Figura 5 – Exemplos de átomos de um dicionário de voz treinado.



Fonte: Elaborado pelo próprio autor.

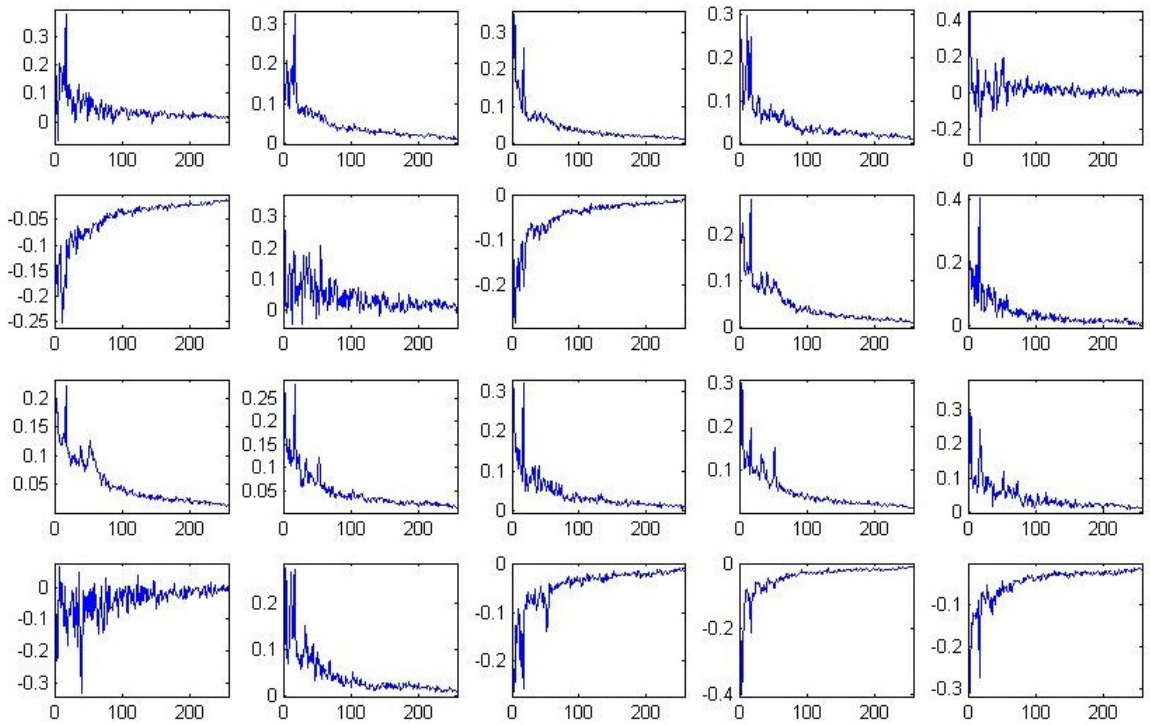
pendente de locutor. O mesmo não ocorre com o ruído estruturado, pois ele varia de acordo com o ambiente onde o áudio é captado, além de geralmente ser não-estacionário. Contudo, as amostras de treinamento do ruído podem ser capturadas nos trechos de pausa ou de silêncio, resultando em um dicionário adaptado para aquele cenário específico. Também é assumido que um detector de atividade de voz (VAD, do inglês *Voice Activity Detector*) ideal é utilizado para obter as amostras de ruído para o treinamento do dicionário de ruído.

4.1 Visão Geral do Método

O diagrama de blocos do GDL é apresentado na Figura 7. O método é dividido em duas etapas: treinamento e melhoramento. Na etapa de treinamento amostras de sinais de voz no domínio do tempo são levadas para o domínio da STFT. Então o espectro de magnitude dos sinais transformados é utilizado para treinar o dicionário de voz. O dicionário de ruído é treinado da mesma maneira, usando amostras do ruído ambiente. Os dois dicionários são treinados a partir dos problemas de otimização das equações (58) e (59).

$$argmin_{\boldsymbol{D}_s,\boldsymbol{C}_s} \|\boldsymbol{S}_t - \boldsymbol{D}_s\boldsymbol{C}_s\|_F^2 \quad \text{sujeito a} \quad \|\boldsymbol{c}_{s,k}\|_1 \leq \tau_s \quad \forall k \tag{58}$$

Figura 6 – Exemplos de átomos de um dicionário de ruído treinado com amostras de ruído de fábrica.



Fonte: Elaborado pelo próprio autor.

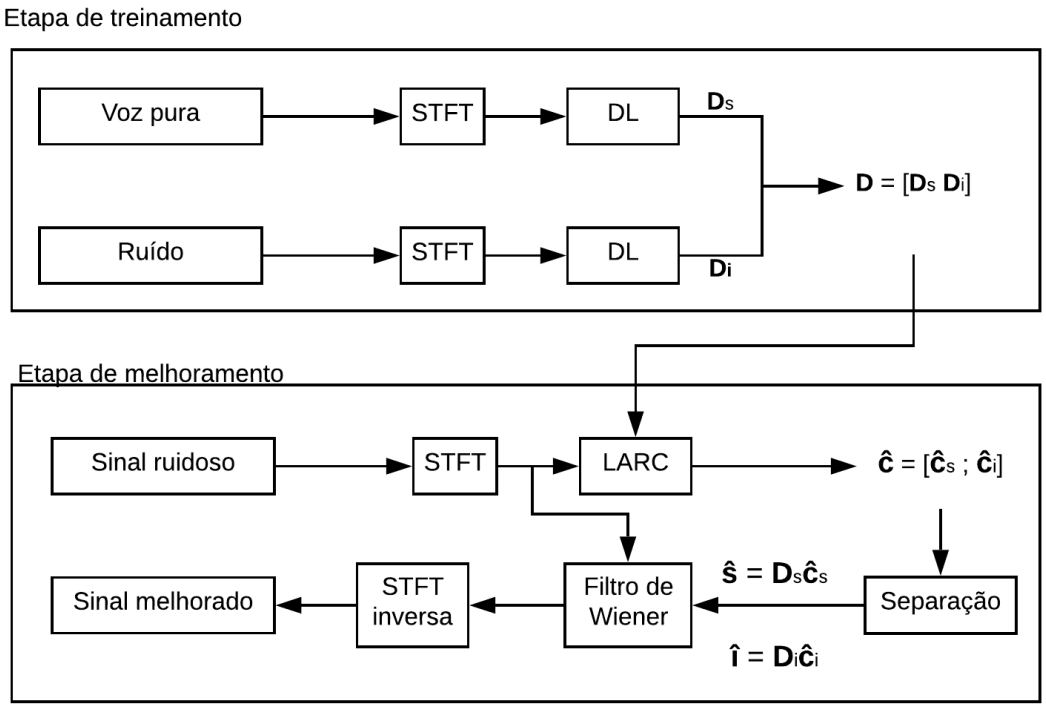
$$argmin_{\boldsymbol{D}_i,\boldsymbol{C}_i}\|\boldsymbol{I}_t - \boldsymbol{D}_i\boldsymbol{C}_i\|_F^2 \quad \text{sujeito a} \quad \|\boldsymbol{c}_{i,k}\|_1 \leq \tau_i \quad \forall k \tag{59}$$

onde $\boldsymbol{D}_s$ é o dicionário de voz e $\boldsymbol{D}_i$ é o dicionário de ruído, $\boldsymbol{S}_t$ e $\boldsymbol{I}_t$ são as matrizes de treinamento dos dicionários de voz e de ruído, respectivamente, $\| \bullet \|_1$ representa a norma ℓ_1, $\boldsymbol{c}_{s,k}$ e $\boldsymbol{c}_{i,k}$ são as k-ésimas colunas de $\boldsymbol{C}_s$ e $\boldsymbol{C}_i$, respectivamente. A constante τ_s controla a esparsidade de $\boldsymbol{C}_s$, e τ_i limita a esparsidade de $\boldsymbol{C}_i$.

