Todo list

Politécnica
UFRJ

ESPECTRO DE MODULAÇÃO APLICADO AO PROCESSAMENTO DE FALA

Miguel Fernandes de Sousa

Projeto de Graduação apresentado ao curso de Engenharia Eletrônica e de Computação da Escola Politécnica, Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Engenheiro.

Orientadores: Luiz Wagner Pereira Biscainho
Nome do Segundo Orientador
Sobrenome
Nome do Terceiro Orientador
Sobrenome

Rio de Janeiro
Julho de 2022

ESPECTRO DE MODULAÇÃO APLICADO AO PROCESSAMENTO DE FALA

Miguel Fernandes de Sousa

PROJETO DE GRADUAÇÃO SUBMETIDO AO CORPO DOCENTE DO CURSO DE ENGENHARIA ELETRÔNICA E DE COMPUTAÇÃO DA ESCOLA POLITÉCNICA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE ENGENHEIRO ELETRÔNICO E DE COMPUTAÇÃO.

Orientadores: Luiz Wagner Pereira Biscainho
Nome do Segundo Orientador Sobrenome
Nome do Terceiro Orientador Sobrenome

Aprovada por: Prof. Nome do Primeiro Examinador Sobrenome
Prof. Nome do Segundo Examinador Sobrenome
Prof. Nome do Terceiro Examinador Sobrenome
Prof. Nome do Quarto Examinador Sobrenome
Prof. Nome do Quinto Examinador Sobrenome

RIO DE JANEIRO, RJ – BRASIL
JULHO DE 2022

Fernandes de Sousa, Miguel

Espectro de Modulação Aplicado ao Processamento de Fala/Miguel Fernandes de Sousa. – Rio de Janeiro: UFRJ/COPPE, 2022.

XIV, 28 p.: il.; 29, 7cm.

Orientadores: Luiz Wagner Pereira Biscainho
Nome do Segundo Orientador Sobrenome
Nome do Terceiro Orientador Sobrenome

Projeto de Graduação – UFRJ/Curso de Engenharia Eletrônica e de Computação, 2022.

Referências Bibliográficas: p. 25 – 28.

1. Espectrograma de Modulação. 2. Filtragem de Modulação. 3. *Speech Enhancement*. I. Pereira Biscainho, Luiz Wagner *et al.* II. Universidade Federal do Rio de Janeiro, Escola Politécnica, Curso de Engenharia Eletrônica e de Computação. III. Título.

"Potiusque sero quam
nunquam."
Lívio (59 a.C. - 17 d.C.)

Agradecimentos

Gostaria de agradecer a todos.

Resumo do Projeto de Graduação apresentado à POLI/UFRJ como parte dos requisitos necessários para a obtenção do grau de Engenheiro.

ESPECTRO DE MODULAÇÃO APLICADO AO PROCESSAMENTO DE FALA

Miguel Fernandes de Sousa

Julho/2022

Orientadores: Luiz Wagner Pereira Biscainho
Nome do Segundo Orientador Sobrenome
Nome do Terceiro Orientador Sobrenome

Programa: Engenharia Eletrônica e de Computação

Apresenta-se, nesta tese,[1] ...

Abstract of Undergraduate Project presented to POLI/UFRJ as a partial fulfillment of the requirements for the degree of Engineer.

MODULATION SPECTRUM APPLIED OVER SPEECH PROCESSING

Miguel Fernandes de Sousa

July/2022

Advisors: Luiz Wagner Pereira Biscainho
Nome do Segundo Orientador Sobrenome
Nome do Terceiro Orientador Sobrenome

Department: Electronic and Computer Engineering

In this work, we present ...

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Símbolos

A_c	amplitude da portadora, p. 4
$M(f)$	moduladora no domínio da frequência, p. 5
$S(f)$	sinal modulado no domínio da frequência, p. 5
$X(t, f)$	sinal no domínio acústico, p. 7
$X(t, f)$	sinal no domínio acústico, p. 7
$*$	operador para convolução, p. 10
$\delta(t)$	impulso unitário, p. 5
$\hat{x}(t)$	sinal em quadratura, p. 10
$\mathbb{C}$	conjunto dos números complexos, p. 7
$\mathbb{R}$	conjunto dos números reais, p. 7
$\mathbb{Z}$	Conjunto dos números inteiros, p. 12
$\mathcal{H}\{\cdot\}$	operador para transformada de Hilbert, p. 10
$\mathcal{X}(\eta, f, t)$	sinal no domínio da modulação, p. 7
ω_c	frequência angular da portadora, p. 11
ω_m	frequência angular de modulação, p. 11
$\phi(t)$	fase do sinal analítico, p. 11
$a(t)$	envoltória no domínio do tempo contínuo, p. 11
$a_k(t)$	envoltória da k-ésima sub-banda, no domínio do tempo contínuo, p. 11
$c[n]$	portadora no domínio discreto, p. 6
$c(t)$	portadora no domínio contínuo, p. 4

$c_k[n]$ portadora da k-ésima sub-banda, no domínio do tempo discreto, p. 8

$c_k(t)$ portadora da k-ésima sub-banda, no domínio do tempo contínuo, p. 11

f_c frequência da portadora, p. 4

k_a sensibilidade à amplitude, p. 4

$m[n]$ moduladora no domínio discreto, p. 6

$m(t)$ moduladora no domínio contínuo, p. 4

$m_k[n]$ moduladora da k-ésima sub-banda, no domínio do tempo discreto, p. 8

$m_k(t)$ moduladora da k-ésima sub-banda, no domínio do tempo contínuo, p. 11

$s[n]$ sinal modulado no domínio discreto, p. 6

$s(t)$ sinal modulado no domínio contínuo, p. 4

$x[n]$ sinal de valores reais, no domínio do tempo discreto, p. 8

$x_k[n]$ k-ésima sub-banda do sinal, no domínio do tempo discreto, p. 8

$x_+(t)$ sinal analítico no domínio do tempo, p. 10

Lista de Abreviaturas

AM	modulação em amplitude, p. 4
DFT	transformada discreta de Fourier, p. 1
DSB-SC	banda lateral dupla com supressão de portadora, p. 5
FFT	transformada rápida de Fourier, p. 1
MTF	função de transferência em modulação, p. 2
SSB	banda lateral suprimida, p. 5
STFT	transformada de Fourier de curta duração, p. 1
STI	índice de transmissão de fala, p. 2
VSB	banda lateral vestigial, p. 5
VoIP	voz sobre IP, p. 1

Capítulo 1

Introdução

1.1 Contexto e Motivação

Aplicações que oi dependam de captação de voz, como chamadas de voz sobre IP (VoIP, do inglês *voice over IP*) e reconhecimento de fala, podem ter sua experiência de uso e assertividade melhoradas quando a fala captada é mais inteligível, o que pode ser potencializado com técnicas de *speech enhancement* aplicadas sobre o áudio analisado.

No conjunto das técnicas de *speech enhancement*, as que utilizam a transformada discreta de Fourier (DFT, do inglês *discrete Fourier transform*) na forma de uma transformada rápida de Fourier (FFT, do inglês *fast Fourier transform*) apresentam baixa complexidade computacional, além de permitirem a modificação dos coeficientes do sinal no domínio da frequência e sua ressíntese no domínio do tempo. Entretanto, o uso da DFT é mais adequado para sinais estacionários. Sinais quase estacionários, isto é, sinais cuja estatística é aproximadamente constante em curtos períodos, tais como a fala — em janelas da ordem de milissegundos —, podem ser bem representados pela transformada de Fourier de curta duração (STFT, do inglês *short-time Fourier transform*). Sobre esta, é possível utilizar algoritmos de redução de ruído, como a subtração espectral e o filtro de Wiener [4]. Uma das técnicas evoluídas a partir da STFT é o espectrograma de modulação — objeto da presente pesquisa —, que representa oscilações de segunda ordem no sinal de áudio analisado, no chamado domínio da modulação.

O sinal de fala pode ser representado como a sobreposição de portadoras geradas pelas cordas vocais, cujas amplitudes e frequências variam lentamente em consequência das mudanças provocadas pelo trato vocal e de seus articuladores, durante a fonação. A bibliografia demonstra que a componente AM contribui para a inteligibilidade do sinal de fala, uma vez que a envoltória quantiza a estrutura temporal de fonemas, sílabas e frases, atribuindo ritmicidade a essas unidades [5, 6].

Quanto à percepção auditiva, atribui-se à cóclea a capacidade de filtrar o som (de banda larga), em diversas sub-bandas de banda estreita, de forma que modulações em amplitude sobre cada sub-banda sejam passadas adiante no sistema auditivo. Portanto, o espectrograma de modulação demonstra-se adequado para aplicações de *speech enhancement*, pois permite a representação da modulação em um domínio que considere diferentes frequências de portadora — de forma análoga ao "banco de filtros" da cóclea — e representa, tal como a STFT, a quase estacionariedade em curtos períodos de tempo característica em sinais de fala. Avaliações do desempenho de técnicas de *speech enhancement* que atuam no domínio do espectrograma de modulação reforçam essa justificativa, indicando bom desempenho em índices de inteligibilidade [7].

Uma segunda técnica disponível é a filtragem de modulação [8]. Essa técnica se baseia nas evidências de que, ao aumentar gradualmente a frequência de corte para um filtro passa-baixas aplicado sobre oscilações da envoltória de um sinal, as componentes acima de 16Hz apresentam apenas um incremento marginal na inteligibilidade [9]. Dessa forma, as componentes de modulação abaixo do limiar de 16Hz são suficientes para uma boa compreensão da fala. A partir dessa característica da percepção, o filtro de modulação é capaz de limitar ruídos sobre a envoltória que incorrem fora da região filtrada, dessa forma aumentando a inteligibilidade.

Por fim, a bibliografia que aborda essas técnicas, apesar de circular há cerca de 20 anos em pesquisa, possui poucos documentos que agregam, condensam e abordam sua evolução e suas formas mais sofisticadas, que carecem de uma apresentação da teoria em que se baseiam ao alcance de um leitor interessado [4, 10].

1.2 Trabalhos Relacionados

HOUTGAST e STEENEKEN [11] descrevem a função de transferência em modulação (MTF, do inglês *modulation transfer function*) para ambientes fechados como forma de medição de inteligibilidade da fala nesses locais, degradada por ruídos quasi-estacionários ou reverberação. A dupla de autores também introduziu o índice de transmissão de fala (STI, do inglês *speech transmission index*) [12], uma medida mais objetiva para a capacidade de um canal preservar a inteligibilidade da fala, cujo cálculo é efetuado a partir da MTF.

DRULLMAN *et al.* [9], a partir de experimento em que varia a frequência de corte de um filtro passa-baixas sobre parcelas moduladoras em sub-bandas críticas de sinais de fala, identificaram, entre outras evidências, que as componentes abaixo de 16Hz são suficientes para uma boa compreensão da fala. Essas evidências justificam a aplicabilidade da filtragem de modulação.

SCHIMMEL [13] traz um panorama da teoria e prática das ferramentas de análise, síntese e filtragem em modulação, com argumentos sustentados sobre as publicações de SCHIMMEL e ATLAS [8], ATLAS e SHAMMA [14] acerca de demodulação coerente, abordada com maior profundidade por CLARK [15].

Algumas aplicações que utilizam da análise no domínio da modulação realizam tarefas como [10]: classificação [16], reconhecimento de fala [17] [18] [19] [20] [21] [22] [23].

1.3 Escopo e Objetivos

O objetivo geral do projeto é apresentar de forma acessível o espectro de modulação e a filtragem de modulação, juntamente com uma aplicação prática sua em processamento de áudio.

Os objetivos específicos são: (1) Apresentar a filtragem de modulação. (2) Apresentar a teoria do espectro de modulação. (2) Apresentar técnicas de *speech enhancement* no domínio da STFT. (3) Comparar seu desempenho quando aplicadas no espectrograma de amplitude e no espectrograma de modulação. (4) Apresentar uma aplicação, já presente na bibliografia, que aborde as duas técnicas comentadas. Dessa forma, o trabalho servirá como material de consulta para pesquisadores que se interessem pelo tema.

1.4 Materiais e Organização do Texto

Capítulo 2

Fundamentação Teórica

2.1 Modulação em Amplitude

A modulação em amplitude (AM, do inglês *Amplitude Modulation*) é o processo no qual a envoltória de uma onda portadora oscila em torno de um valor médio, conforme um sinal modulador em baixa frequência [2]. No contexto de telecomunicações, a moduladora é um sinal em banda base, isto é, contém a informação em um espectro delimitado a ser transmitida.