Normalmente, na etapa de codificação esparsa do DL, usa-se algum algoritmo de busca que limita a norma ℓ_0 de cada coluna da matriz de representação. Isto pode ser vantajoso em tarefas como compressão, contudo em tarefas como filtragem, melhores resultados podem ser atingidos permitindo que cada coluna tenha um número diferente de elementos não-nulos. O LARC é escolhido para efetuar a codificação esparsa por flexibilizar a esparsidade controlando a norma ℓ_1 da matriz de representação, além de viabilizar o uso de um único limiar de coerência para o treinamento tanto do dicionário de voz quanto dos dicionários de diferentes ruídos. A atualização do dicionário é realizada pelo AK-SVD. Por fim, os dicionários são concatenados formando o dicionário conjunto $\boldsymbol{D} = \begin{bmatrix} \boldsymbol{D}_s & \boldsymbol{D}_i \end{bmatrix}$.

Na etapa de melhoramento, a STFT do sinal a ser melhorado é calculada, a matriz de fase do espectro complexo é guardada enquanto a matriz de magnitude $\boldsymbol{X}$ é codificada com $\boldsymbol{D}$.

Figura 7 – Diagrama de blocos do algoritmo baseado em dicionários treinados.



Fonte: Elaborado pelo próprio autor.

O espectro de magnitude do sinal ruidoso é então expresso como uma combinação linear de átomos dos dicionários de voz e de ruído. A matriz de representação é dada pelo problema de otimização

$$\hat{\boldsymbol{C}} = \quad argmin_{\boldsymbol{C}} \|\boldsymbol{X} - \boldsymbol{D}\,\boldsymbol{C}\|_2^2 \quad \text{sujeito a} \quad \|\boldsymbol{c}_k\|_1 \leq \tau \quad \forall k \tag{60}$$

onde $\boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_s & \boldsymbol{C}_i \end{bmatrix}^T$ é a concatenação de $\boldsymbol{C}_s$, os coeficientes correspondentes à $\boldsymbol{D}_s$, e $\boldsymbol{C}_i$, coeficientes que correspondem à $\boldsymbol{D}_i$. A estimação da voz é obtida pela representação gerada pelo dicionário de voz, isto é,

$$\hat{\boldsymbol{S}} = \boldsymbol{D}_s \hat{\boldsymbol{C}}_s \tag{61}$$

e a estimação do ruído é produzida pela representação fornecida pelo dicionário de ruído:

$$\hat{\boldsymbol{I}} = \boldsymbol{D}_i \hat{\boldsymbol{C}}_i \tag{62}$$

As estimações $\hat{\boldsymbol{S}}$ e $\hat{\boldsymbol{I}}$ podem ser utilizadas como estimações finais tanto da voz pura quanto do ruído, contudo a soma dos dois espectros de magnitude podem não resultar exatamente em

$\boldsymbol{X}$, ou seja, a codificação esparsa fornecida pelo dicionário conjunto leva a aproximação

$$\boldsymbol{X} \approx \hat{\boldsymbol{S}} + \hat{\boldsymbol{I}}. \tag{63}$$

Se a contribuição do ruído no sinal ruídoso é pequena, a soma matricial da Equação (63) deve ser igual a $\boldsymbol{X}$. Para zerar o erro de aproximação, as estimações iniciais são usadas para construir a máscara definida na equação (64).

$$\boldsymbol{H} = \frac{(\hat{\boldsymbol{S}})^p}{(\hat{\boldsymbol{S}})^p + (\hat{\boldsymbol{I}})^p} \tag{64}$$

onde $p > 0$, $(\bullet)^p$ e a divisão são operações elemento a elemento entre matrizes. A matriz $\boldsymbol{H}$ assume diferentes valores de acordo com a escolha de p. Quando a máscara é calculada com $p = 2$, $\boldsymbol{H}$ torna-se um filtro de Wiener. Assumindo que a máscara seja construída como um filtro de Wiener, as estimativas geradas pelo filtro são dadas por:

$$\tilde{\boldsymbol{S}} = \boldsymbol{H} \otimes \boldsymbol{X} \tag{65}$$

sendo $\tilde{\boldsymbol{S}}$ a estimativa final da voz e $\otimes$ é a multiplicação elemento a elemento (GRAIS; ERDOGAN, 2011). Com a estimação final do espectro de maginitude da voz, o sinal de voz melhorado $\tilde{\boldsymbol{s}}(n)$ é dado pela STFT inversa da matriz complexa formada pela magnitude $\tilde{\boldsymbol{S}}$ e pela fase da voz ruidosa.

4.2 Garantias

Duas características do DL influenciam diretamente o desempenho do GDL: a coerência do dicionário com a classe de sinais de treinamento e a coerência mútua entre os átomos do dicionário. Segundo Donoho e Huo (2001) a coerência mútua $\mu(\boldsymbol{D})$ é estabelecida como

$$\mu(\boldsymbol{D}) = \max_{i \neq j} |\langle \boldsymbol{d}_i, \boldsymbol{d}_j \rangle|. \tag{66}$$

A coerência mútua mensura quão semelhantes dois átomos são. Dicionários ortogonais são altamente incoerentes, com $\mu(\boldsymbol{D}) = 0$, já dicionários sobrecompletos possuem obrigatóriamente coerência mútua maior que 0, e no pior cenário $\mu(\boldsymbol{D})$ se aproxima de 1, quando dois átomos são aproximadamente colineares. Um dicionário $\boldsymbol{D} \in \mathbb{R}^{M \times K}$ que possui coerência mútua mínima é conhecido como *Grassmannian frame* e este mínimo depende M e K, que são o número de linhas e de colunas de $\boldsymbol{D}$, respectivamente. O limite mínimo da coerência de um *Grassmannian frame* é imposto pela inequação

$$\mu(\boldsymbol{D}) \geq \sqrt{\frac{K - M}{M(K-1)}}. \tag{67}$$

Na Equação (67), a expressão a direita é conhecida como limite de Welch (WELCH, 1974). A igualdade na equação (68) só existe para uma particular classe de *Grassmannian frames*, os *equiangular tight frames*. Eles podem ser compreendidos como uma generalização dos dicionários ortogonais, por apresentar a melhor garantia de recuperação, mas possuem flexibilidade limitada para se adaptar a certas classes de sinais (SUSTIK *et al.*, 2007).

Segundo Donoho e Elad (2003), a condição de recuperação exata (ERC, do inglês *Exact Recovery Condition*) garante que a solução encontrada é a mais esparsa possível se

$$\|\boldsymbol{c}\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\boldsymbol{D})}\right). \tag{68}$$

Quando um dicionário é treinado para representar amostras semelhantes, a densidade de átomos é maior em certas regiões. Consequentemente alguns átomos são semelhantes entre si, aumentando a coerência mútua do dicionário. Porém, se a coerência mútua é elevada, as propriedades de recuperação, de acordo com a equação (67), só são garantidas para representações muito esparsas, caso contrário não há garantia teórica de recuperação. Portanto os dicionários treinados possuem uma melhor capacidade de representação por se adaptarem a distribuição das amostras de treinamento, mas não garantem que o potencial desses dicionários possam ser de fato aproveitados. Resultados empíricos mostram que desempenhos melhores que os teóricos são obtidos com os algoritmos de codificação esparsa, embora o desempenho seja comprometido quando o dicionário possui propriedades ruins (DUMITRESCU; IROFTI, 2018).