Considerando a portadora como um tom puro de frequência f_c e amplitude A_c:

$$c(t) = A_c \cos(2\pi f_c t). \tag{2.1}$$

Um sinal $s(t)$ modulado em amplitude é definido pela composição de suas parcelas portadora $c(t)$ e moduladora $m(t)$:

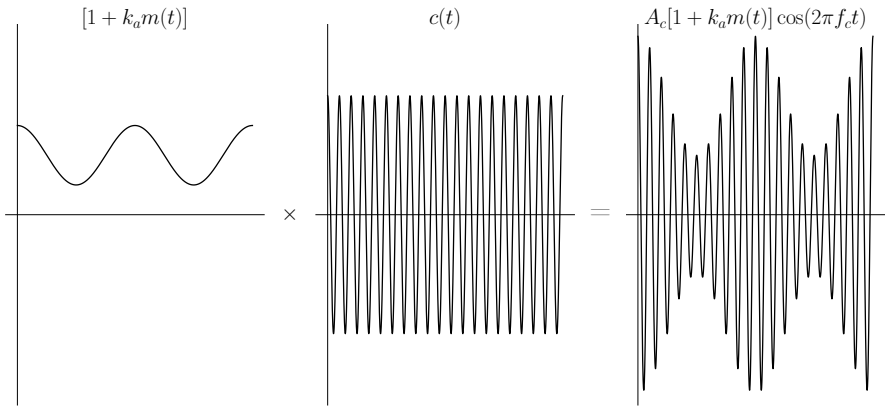

Figura 2.1: Ilustração para equação (2.2) quando (2.3) é satisfeita.

$$s(t) = [1 + k_a m(t)]c(t) = A_c[1 + k_a m(t)]\cos(2\pi f_c t) \tag{2.2}$$

em que k_a é a sensibilidade à amplitude da moduladora. Para que a envoltória dada pela função $1+k_a m(t)$ seja sempre positiva, evitando reversão de fase em decorrência de sobremodulação, basta que

$$|k_a m(t)| < 1, \quad \forall t. \tag{2.3}$$

No domínio da frequência, a equação (2.2) é dada por:

$$S(f) = \frac{A_c}{2}[\delta(f - f_c) + \delta(f + f_c)] + \frac{k_a A_c}{2}[M(f - f_c) + M(f + f_c)], \tag{2.4}$$

em que δ é o impulso unitário e $M(f)$ é a parcela moduladora no domínio da frequência.

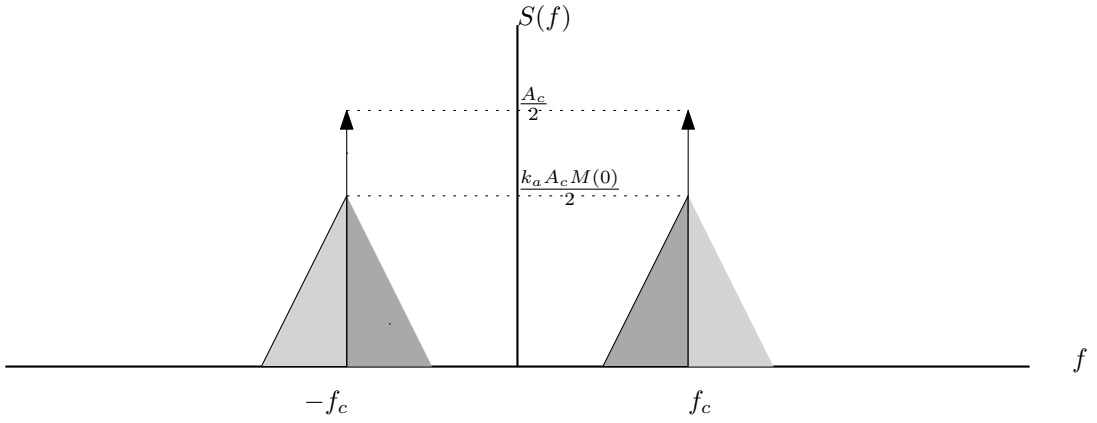

Figura 2.2: Ilustração para equação (2.4), adaptado de [2].

Pela equação (2.4), observa-se que a parcela com impulsos unitários não contribui diretamente para a transmissão da informação contida em $M(f)$, além de despender grande parte da potência do sinal. Uma segunda característica de $S(f)$ é a existência de dois lóbulos simétricos no domínio da frequência, o que confere redundância na transmissão do sinal e maior largura de banda. Para solucionar essas duas características indesejáveis em telecomunicações, foram criadas as modulações com banda lateral dupla e supressão de portadora (DSB-SC, do inglês *double-sideband suppressed-carrier*), banda lateral vestigial (VSB, do inglês *vestigial sideband modulation*) e banda lateral suprimida (SSB, do inglês *single-sideband modulation*), que vão além do escopo desse trabalho.

Uma notação mais simples para a equação (2.2), presente na maior parte da bibliografia, em que $k_a = 1$ e que desconsidera o *offset*, é dada pelo produto dos

termos:

$$s(t) = m(t)c(t). \tag{2.5}$$

No domínio do tempo discreto,

$$s[n] = m[n]c[n]. \tag{2.6}$$

As parcelas moduladoras $m(t)$ e $m[n]$ podem assumir valores complexos [13], apesar do termo *envoltória* referir-se, geralmente, à parcela de valores reais não-negativos [2] com lenta variação sobre um sinal. No caso particular em que $m(t)$ é um sinal de valores reais não-negativos, para a equação (2.5), temos que:

$$m(t) \geq s(t), \forall t. \tag{2.7}$$

A decomposição do sinal em envoltória e portadora em valores complexos é abordada por ATLAS *et al.* [24].

explorar mais

2.2 Espectro de Modulação

O espectro de modulação, por vezes chamado de espectrograma de modulação, é o nome dado a representações que evidenciam frequências associadas a modulações presentes no sinal. GREENBERG e KINGSBURY [25] apresentam o espectrograma de modulação em 1997 como "*um novo formato de representação para a fala* [...] *que exibe e codifica o sinal em termos da distribuição de modulações lentas em função do tempo e da frequência*". Trabalhos anteriores fizeram uso do espectro da envoltória do sinal e sua correspondência com a inteligibilidade de fala, como HOUTGAST e STEENEKEN [11] STEENEKEN e HOUTGAST [12] [26] a partir de 1973 e DRULLMAN *et al.* [9] em 1994.

O método apresentado por GREENBERG e KINGSBURY [25] consiste em obter a envoltória nas parcelas filtradas de um sinal por sub-bandas de um banco de filtros, para então calcular suas transformadas de Fourier. Essa abordagem também é utilizada para o espectro de modulação apresentado por ATLAS *et al.* [24] [14] a partir de 2003, definido como "*uma representação bidimensional* [...] *onde a primeira dimensão é a célebre frequência acústica e a segunda dimensão é a frequência de modulação*". A principal novidade, nesse segundo caso, é a estimação e decomposição das partes moduladora e portadora de forma coerente, em vez da decomposição por módulo e fase.

Em uma abordagem distinta da tradicional, PALIWAL *et al.* [27][28][29] utiliza a STFT em vez do banco de filtros, de forma que as trajetórias dos módulos do espectro acústico no tempo são adequadas para o cálculo do espectro de modulação

de tempo curto.

Para compreender a representação do espectrograma de modulação, definem-se dois domínios, na bibliografia [10]. O domínio acústico é uma representação bidimensional de um sinal na qual, para as frequências f e os instantes de tempo t, mapeiam-se os respectivos coeficientes complexos de Fourier [30]:

chamo de STFT?

$$X(t, f) : \mathbb{R} \times \mathbb{R} \to \mathbb{C}. \tag{2.8}$$

O chamado espectrograma de amplitude é a representação do módulo desse domínio:

modulo?

$$\mathrm{Spec}(t, f) = |X(t, f)|\,. \tag{2.9}$$

Por sua vez, o domínio da modulação é capaz de representar esse mesmo sinal a partir das mesmas coordenadas do domínio acústico, acrescidas da frequência η de oscilação de sua envoltória, mapeando-o também em coeficientes complexos de Fourier:

$$\mathcal{X}(\eta, f, t) : \mathbb{R} \times \mathbb{R} \times \mathbb{R} \to \mathbb{C}. \tag{2.10}$$

O correspondente espectrograma de modulação é dado por

$$\mathrm{ModSpec}(\eta, f, t) = |\mathcal{X}(\eta, f, t)|\,. \tag{2.11}$$

A seguir, serão apresentadas duas possibilidades de implementação: uma com banco de filtros, e outra com a STFT.

Na implementação com banco de filtros, o espectro de modulação $P_k[\mu]$ segue da seguinte forma [31]:

$$P_k[\mu] = \sum_{n=0}^{N-1} w[n] m_k[n] e^{-j\frac{2\pi\mu}{N}n} \tag{2.12}$$

em que k é a k-ésima sub-banda do banco de filtros, μ é a frequência de modulação, $w[n]$ é a função-janela com tamanho de N amostras e $m_k[n]$ é a parcela moduladora obtida pela demodulação da k-ésima sub-banda. A estimação das partes portadora e moduladora também é utilizada na filtragem de modulação, descrita na secção 2.3.

Ao se utilizar a STFT, calcula-se $X[l, k]$ a partir do sinal $x[n]$, da taxa de amostragem do sinal, da função janela $w[n]$ de análise, do comprimento M do *frame* em amostras, do número N de amostras da DFT, e do salto da janela em H amos-

tras [10, 32]:

$$X[l,k] = \sum_{n=0}^{M-1} x[n+lH]w[n]\mathrm{e}^{-2\mathrm{j}\pi nk/N}. \tag{2.13}$$

Dessa forma, obtém-se os coeficientes de Fourier para cada *bin* k na frequência e *frame* l no tempo, análogos às dimensões de $f(t,\omega)$.

Finalmente, obtêm-se os coeficientes complexos no domínio da modulação aplicando-se uma segunda STFT sobre o módulo das séries de *frames*, para cada *bin* k do domínio acústico [10]:

$$\mathcal{X}[\ell,k,\mu] = \sum_{l=0}^{\mathcal{M}-1} |X[l+\ell\mathcal{H},k]|\, v[l]\mathrm{e}^{-2\mathrm{j}\pi\mu l/\mathcal{N}}, \tag{2.14}$$

no qual ℓ é o índice do *frame* no domínio da modulação, k é o índice da frequência advinda do domínio acústico, μ é o índice do *bin* da frequência de modulação, $\mathcal{N}$ é o número de amostras da DFT, $\mathcal{M}$ é o tamanho do *frame* em amostras e $\mathcal{H}$ é o deslocamento da janela de análise $v[l]$ no domínio da modulação, em amostras. O espectrograma de modulação é dado por

$$\mathrm{ModSpec}[\ell,k,\mu] = |\mathcal{X}[\ell,k,\mu]|\,. \tag{2.15}$$

Nas ilustrações a seguir, é possível visualizar a forma de onda no domínio do tempo, a STFT, o espectro de modulação obtido a partir do banco de filtros e o espectro de modulação obtido pela STFT sobre o domínio acústico. O trecho de fala masculina foi retirado de https://ufpafalabrasil.gitlab.io/.

inversa?

imagens?

specs de implementação

2.3 Filtragem de Modulação

A filtragem de modulação é uma técnica para filtragem de envoltórias com oscilações lentas, presentes em sub-bandas de frequência, aplicada sobre sinais não-estacionários.[33] [3]. O objetivo, em aplicações de fala, é obter um sinal mais inteligível na saída, composto apenas pela composição de tons com modulações de baixa frequência associadas à articulação vocal, eliminando as demais parcelas, as quais, quando presentes, podem conter ruído e artefatos que degradam a compreensão da fala.

Essa técnica assume que um sinal $x[n]$ de valores reais é representado pela soma dos produtos entre moduladora e portadora que compõem cada sub-banda $x_k[n]$, obtidas a partir da filtragem por um filtro passa-bandas $h_k[n]$:

$$x_k[n] = h_k[n] * x[n] \tag{2.16}$$

$$x[n] = \sum_{k=0}^{K-1} x_k[n] = \sum_{k=0}^{K-1} m_k[n]c_k[n] \tag{2.17}$$

A implementação genérica da filtragem de modulação está representada no diagrama de blocos abaixo.

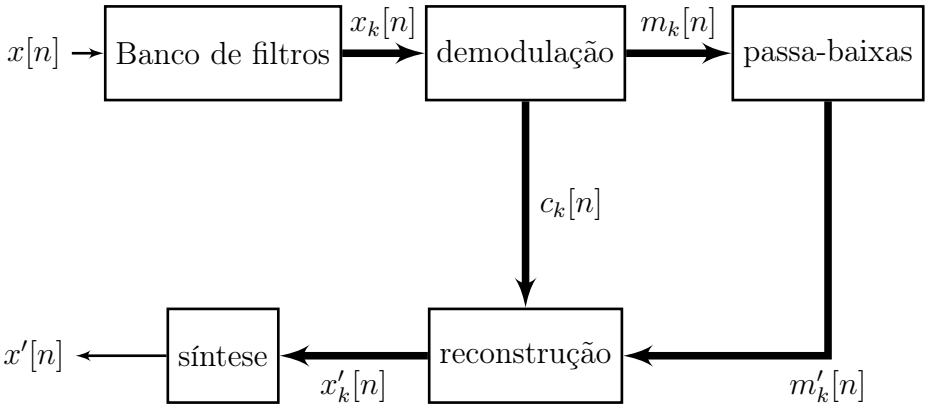

Figura 2.3: Filtro de modulação, em diagrama de blocos. Adaptado de [3].