A representação esparsa de um segmento de voz degradado por ruído coerente é feita usando o dicionário conjunto formado pela união do dicionário de voz e do dicionário de ruído. O melhoramento da voz é dado pela separação das fontes através da representação das componentes de voz pelo dicionário de voz e das componentes do ruído pelo dicionário de ruído. Se tanto a voz quanto o ruído forem codificados apenas por átomos de seus respectivos dicionários, não haverá confusão entre as fontes. Para garantir que voz pura seja descrita como $\hat{\boldsymbol{s}} = \boldsymbol{D}_s\hat{\boldsymbol{c}}_s$ e o ruído seja decomposto como $\hat{\boldsymbol{i}} = \boldsymbol{D}_i\hat{\boldsymbol{c}}_i$, a seguinte condição de reconstrução exata deve ser satisfeita:

$$\|\hat{\boldsymbol{c}}_s\|_0 + \|\hat{\boldsymbol{c}}_i\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\boldsymbol{D}_s, \boldsymbol{D}_i)}\right) \tag{69}$$

A coerência mútua entre os dicionários de voz e ruído, denotada por $\mu(\boldsymbol{D}_s, \boldsymbol{D}_i)$, é dada por

$$\mu(\boldsymbol{D}_s, \boldsymbol{D}_i) = \max_{1\leq j\leq K_s, 1\leq k\leq K_i} |\langle \boldsymbol{d}_{s,j}, \boldsymbol{d}_{i,k}\rangle| \tag{70}$$

onde K_s e K_i são os números de átomos dos dicionários de voz e de ruído, respectivamente, e $\boldsymbol{d}_{s,j}$ é o j-ésimo átomo do dicionário de voz enquanto $\boldsymbol{d}_{i,k}$ é o k-ésimo átomo do dicionário de ruído (DONOHO; HUO, 2001).

Para ruídos com estruturas semelhantes à estrutura da voz, a condição de reconstrução na Equação (69) não pode ser satisfeita. Se o vetor de representação $\hat{\boldsymbol{c}} = \begin{bmatrix} \hat{\boldsymbol{c}}_s & \hat{\boldsymbol{c}}_i \end{bmatrix}^T$ for muito esparso, a participação do dicionário de ruído na construção da voz e a contribuição do dicionário de voz na representação do ruído tornam-se improváveis, considerando um $\mu(\boldsymbol{D}_s, \boldsymbol{D}_i)$ suficientemente pequeno, mas o erro $\|\boldsymbol{x} - \boldsymbol{D}\hat{\boldsymbol{c}}\|_2^2$ torna-se grande devido a liberdade limitada de $\hat{\boldsymbol{c}}$ em representar $\boldsymbol{x}$. A estimação da voz resultante sofre uma distorção de fonte, pelo fato da representação correspondente a voz pura ser muito esparsa. Se a norma ℓ_0 de $\hat{\boldsymbol{c}}$ for aumentada, o erro de aproximação $\|\boldsymbol{x} - \boldsymbol{D}\hat{\boldsymbol{c}}\|_2^2$ tende a diminuir, porém a confusão de fonte torna-se mais provável como é observado na Equação (69), resultando em partes da voz sendo representadas pelo dicionário de ruído e componentes do ruído sendo representadas pelo dicionário de voz, e este fenômeno é diretamente refletido quando a representação $\boldsymbol{D}\hat{\boldsymbol{c}}$ é separada em $\boldsymbol{D}_s\hat{\boldsymbol{c}}_s$ e $\boldsymbol{D}_i\hat{\boldsymbol{c}}_i$. A esparsidade de $\hat{\boldsymbol{c}}$ pode ser regulada afim de minimizar tanto a distorção de fonte quanto a confusão de fonte, variando o limiar de coerência do LARC. Um limiar de coerência alto leva a soluções mais esparsas. Já soluções mais densas são obtidas com o uso de um limiar de coerência mais baixo.

4.3 Avaliação

O desempenho do GDL é avaliado e comparado com os métodos de subtração espectral (BOLL, 1979), filtro de Wiener (SCALART *et al.*, 1996) e MMSE-STSA (EPHRAIM; MALAH, 1984), bem estabelecidos e eficientes ná área de melhoramento de voz. A eficiência do método é examinada tanto para o cenário dependente de locutor, onde tanto as amostras de treinamento quanto de avaliação são fornecidas pelo mesmo locutor, quanto para o cenário independente de locutor, com as amostras de treinamento fornecidas por locutores diferentes daqueles que fornecem as amostras de avaliação.

4.3.1 Dados

Os sinais de voz foram obtidos da base de dados GRID *audio-visual corpus*, que é composta por 34 locutores, 18 do gênero masculino e 16 do gênero feminino. Cada locutor possui 1000 falas com estrutura gramatical simples, por exemplo "Place green at B 4 now"(COOKE *et al.*, 2006). Foram selecionados 9 locutores do gênero masculino e 9 locutores do gênero feminino para os experimentos.

Ruídos de diferentes ambientes relevantes foram incluídos nos testes. Os ruídos estruturados utilizados pertencem ao banco de dados NOISEX-92 (VARGA, 1992). Dos ruídos disponíveis foram escolhidos o ruído de fábrica (fab), interior de um veículo Volvo (vol), interior de uma aeronave Buccaneer (buc), interior da sala de máquinas de um Destroyer (des) e ruído proveniente de um canal de alta frequência (hfc). Como ruído não estruturado, ruído branco gaussiano

sintético (whi). Os sinais de voz foram subamostrados de 25kHz para 8kHz, enquanto os ruídos estruturados foram subamostrados de 20kHz para 8kHz.

Os dados utilizados foram separados em conjuntos de treinamento e de avaliação. Para o cenário independente de locutor, o conjunto de amostras de treinamento foi composto por 10 amostras selecionadas de cada um dos 6 locutores masculinos e 6 locutores femininos escolhidos, totalizando 120 amostras. No cenário dependente de locutor 3 locutores masculinos e 3 locutores femininos foram selecionados e 120 amostras de cada um foram utilizadas para treinar um dicionário para cada locutor escolhido, e o conjunto de avaliação foi formado por 4 amostras de cada locutor, totalizando 24 amostras. O mesmo conjunto é utilizado para a avaliação no cenário independente de locutor, possibilitando a comparação direta dos dois cenários. Todos os conjuntos são disjuntos e foram formados por amostras selecionadas de forma aleatória. A partir dos dados dos ruídos, foram criados dois conjuntos distintos, um conjunto de treinamento e conjunto de avaliação, para cada um dos ruídos selecionados. Um dicionário foi treinado para cada ruído escolhido com 60 segundos consecutivos de dados, e cerca de 40 segundos de cada ruído foi separado para a etapa de avaliação.

4.3.2 Medidas objetivas de desempenho

Métodos de melhoramento de voz buscam melhorar tanto a qualidade quanto a inteligibilidade da voz. A qualidade da voz está associada a como a voz é produzida e a qualidade alta indica que a voz é boa e agradável para ouvir. Já a inteligibilidade da voz é vinculada com a compreensão das palavras faladas e uma voz com alta inteligibilidade garante que o erro no entendimento das palavras contidas seja baixo. A qualidade da voz é subjetiva, enquanto que a inteligibilidade é objetiva, já que o número de palavras não compreendidas pode ser contabilizado.

A performance de algoritmos de melhoramento de voz pode ser medida de forma objetiva ou subjetiva. Medições subjetivas são realizadas por ouvintes humanos e são uma forma confiável de verificar a qualidade e inteligibilidade do sinal melhorado. Contudo, medições subjetivas são demoradas, caras e demandam ouvintes treinados. Medições objetivas podem ser usadas como alternativa às avaliações subjetivas, pois fornecem modelos matemáticos capazes de medir algum aspecto perceptivo do ouvido humano (LOIZOU, 2013).

As estimações de qualidade e inteligibilidade dos resultados gerados foram medidas com a SNR segmentada no tempo (SSNR), a SNR segmentada ponderada na frequência (fwSSNR), a PESQ (do inglês *Perceptual Evaluation of Speech Quality*), e a STOI (do inglês *short-time objective intelligibility*).

A SNR segmentada é uma das medições objetivas mais simples entre os métodos usados para avaliar os algoritmos de melhoramento de voz, e é dada pela média da SNRs calculadas

para cada quadro no domínio do tempo, conforme a Equação (71):

$$\text{SSNR} = \frac{10}{M}\sum_{m=0}^{M-1} log_{10}\frac{\sum_{Nm}^{Nm+N-1} \boldsymbol{s}^2(n)}{\sum_{Nm}^{Nm+N-1}(\boldsymbol{s}(n) - \tilde{\boldsymbol{s}}(n))^2} \tag{71}$$

onde N é o comprimento do quadro, geralmente definido como 20ms, e M é o número de quadros do sinal.

A SNR também pode ser avaliada no domínio da frequência. Dados os espectros de potência da voz pura e da voz melhorada, a SNR segmentada ponderada na frequência é definida por

$$\text{fwSSNR} = \frac{10}{M}\sum_{m=0}^{M-1} \frac{\sum_{j=1}^{K} W_j \log_{10}[|S^2(j,m)|/(|S(j,m)| - |\tilde{S}(j,m)|)^2]}{\sum_{j=1}^{K} W_j} \tag{72}$$

onde W_j é o peso aplicado na j-ésima banda de frequência, K é o número de índices de frequência, M é o número de índices no tempo, $|S(j,m)|$ e $|\tilde{S}(j,m)|$ são as amplitude da voz pura e da voz melhorada na j-ésima frequência e m-ésimo segmento janelado, respectivamente. Diferente da SNR segmentada no tempo, a fwSSNR possibilita que diferentes pesos sejam aplicados a cada banda de frequência dos espectros. Os pesos utilizados foram sugeridos por Quackenbush, Barnwell e Clements (1988) e baseiam-se em estudos sobre índices de articulação realizados por Kryter (1962).

A PESQ foi desenvolvida para realizar testes subjetivos frequentemente usados em telecomunicações e avaliar a qualidade de voz em uma grande diversidade de condições. As medições da PESQ são mais elaboradas e são detalhadas por Rix *et al.* (2001). Já a STOI é uma medida de inteligibilidade que se mostra uma boa alternativa para avaliar algoritmos baseados no processamento de coeficientes tempo-frequência, como redução de ruído e separação de fontes (TAAL *et al.*, 2010).

4.3.3 Treinamento dos dicionários

Para o treinamento dos dicionários, o LARC foi utilizado na etapa de codificação esparsa enquanto o AK-SVD foi escolhido para a atualização dos átomos. A escolha do LARC foi motivada pela falta de conhecimento prévio a respeito das propriedades de cada ruído. Controlar a esparsidade de forma indireta pelo limiar de coerência mostra-se eficiente por controlar o número de coeficientes de acordo com a coerência entre as amostras de sinais de treinamento e o dicionário. Desta forma, amostras incoerentes são ignoradas, o que não ocorreria se o mesmo número de elementos não nulos fosse imposto para todas as amostras de treinamento. Além disso o limiar de coerência não é sensível a distribuição de energia das amostras, permitindo que apenas um limiar seja utilizado para diferentes classes de sinais. O dicionário inicial utilizado foi composto por números aleatórios uniformemente distribuídos no intervalo (0,1) e suas colunas foram normalizadas.

4.3.4 Melhoramento de voz

O desempenho dos métodos escolhidos foi avaliado usando grupos de sinais ruidosos gerados pela mistura das amostras de avaliação somadas aos ruídos selecionados, com SNR de entrada de 0dB, 5dB e 10dB.

A dimensão dos sinais a serem representados de forma esparsa no domínio da STFT depende da largura da janela adotada. A capacidade do GDL em estimar a voz a partir da mistura só é viável se o dicionário de voz for coerente com a voz e, ao mesmo tempo, incoerente com o ruído. A coerência entre a voz e seu dicionário é maior quando janelas de menor largura são utilizadas na STFT, ou seja, a coerência é aumentada em espaços dimensionais menores. Isto leva a uma menor distorção de fonte, porém a estrutura da voz torna-se mais semelhante ao do ruído em dimensões menores, o que aumenta a confusão de fonte. Já dimensões maiores diminuem as semelhanças entre a estrutura da voz e do ruído, minimizando a confusão de fonte, mas a maior complexidade dos sinais em espaços dimensionais maiores dificulta a representação via DL. Empiricamente os melhores resultados foram alcançados com o uso da janela de Hanning de 64ms, isto é, com sinais com dimensão de ordem 257.

O número de átomos do dicionário inicial influencia a coerência entre o dicionário treinado e sua respectiva classe de sinais. A coerência aumenta se o número de átomos for grande. Contudo, dicionários grandes exigem maior tempo de treinamento. Tanto os dicionários de voz quanto os dicionários de ruído foram treinados com 512 átomos, mostrando-se um bom balanço entre erro de aproximação e custo computacional. O limiar de coerência no treinamento foi fixado em 0.2. Na etapa de melhoramento o limiar foi fixado em 0.15 e 0.1 nos cenários dependente e independente de locutor, respectivamente. Todos os parâmetros foram escolhidos empiricamente.