O sinal banda-larga $x[n]$ passa por um banco de filtros $h_k[n]$, no qual as $x_k[n]$ saídas do banco de filtros passam por um módulo de demodulação, no qual são separadas as parcelas moduladoras $m_k[n]$ e portadoras $c_k[n]$ de cada sub-banda. Cada parcela $m_k[n]$ é filtrada por um filtro passa-baixas, cuja resposta ao impulso é $g[n]$. Em seguida, cada sub-banda é reconstruída a partir da portadora $c_k[n]$ e da modulação filtrada $m'_k[n]$. Finalmente, o sinal de banda larga $x'_k[n]$ é sintetizado com as sub-bandas $x'_k[n]$ reconstruídas, obtidas a partir das modulações filtradas.

Em aplicações de *speech enhancement*, o projeto do banco de filtros emula a resposta em frequência da cóclea no sistema auditivo. Por sua vez, a frequência de corte do filtro passa-baixas é tipicamente de 16Hz [9], tal que frequências abaixo desse valor estão associadas às periodicidades na construção de frases, palavras e sílabas.

Uma forma de calcular $m'_k[n]$ é a partir do espectro de modulação $P_k[\mu]$, apresentado na equação 2.12:

$$m'_k[n] = g[n] * m_k[n] = \sum_{q=0}^{n} g[n-q]m_k[q] \tag{2.18}$$

$$m'_k[n] = \sum_{\mu=0}^{N-1} G[\mu]P_k[\mu]e^{j\frac{2\pi\mu}{N}n} \tag{2.19}$$

janela na eq?

Dessa forma, obtém-se $m'_k[n]$ a partir da transformada inversa do espectro de modulação filtrado por $G[\mu]$ [31].

specs de implementação

2.4 Demodulação Incoerente e Coerente

A demodulação, no contexto de telecomunicações, consiste em recuperar a parcela original em banda base de um sinal modulado, transladando seu espectro para sua posição original [34]. Os métodos disponíveis na bibliografia de filtragem de modulação são separados em demodulações incoerentes, que estimam a moduladora a partir de operações de módulo, e em demodulações coerentes, que estimam a portadora a partir da frequência instantânea do sinal [35].

Uma demodulação incoerente é a detecção de envoltória feita a partir da transformada de Hilbert. A transformada de Hilbert é capaz de gerar o sinal em quadratura, deslocando sua fase em $\pi/2$ radianos. Sua definição, no domínio do tempo contínuo e seu espectro [36][37], são dados por:

$$\mathcal{H}\{x(t)\} = x(t) * \frac{1}{\pi t} = \frac{1}{\pi}\int_{-\infty}^{+\infty} \frac{x(\tau)}{t-\tau}\,dx \tag{2.20}$$

$$\hat{x}(t) = \mathcal{H}\{x(t)\} \tag{2.21}$$

$$\operatorname{sgn}(\omega) = \begin{cases} 1, & \omega < 0 \\ 0, & \omega = 0 \\ -1, & \omega < 0 \end{cases} \tag{2.22}$$

$$\hat{X}(\omega) = -j\operatorname{sgn}(\omega)X(\omega) \tag{2.23}$$

em que $\mathcal{H}\{\cdot\}$ é o operador para a transformada de Hilbert, $x(t)$ é um sinal de valores reais, $*$ é o operador para convolução, $\hat{x}(t)$ é o sinal em quadratura, $\operatorname{sgn}(\omega)$ é a função sinal e $\hat{X}(\omega)$ é o espectro do sinal em quadratura.

O método consiste em decompor um sinal real em partes moduladora e portadora a partir de um sinal analítico, obtido pela soma do sinal em questão com sua transformada de Hilbert:

hilbert discreto?

$$x_+(t) = x(t) + \mathrm{j}\mathcal{H}\{x(t)\} \tag{2.24}$$

$$|x_+(t)| = \sqrt{x^2(t) + \hat{x}^2(t)} \tag{2.25}$$

$$\phi(t) = \arctan\left(\frac{\hat{x}(t)}{x(t)}\right) \tag{2.26}$$

$$X_+(\omega) = X(\omega) + \operatorname{sgn}(\omega)\hat{X}(\omega) \tag{2.27}$$

$$X_+(\omega) = \begin{cases} 2X(\omega), & \omega < 0 \\ X(0), & \omega = 0 \\ 0, & \omega < 0 \end{cases} \tag{2.28}$$

em que $x_+(t)$, $|x_+(t)|$ e $\phi(t)$ são, respectivamente, o sinal analítico, seu módulo e

fase, $X_+(\omega)$ é o espectro do sinal analítico.

Em seguida, decompõem-se a moduladora e a portadora a partir do módulo e fase do sinal analítico:

$$a(t) = |x_+(t)| \tag{2.29}$$

$$c(t) = \cos\phi(t) = \mathrm{e}^{\mathrm{j}\phi(t)} \tag{2.30}$$

em que, $a(t)$ é a envoltória, $c(t)$ é a portadora. Note que, ao estimar a envoltória a partir do módulo do sinal analítico, seus valores serão reais não-negativos.