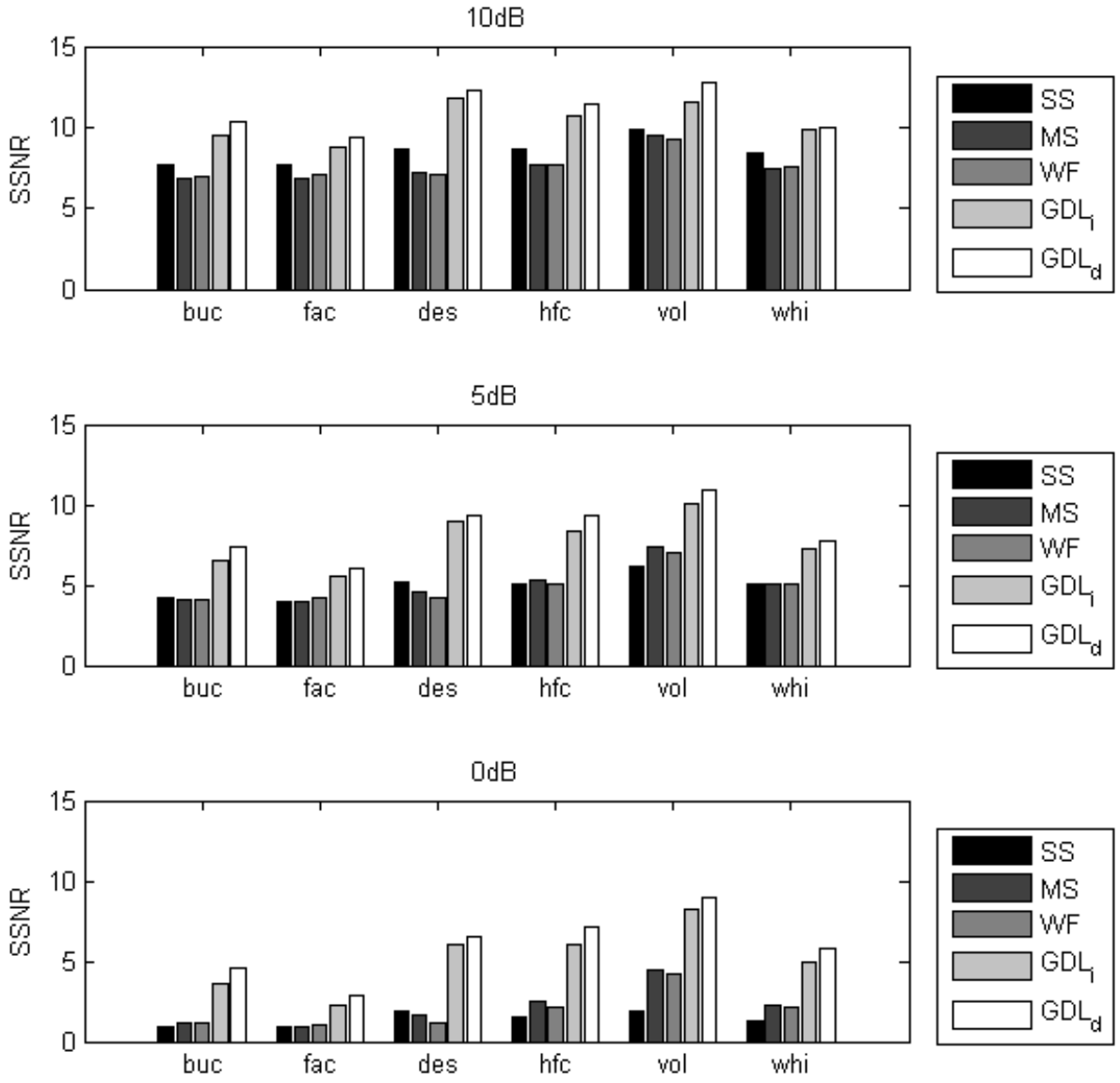
4.3.5 Resultados e discussões

Nesta seção são apresentados os resultados das avaliações de desempenho dos métodos selecionados. A Subtração espectral é denotada por SS, a filtragem baseada em filtro de Wiener é denotada por WF, MS denota a filtragem por estimadores MMSE de espectro de magnitude, GDL_i denota o GDL no cenário independente de locutor enquanto GDL_d denota o mesmo método no cenário dependente de locutor.

Na figura 8 são apresentas as médias de SSNR considerando os 6 ruídos escolhidos em 3 níveis de SNR de entrada distintos. Tanto o GDL_i quanto o GDL_d alcançaram resultados superiores em SNR segmentada em relação aos outros algoritmos em todos os diferentes cenários de ruídos. Por exemplo, com SNR de entrada de 0dB, o GDL_i obteve ganho médio de 4.58, 3.81 e 3.99, respectivamente, sobre o SS, MS e WF. O GDL_i alcançou ganho médio de 3.80, 3.03 e 3.21 sobre o SS, MS e WF, respectivamente. Melhores resultados foram registrados em cenários

com a presença de ruídos muito estruturados, como o ruído de Destroyer. O desempenho do GDL_i foi sempre inferior ao do GDL_d, o que se justifica pelos dicionários utilizados no cenário dependente de locutor serem mais coerentes com as amostras de avaliação.

Figura 8 – Médias de SSNR resultante do processamento do conjunto de avaliação para diferentes métodos de melhoramento de voz e diferentes SNR de entrada.

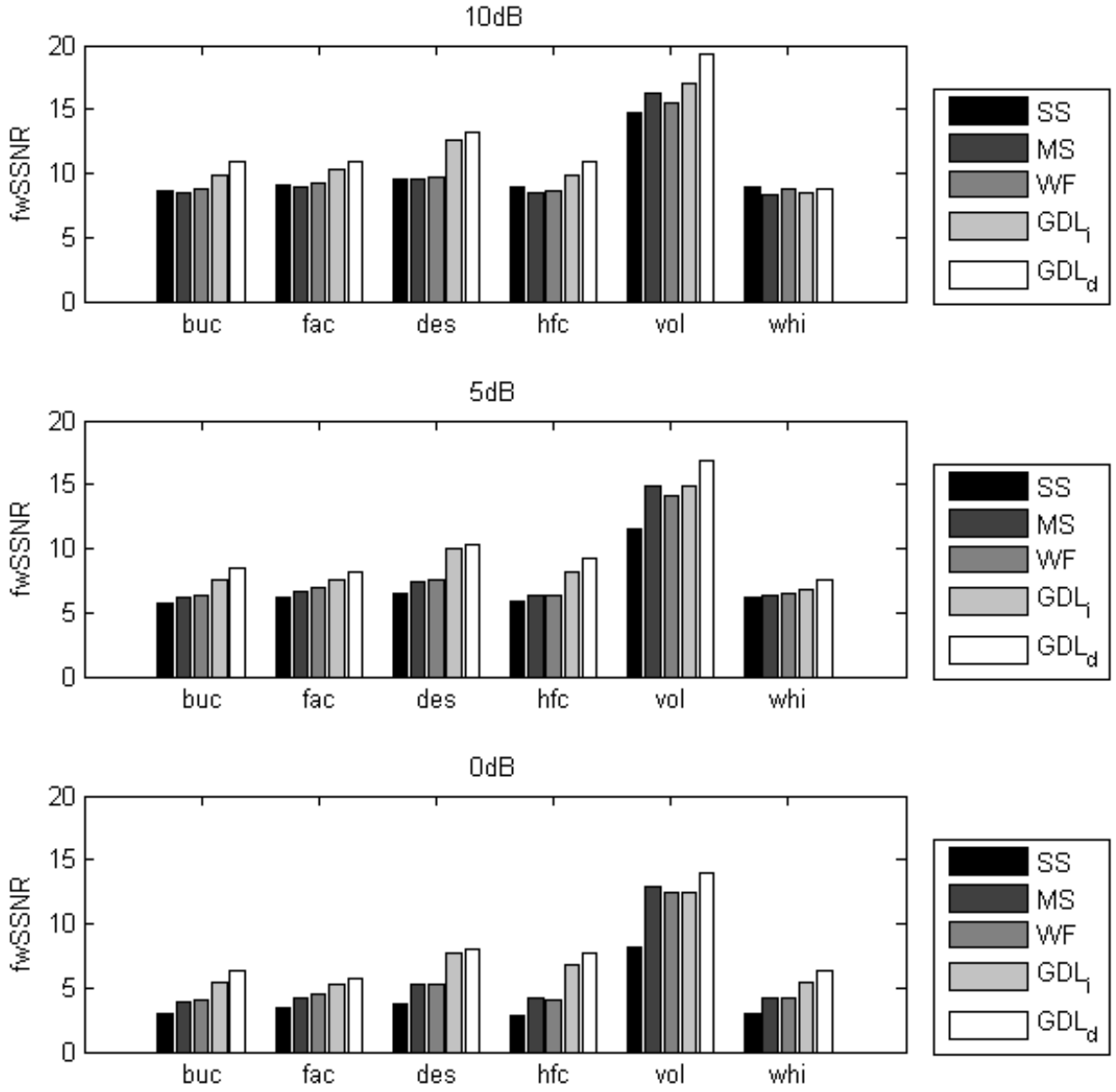


Fonte: Elaborado pelo próprio autor.

O desempenho mensurado pela fwSSNR é mostrado na Figura 9. O GDL, tanto no cenário dependente quanto no independente de locutor, apresentou desempenho superior aos outros métodos em 17 dos 18 cenários ruidosos. Contudo os ganhos médios do GDL_i e GDL_d foram positivos em todos os níveis de SNR. Por exemplo, com SNR de entrada de 5dB os ganhos médios do GDL_i foram de 2.17, 1.19 e 1.16 sobre o SS, MS e WF, respectivamente, enquanto

o GDL_d obteve ganhos de 3.13, 2.15 e 2.12. Os ganhos médios foram mais expressivos com a SNR de entrada de 0dB, com o GDL_i alcançando ganhos médios de 3.14, 1.41 e 1.44 sobre o SS, WF e MS, e 3.97, 2.24 e 2.27 obtidos pelo GDL_d.

Figura 9 – Médias de fwSSNR resultante do processamento do conjunto de avaliação para diferentes métodos de melhoramento de voz e diferentes SNR de entrada.



Fonte: Elaborado pelo próprio autor.

A Tabela 1 contém os resultados da avaliação de desempenho medidos com a PESQ. O GDL alcançou desempenho superior em 14 dos 18 cenários de ruídos implementados. Ambas as versões do GDL obtiveram ganhos médios superiores aos métodos restantes, com maior destaque no cenário de 0dB, com ganhos médios de 0.41, 0.15 e 0.17 obtidos pelo GDL_i sobre o SS, MS e WF, respectivamente e ganhos de 0.46, 0.20 e 0.22 pelo GDL_d. Além disso, como

mostrado na Tabela 2, tanto o GDL_i quanto o GDL_d apresentaram inteligibilidade superior, indicado pelos resultados medidos pela STOI, aos outros algoritmos avaliados. No cenário com SNR de 0dB, por exemplo, o GDL_i atingiu ganhos médios de 0.12, 0.03 e 0.01 sobre o SS, MS e WF respectivamente, enquanto o GDL_d apresentou ganhos médios 0.01 superiores aos ganhos médios obtidos pelo GDL_i.

Em geral, o GDL apresentou resultados iguais ou, na maior parte das médias obtidas, superiores aos métodos clássicos, principalmente nos cenários com menor SNR de entrada.

Na figura 10 é possível comparar os espectrogramas de um sinal de voz processado pelos métodos avaliados. O sinal foi misturado com ruído de fábrica com SNR de entrada de 0dB. É possível notar que a subtração espectral é o método que preservou menos a estrutura espectral da voz no domínio tempo-frequência. O estimador MMSE-STSA preservou conteúdo espectral nas frequências baixas. Contudo, em frequências altas houve maior deterioração da voz. O filtro de Wiener degradou menos a estrutura da voz que a subtração espectral e o estimador MMSE-STSA, contudo, o método criou fortes artefatos, principalmente na faixa de frequências entre 0Hz e 2000Hz. O GDL_i degradou menos informação espectral da voz e gerou menos artefatos que os métodos clássicos, já o GDL_d, por usar um dicionário mais coerente com a estrutura espectral do sinal selecionado, gerou menos degradação e menos artefatos que o GDL_i.

Tabela 1 – Médias de PESQ resultantes do processamento do conjunto de avaliação para diferentes métodos de melhoramento de voz e diferentes SNR de entrada.

SNR	Ruído	SS	MS	WF	GDL_i	GDL_d
10dB	buc	2.87	2.86	2.85	2.81	**2.93**
	fac	**3.02**	2.99	3.01	3.00	3.00
	des	3.08	2.94	2.89	3.27	**3.34**
	hfc	**2.87**	2.74	2.73	2.67	2.81
	vol	3.77	3.91	3.87	4.20	**4.26**
	whi	**2.89**	2.79	2.80	2.43	2.30
	média	3.08	3.04	3.03	3.06	**3.11**
5dB	buc	2.40	2.50	2.48	2.53	**2.65**
	fac	2.57	2.63	2.64	2.66	**2.68**
	des	2.65	2.63	2.59	2.99	**3.05**
	hfc	2.40	2.38	2.38	2.50	**2.66**
	vol	3.43	3.75	3.73	3.97	**3.99**
	whi	**2.44**	2.41	2.43	2.21	2.09
	média	2.65	2.72	2.71	2.81	**2.85**
0dB	buc	1.86	2.14	2.11	2.22	**2.32**
	fac	2.07	2.23	2.23	2.28	**2.31**
	des	2.20	2.33	2.29	2.70	**2.77**
	hfc	1.86	2.06	2.06	2.33	**2.50**
	vol	2.99	3.56	3.54	**3.69**	3.68
	whi	1.88	2.06	**2.07**	**2.07**	2.03
	média	2.14	2.40	2.38	2.55	**2.60**

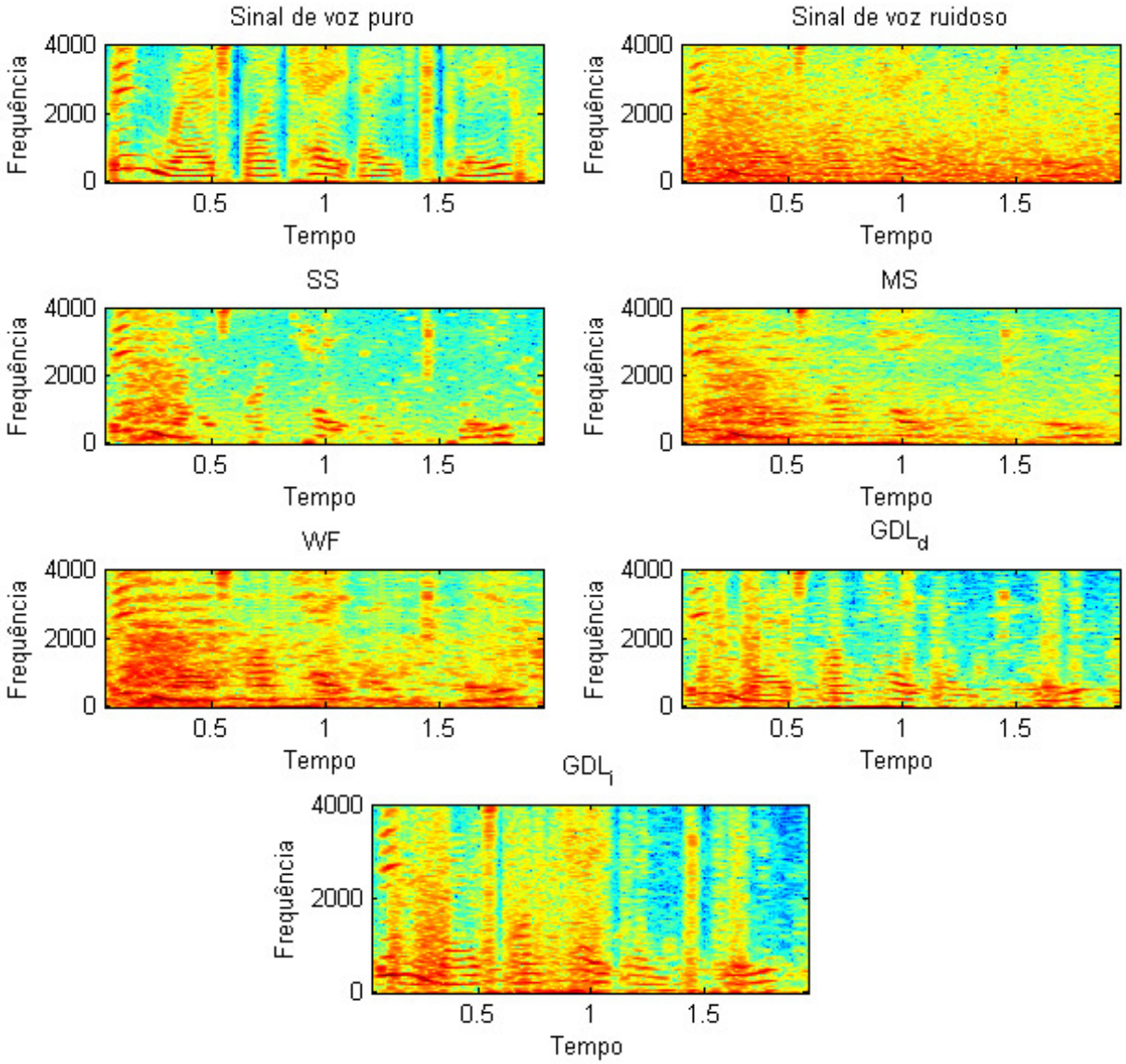
Fonte: Elaborado pelo próprio autor.