discreto e sub banda

No contexto de sub-bandas, em que as envoltórias e portadoras são estimadas para cada sub-banda k, tal como na etapa de demodulação na filtragem de modulação, as equações 2.24, 2.29 e 2.30 são aplicadas em cada sub-banda:

$$x_{k,+}(t) = x_k(t) + \mathrm{j}\mathcal{H}\{x_k(t)\} \tag{2.31}$$

$$m_k(t) = a_k(t) = |x_{k,+}(t)| \tag{2.32}$$

$$c_k(t) = \cos\phi_k(t) = \mathrm{e}^{\mathrm{j}\phi_k(t)} \tag{2.33}$$

O método acima possui limitações importantes [13]. A largura de banda da portadora estimada espalha-se além dos limites da banda original do sinal. Esse aspecto inviabiliza a etapa de reconstrução em bancos de filtros, como na filtragem de modulação. Para exemplificar essa limitação, considere o sinal complexo $y(t)$:

$$y(t) = \cos(\omega_m t)e^{j\omega_c t}, \quad \omega_c > \omega_m \tag{2.34}$$

em que ω_c e ω_m são as frequências da portadora e moduladora. O módulo ao quadrado do sinal é dado por

$$|y(t)|^2 = y(t)y^*(t) \tag{2.35}$$

$$|y(t)|^2 = \cos^2(\omega_m t) \tag{2.36}$$

$$|y(t)|^2 = \frac{1}{2}\cos(2\omega_m t) + \frac{1}{2} \tag{2.37}$$

cujo espectro é limitado em banda. Por sua vez, o módulo de $y(t)$ é dado por:

$$|y(t)| = |e^{j\omega_c t}\cos(\omega_m t)| \tag{2.38}$$

$$|y(t)| = |\cos(\omega_m t)| \tag{2.39}$$

que contém descontinuidades em:

$$t = \frac{\pi}{\omega_m}(\frac{1}{2} + \mathbf{k}), \quad \mathbf{k} \in \mathbb{Z}. \tag{2.40}$$

Para representar essas descontinuidades, são necessárias infinitas frequências. O espectro do módulo de $y(t)$ é dado pela equação 2.43:

calcular limite?

$$m(t) = |y(t)| \tag{2.41}$$

$$M(\omega) = \mathcal{F}\{|y(t)|\} \tag{2.42}$$

$$M(\omega) = \sum_{\mathbf{k}=1}^{\infty} \frac{4}{-(-1)^{\mathbf{k}}((2\mathbf{k})^2) - 1)} \delta(\omega - 2\mathbf{k}\omega_m) \tag{2.43}$$

em que $\mathcal{F}\{\cdot\}$ é o operador para a transformada de Fourier. COHEN *et al.* [38] demonstra que a representação de módulo e fase não é única, de forma que um sinal complexo possui múltiplos pares módulo-fase que o representam. Tal como na demodulação incoerente, haverão descontinuidades nessas representações caso sua amplitude não seja positiva definida.

eq acima apêndice

adicionar grafico exemplificando

Como alternativa, há a detecção coerente de portadora [8, 33, 35], que visa à detecção de uma portadora com banda estreita.

Para compreender as diferenças entre demodulação coerente e incoerente, representamos a equação 2.5 na forma polar [13]:

$$a_x(t)e^{j\phi_x(t)} = \left[a_m(t)e^{j\phi_m(t)}\right]\left[a_c(t)e^{j\phi_c(t)}\right] \tag{2.44}$$

nessa representação, a envoltória $a_x(t)$ e fase $\phi_x(t)$ do sinal são dados pela composição das envoltórias e fases da portadora e moduladora:

$$a_x(t) = a_m(t)a_c(t) \tag{2.45}$$

$$\phi_x(t) = \phi_m(t) + \phi_c(t) \tag{2.46}$$

Como discutido por COHEN *et al.* [38], há múltiplas soluções para a equação 2.44. A detecção incoerente assume que:

falar da fase

$$\phi_m(t) = 0 \tag{2.47}$$

$$a_c(t) = 1 \tag{2.48}$$

$$a_m(t) = a_x(t) \tag{2.49}$$

$$\phi_c(t) = \phi_x(t) \tag{2.50}$$

Diferentemente, a detecção coerente considera que a parcela moduladora assuma valores complexos, de forma que $\phi_m(t)$ tenha valores não-nulos. Para resolver a ambiguidade na decomposição das fases de $\phi_m(t)$ e $\phi_c(t)$, impõe-se a restrição de que $\phi_c(t)$ recebe apenas componentes de fase com lenta variação, em relação à fase

do sinal original.

A fase $\phi_c(t)$ da portadora é estimada a partir da frequência instantânea do sinal $x(t)$, na qual aplica-se um filtro passa-baixas $h_{lp}(t)$, integrando-a no tempo:

$$\alpha_c(t) = \frac{d\phi_x(t)}{dt} * h_{lp}(t) \tag{2.51}$$

$$\phi_c(t) = \int_0^t \alpha_c(t)\, d\tau \tag{2.52}$$

Em que $\alpha_c(t)$ é a frequência instantânea filtrada. Note que a derivada de $\phi_x(t)$ no tempo expressa a frequência instantânea de $x(t)$. Finalmente, a parte moduladora é obtida ao descontar do sinal original a estimativa de sua parcela portadora:

$$c(t) = e^{j\phi_c(t)} \tag{2.53}$$

$$m(t) = \frac{x(t)}{c(t)} = x(t) \cdot c^*(t) \tag{2.54}$$

Substituindo a equação:

$$m(t) = \frac{x(t)}{e^{j\phi_c(t)}} = x(t)e^{-j\phi_c(t)}. \tag{2.55}$$

Existem formas distintas de realizar a demodulação coerente. Em comum, compartilham a estimativa das componentes a partir da fase do sinal, preservando a largura de banda final.

A demodulação coerente por centro de gravidade espectral [35] [33] [31] estima a portadora da equação 2.53 a partir da frequência que corresponde ao centróide da distribuição, denotada por ω_0:

$$\omega_0 = \frac{\int_{-\infty}^{+\infty} \omega S_{xx}(\omega)\, d\omega}{\int_{-\infty}^{+\infty} S_{xx}(\omega)\, d\omega} \tag{2.56}$$

$$c(t) = e^{j\omega_0 t} \tag{2.57}$$

em que $S_{xx}(\omega)$ é a densidade espectral de potência do sinal original $x(t)$. Note que a equação 2.56 aplica-se para sinais estacionários, enquanto que a equação 2.57 equivale à equação 2.52, com a fase proporcional ao tempo.

No domínio discreto, para sinais não-estacionários, a frequência instantânea para o l-ésimo *frame* é estimada a partir da STFT. No contexto de filtragem por subbandas:

$$f_k[l] = \frac{\sum_{\mathcal{K}=0}^{M-1} r[l]|X_k[l,\mathcal{K}]|^2}{|X_k[l,\mathcal{K}]|^2} \tag{2.58}$$

em que $f_k[n]$ é a frequência instantânea da k-ésima sub-banda, $\mathcal{K}$ é o índice do *bin* e M é metade do tamanho do *frame* N, sendo par, em amostras.

Uma segunda demodulação coerente, também presente na bibliografia, estima a portadora a partir da detecção de frequência fundamental F_0 do sinal original, de forma análoga à equação 2.52, no domínio do tempo discreto:

$$\phi_0[n] = \sum_{p=0}^{n} F_0[n] \tag{2.59}$$

Em que F_0 é o *pitch* estimado para $x[n]$. Esse método, chamado demodulação coerente harmônica, considera que cada sub-banda está relacionada a um harmônico múltiplo da frequência fundamental, tal que a portadora seja obtida por:

$$c_k[n] = e^{jk\phi_0[n]} \tag{2.60}$$

$$m_k[n] = \sum_{p=0}^{n} h_{lp}[n-p] \cdot x[p] \cdot c_k^*[p]. \tag{2.61}$$

Tal que a parcela moduladora seja obtida ao aplicar filtragem passa-baixas no sinal original, previamente deduzido de sua portadora.

Capítulo 3

Demodulação

3.1 Demodulação por Banco de Filtros

3.1.1 Filtragem e Ressíntese por Sub-bandas

O pacote confere suporte para dois projetos distintos de bancos de filtros.

O primeiro banco de filtros tem reconstrução perfeita, no qual as sub-bandas de mesma largura são equidistantes.

O segundo banco de filtros permite a escolha arbitraria da posição e largura dos filtros, sem reconstrução perfeita.

3.1.2 Demodulação por Transformada de Hilbert

3.1.3 Demodulação por Centro de Gravidade Espectral

3.1.4 Exemplos

3.2 Demodulação por Conteúdo Harmônico

3.2.1 Detecção de Pitch

3.2.2 Demodulação por Conteúdo Harmônico

3.2.3 Demodulação por Conteúdo Harmônico por Centro de Gravidade Espectral

3.2.4 Exemplos

Capítulo 4

Speech Enhancement

4.1 *Framework* Análise-Modificação-Síntese

O *Framework* análise-modificação-síntese (AMS) é o procedimento que realiza as etapas de análise com a STFT do sinal, modificação no domínio da frequência acústica e ressíntese do sinal modificado com a STFT inversa [10]. Caso o sinal de entrada esteja impregnado por ruído, é possível, de posse de uma estimativa da parcela ruidosa, subtraí-la do sinal impregnado no domínio da frequência.

Há diferentes métodos e algoritmos para a etapa de modificação no domínio acústico, os quais também podem ser aplicados no domínio na modulação. Geralmente, nesse procedimento, a modificação é realizada no módulo do sinal impregnado no domínio da frequência, preservando a fase.

Assumimos um sinal impregnado por ruído:

$$x[n] = s[n] + d[n] \tag{4.1}$$

em que $s[n]$ e $d[n]$ são as parcelas de sinal e de ruído, respectivamente. No contexto de *Speech Enhancement*, $s[n]$ é o sinal de fala. O objetivo final é obter a estimativa $\hat{s}[n]$ mais próxima possível de $s[n]$.

Ao retomar a equação 2.13, temos o sinal impregnado no domínio acústico em função da STFT das parcelas $s[n]$ e $d[n]$, representadas por $S[l,k]$ e $D[l,k]$:

$$X[l,k] = S[l,k] + D[l,k], \tag{4.2}$$

cuja representação em fasor é dada por:

$$X[l,k] = |X[l,k]|e^{j\angle X[l,k]} \tag{4.3}$$

em que $|X[l,k]|$ e $e^{j\angle X[l,k]}$ são módulo e fase do espectro acústico. A subtração

espectral, um dos métodos de *Speech Enhancement*, estima o espectro da parcela ruidosa $d[n]$ a partir de pausas na fala contidas em $s[n]$.

O *framework* AMS para *Speech Enhancement* no domínio acústico é ilustrado no diagrama 4.1. A etapa de modificação no domínio acústico retorna $|\hat{S}[l,k]|$, que representa a estimativa do módulo do sinal desejado. Assume-se que sua fase é igual a de $X[l,k]$. Dessa forma, a estimativa do sinal desejado no domínio acústico é dada por $|\hat{S}[l,k]|e^{j\angle X[l,k]}$. Ao efetuar a STFT inversa desse sinal, obtém-se a estimativa $\hat{s}[n]$ do sinal desejado no domínio do tempo. O procedimento é ilustrado no diagrama 4.1.