Tabela 2 – Médias de STOI resultantes do processamento do conjunto de avaliação para diferentes métodos de melhoramento de voz e diferentes SNR de entrada.

SNR	Ruído	SS	MS	WF	GDL_i	GDL_d
10dB	buc	0.87	0.86	0.87	0.87	**0.89**
	fac	0.89	0.87	0.89	0.89	**0.90**
	des	0.91	0.89	0.90	0.93	**0.94**
	hfc	0.87	0.86	0.87	0.86	**0.88**
	vol	0.97	0.95	0.96	0.98	**0.99**
	whi	0.86	0.85	**0.87**	0.78	0.79
	média	**0.90**	0.88	0.89	0.89	**0.90**
5dB	buc	0.74	0.77	0.79	0.79	**0.81**
	fac	0.79	0.79	**0.82**	0.80	**0.82**
	des	0.82	0.82	0.84	0.88	**0.89**
	hfc	0.76	0.79	0.81	0.81	**0.83**
	vol	0.94	0.95	0.95	0.97	**0.98**
	whi	0.75	0.77	**0.79**	0.73	0.73
	média	0.80	0.82	0.83	0.83	**0.84**
0dB	buc	0.56	0.66	0.67	0.69	**0.70**
	fac	0.61	0.67	**0.70**	0.69	**0.70**
	des	0.66	0.73	0.75	**0.81**	**0.81**
	hfc	0.59	0.71	0.72	0.75	**0.76**
	vol	0.85	0.93	0.94	0.95	**0.96**
	whi	0.56	0.68	**0.70**	0.67	0.66
	média	0.64	0.73	0.75	0.76	**0.77**

Fonte: Elaborado pelo próprio autor.

Figura 10 – Comparação dos espectrogramas gerados pelos métodos avaliados.



Fonte: Elaborado pelo próprio autor.

5 Conclusões

Neste trabalho se propôs explorar a utilização de representações esparsas para o melhoramento de voz em cenários com ruídos não-estacionários. Para isso foram apresentados os conceitos essenciais envolvendo representações esparsas, seus algoritmos de implementação e de treinamento de dicionários que possibilitam sua aplicação no melhoramento de sinais de voz.

Representar esparsamente um sinal significa aproximá-lo de uma combinação linear de um pequeno número de átomos que fazem parte de um dicionário sobrecompleto. Algoritmos como o OMP e o LASSO encontram soluções suficientemente esparsas em meio à um infinito número de vetores-solução possíveis.

Uma das aplicações desses algoritmos é o treinamento de dicionários sobrecompletos, que adaptam um dicionário inicial para representar uma determinada classe de sinais, modificando os átomos do dicionário inicial de acordo com as diferentes estruturas que caracterizam esta classe. Desta forma, dicionários treinados possuem a capacidade de decompor sinais estruturalmente complexos de forma mais eficaz que dicionários analíticos.

O *Generative Dictionary Learning* (GDL) é um método de melhoramento de voz baseado no treinamento de dicionários sobrecompletos usando amostras de voz limpa e de ruídos de diferentes ambientes. O sinal ruidoso é codificado com a concatenação dos dicionários treinados e a estimação da voz pura é dada pelos coeficientes associados ao dicionário de voz, enquanto que o ruído é estimado pelos coeficientes referentes ao dicionário de ruído. Dicionários analíticos são coerentes tanto com a voz quanto com o ruído, já dicionários de voz e de ruído treinados têm a coerência maximizada com suas respectivas classes de sinais, permitindo uma melhor separação mesmo em condições com baixa SNR de entrada. Controlar a esparsidade desta representação possibilita obter menores distorção de fonte e confusão de fonte, diminuindo a degradação e a presença de componentes ruidosos na estimação da voz. A utilização de um filtro tipo Wiener com as estimações de voz e ruído conduzem a uma estimação ótima da voz no sentido dos mínimos quadrados. Este método mostra-se em média superior aos métodos comparados, tanto em cenários com ruído estruturado quanto com ruído não-estruturado, como mostrado nos resultados.

Neste trabalho, por questões de limitação de máquina, não foi levado em conta o tempo de processamento de cada método de melhoramento de voz estudado. Porém, observações levaram a acreditar que o GDL demanda mais tempo que os demais métodos, devido a sua fase de treinamento. Além disso a etapa de melhoramento do GDL se mostrou mais lenta que os demais métodos, porém o tempo adicional durante o melhoramento não compromete a aplicabilidade do GDL. Isto o torna muito útil em aplicações com reconhecimento automático de voz, biometria de voz, robótica e outras.

O AK-SVD produz dicionários coerentes com a voz e com o ruído, o que permite separar diretamente o sinal ruidoso pela sua codificação esparsa com a concatenação dos dicionários treinados. Porém deve ser observado que o treinamento do dicionário de ruído é feito a partir dos dados fornecidos por um detector de atividade de voz ideal. Treinar dicionários com o uso de um VAD não-ideal fugiu do propósito deste trabalho, mas é algo possível de ser abordado em trabalhos futuros.

Referências bibliográficas

AMALDI, E.; KANN, V. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science*, Elsevier, v. 209, n. 1-2, p. 237–260, 1998.

BOLL, S. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on acoustics, speech, and signal processing*, IEEE, v. 27, n. 2, p. 113–120, 1979.

BRASSARD, G.; BRATLEY, P. *Algorithmics*: Theory & Practice. USA: Prentice-Hall, Inc., 1988.

BRUCKSTEIN, A. M.; DONOHO, D. L.; ELAD, M. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM review*, SIAM, v. 51, n. 1, p. 34–81, 2009.

CANDES, E. J.; ROMBERG, J. K.; TAO, T. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, Wiley Online Library, v. 59, n. 8, p. 1207–1223, 2006.

CAPPÉ, O. Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor. *IEEE transactions on Speech and Audio Processing*, IEEE, v. 2, n. 2, p. 345–349, 1994.

COOKE, M. *et al.* An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, ASA, v. 120, n. 5, p. 2421–2424, 2006.

DONOHO, D. L. For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, Wiley Online Library, v. 59, n. 6, p. 797–829, 2006.

DONOHO, D. L.; ELAD, M. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 100, n. 5, p. 2197–2202, 2003.

DONOHO, D. L.; HUO, X. Uncertainty principles and ideal atomic decomposition. *IEEE transactions on information theory*, v. 47, n. 7, p. 2845–2862, 2001.

DONOHO, D. L. *et al.* Compressed sensing. *IEEE Transactions on information theory*, Citeseer, v. 52, n. 4, p. 1289–1306, 2006.

DUMITRESCU, B.; IROFTI, P. *Dictionary learning algorithms and applications*. [S.l.]: Springer, 2018.

EFRON, B. *et al.* Least angle regression. *The Annals of statistics*, Institute of Mathematical Statistics, v. 32, n. 2, p. 407–499, 2004.

ELAD, M.; FIGUEIREDO, M. A.; MA, Y. On the role of sparse and redundant representations in image processing. *Proceedings of the IEEE*, IEEE, v. 98, n. 6, p. 972–982, 2010.

EPHRAIM, Y. Statistical-model-based speech enhancement systems. *Proceedings of the IEEE*, IEEE, v. 80, n. 10, p. 1526–1555, 1992.

EPHRAIM, Y.; MALAH, D. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on acoustics, speech, and signal processing*, IEEE, v. 32, n. 6, p. 1109–1121, 1984.

EPHRAIM, Y.; MALAH, D. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE transactions on acoustics, speech, and signal processing*, IEEE, v. 33, n. 2, p. 443–445, 1985.

GERKMANN, T.; KRAWCZYK-BECKER, M.; ROUX, J. L. Phase processing for single-channel speech enhancement: History and recent advances. *IEEE signal processing Magazine*, IEEE, v. 32, n. 2, p. 55–66, 2015.

GERKMANN, T.; KRAWCZYK, M.; REHR, R. Phase estimation in speech enhancement—unimportant, important, or impossible? In: CONVENTION OF ELECTRICAL AND ELECTRONICS ENGINEERS IN ISRAEL, 27th, 2012, [S.l.]. Proceedings [...][S.I.: s.n.], 2012. p. 1-5.

GRAIS, E. M.; ERDOGAN, H. Single channel speech music separation using nonnegative matrix factorization and spectral masks. In: INTERNATIONAL CONFERENCE ON DIGITAL SIGNAL PROCESSING, 17th, 2011, [S.I.]. Proceedings [...][S.l.: s.n.], 2011. p.1-6.

HENDRIKS, R. C.; GERKMANN, T.; JENSEN, J. Dft-domain based single-microphone noise reduction for speech enhancement: A survey of the state of the art. *Synthesis Lectures on Speech and Audio Processing*, Morgan & Claypool Publishers, v. 9, n. 1, p. 1–80, 2013.

HU, Y.; LOIZOU, P. C. A generalized subspace approach for enhancing speech corrupted by colored noise. *IEEE Transactions on Speech and Audio Processing*, IEEE, v. 11, n. 4, p. 334–341, 2003.

JABLOUN, F.; CHAMPAGNE, B. Incorporating the human hearing properties in the signal subspace approach for speech enhancement. *IEEE Transactions on Speech and Audio Processing*, IEEE, v. 11, n. 6, p. 700–708, 2003.

KRYTER, K. D. Methods for the calculation and use of the articulation index. *The Journal of the Acoustical Society of America*, Acoustical Society of America, v. 34, n. 11, p. 1689–1697, 1962.

LIM, J. S.; OPPENHEIM, A. V. Enhancement and bandwidth compression of noisy speech. *Proceedings of the IEEE*, IEEE, v. 67, n. 12, p. 1586–1604, 1979.

LIU, L. *et al.* Realistic action recognition via sparsely-constructed gaussian processes. *Pattern Recognition*, Elsevier, v. 47, n. 12, p. 3819–3827, 2014.

LOIZOU, P. C. *Speech enhancement: theory and practice*. [S.l.]: CRC press, 2013.

LU, Y.; LOIZOU, P. C. A geometric approach to spectral subtraction. *Speech communication*, Elsevier, v. 50, n. 6, p. 453–466, 2008.

MALLAT, S. G.; ZHANG, Z. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on signal processing*, v. 41, n. 12, p. 3397–3415, 1993.

PAPADIMITRIOU, C. H. *Computational complexity*. Reading: Addison-Wesley, 1994.

PATEL, V. M.; CHELLAPPA, R. Sparse representations, compressive sensing and dictionaries for pattern recognition. In: THE FIRST ASIAN CONFERENCE ON PATTERN RECOGNITION, 1st, 2011. Proceedings [...][S.l.: s.n.], 2011. p. 325-329.

QUACKENBUSH, S. R.; BARNWELL, T. P.; CLEMENTS, M. A. *Objective measures of speech quality*. [S.l.]: Prentice Hall, 1988.

RIX, A. W. *et al.* Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In: IEEE. *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*. [S.l.], 2001. v. 2, p. 749–752.

RUBINSTEIN, R.; BRUCKSTEIN, A. M.; ELAD, M. Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, IEEE, v. 98, n. 6, p. 1045–1057, 2010.

RUBINSTEIN, R.; ZIBULEVSKY, M.; ELAD, M. *Efficient implementation of the K-SVD algorithm using batch orthogonal matching pursuit*. [S.l.: s.n.], 2008.

SCALART, P. *et al.* Speech enhancement based on a priori signal to noise estimation. In: IEEE. INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING CONFERENCE PROCEEDINGS, 21th, 1996. Proceedings [...][S.l.: s.n.], 1996. v. 2, p. 629–632.

SIGG, C. D.; DIKK, T.; BUHMANN, J. M. Speech enhancement with sparse coding in learned dictionaries. In: IEEE. INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, 35th, 2010. Proceedings [...][S.l.: s.n.], 2010. p. 4758–4761.

SIGG, C. D.; DIKK, T.; BUHMANN, J. M. Speech enhancement using generative dictionary learning. *IEEE Transactions on Audio, Speech, and Language Processing*, IEEE, v. 20, n. 6, p. 1698–1712, 2012.

SUSTIK, M. A. *et al.* On the existence of equiangular tight frames. *Linear Algebra and its applications*, Elsevier, v. 426, n. 2-3, p. 619–635, 2007.

TAAL, C. H. *et al.* A short-time objective intelligibility measure for time-frequency weighted noisy speech. In: INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, 35th, 2010. Proceedings [...][S.l.: s.n.], 2010. p. 4214–4217.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 58, n. 1, p. 267–288, 1996.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Wiley Online Library, v. 73, n. 3, p. 273–282, 2011.

VARGA, A. The noisex-92 study on the effect of additive noise on automatic speech recognition. *ical Report, DRA Speech Research Unit*, 1992.

WELCH, L. Lower bounds on the maximum cross correlation of signals (corresp.). *IEEE Transactions on Information theory*, IEEE, v. 20, n. 3, p. 397–399, 1974.

WRIGHT, S. J. Coordinate descent algorithms. *Mathematical Programming*, Springer, v. 151, n. 1, p. 3–34, 2015.

YUAN, Y.; LU, X.; LI, X. Learning hash functions using sparse reconstruction. In: INTERNATIONAL CONFERENCE ON INTERNET MULTIMEDIA COMPUTING AND SERVICE, 6th, 2014. Proceedings [...][S.l.: s.n.], 2014. p. 14–18.

ZHANG, Z. *et al.* A survey of sparse representation: algorithms and applications. *IEEE access*, IEEE, v. 3, p. 490–530, 2015.