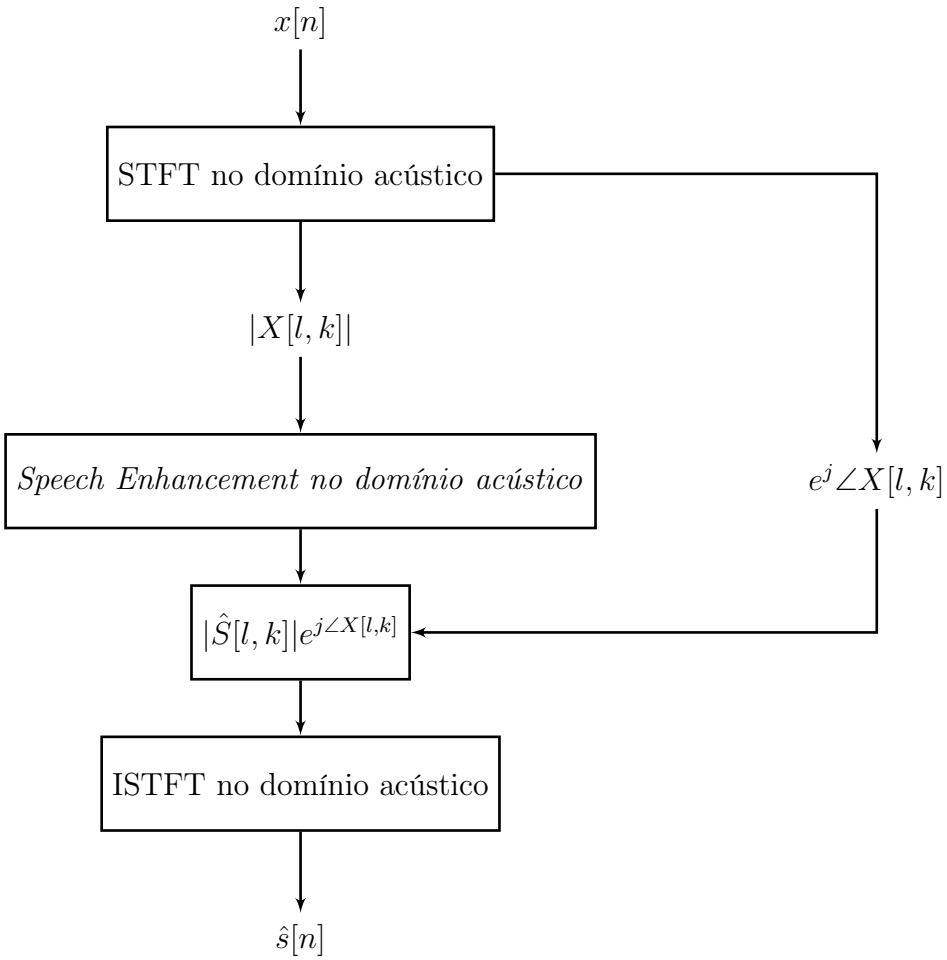

Figura 4.1: *Framework AMS para Speech Enhancement no domínio acústico.*

Por sua vez, o *framework AMS* para o domínio da modulação é ilustrado no diagrama 4.1. Aplica-se a STFT no domínio da modulação ao módulo do espectro no domínio acústico, obtendo $|\mathcal{X}[l,k,\mu]|$ e $e^{j\angle\mathcal{X}[l,k,\mu]}$, que são o módulo e fase do espectro de modulação.

4.2 *Speech Enhancement* no Domínio Acústico

enhancement por melhora-mento

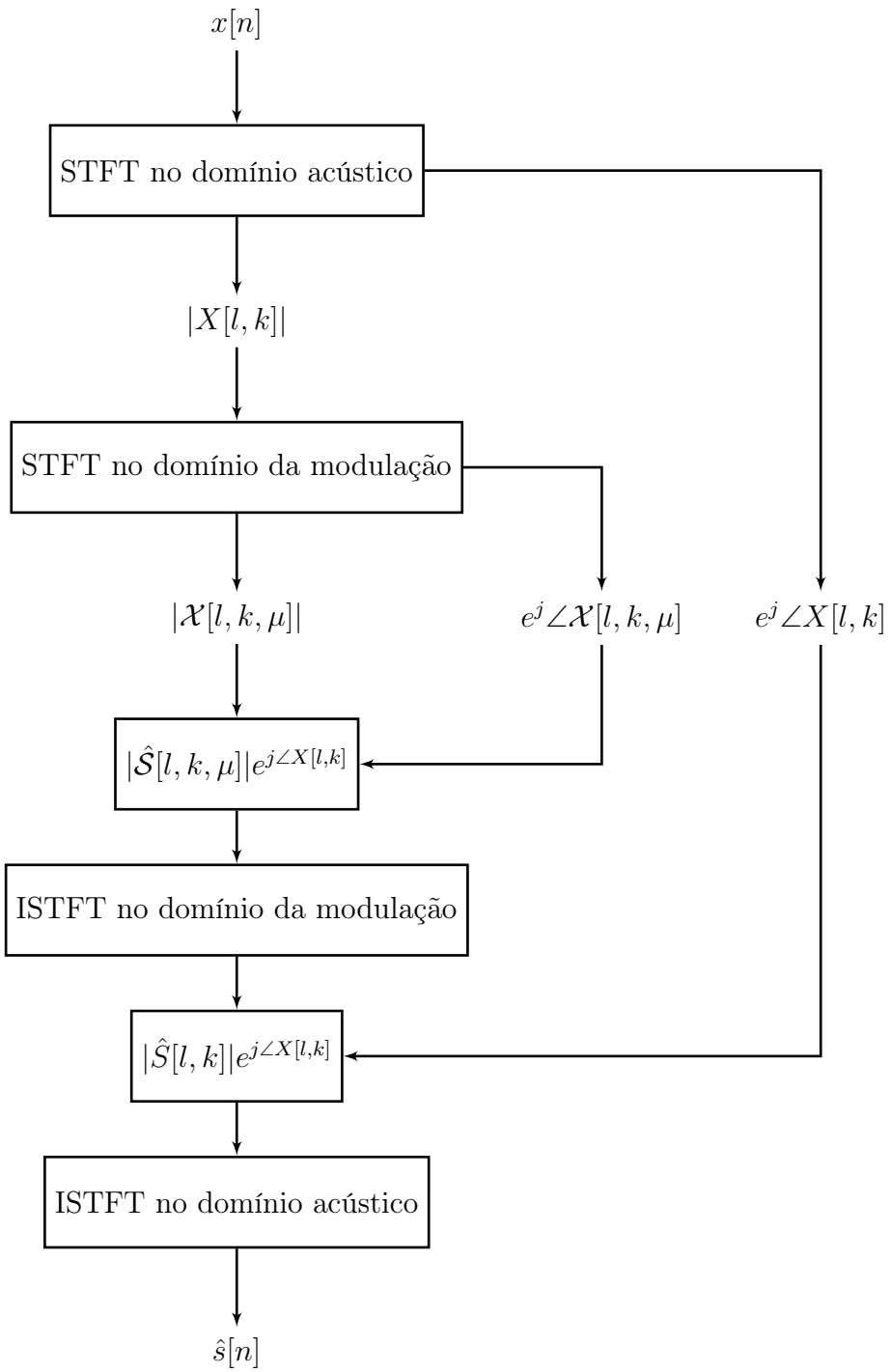

Figura 4.2: *Framework AMS para Speech Enhancement no domínio acústico.*

4.2.1 Subtração Espectral no Domínio Acústico

A subtração espectral no domínio acústico, em sua forma mais simples, estima que o espectro do sinal limpo é dado pela espectro do sinal de entrada subtraído da estimativa do espectro do ruído. Assume-se também que a fase da parcela ruidosa é substituível pela fase do sinal de entrada [39]. Dessa forma, temos:

$$\hat{S}[l,k] = \begin{cases} [|X[l,k]| - |\hat{D}[l,k]|]e^{j\angle X[l,k]}, & |X[l,k]| > |\hat{D}[l,k]| \\ 0, & |X[l,k]| \leq |\hat{D}[l,k]| \end{cases} \tag{4.4}$$

em que $\hat{S}[l,k]$ é a estimativa do sinal de fala desejado. Note que a parcela de módulo $|\hat{S}[l,k]|$ na expressão seria negativa caso o módulo do ruído fosse superior ao do sinal. Para tratar esse caso, retifica-se a parcela de módulo obtida, tal que todos os valores negativos sejam truncados a zero.

Para descrever a subtração espectral em termos de densidade espectral de potência, obtém-se a densidade espectral do sinal de fala a partir da equação (4.2):

$$|X[l,k]|^2 = |S[l,k]|^2 + |D[l,k]|^2 + S[l,k] \cdot D^*[l,k] + S^*[l,k] \cdot D[l,k] \tag{4.5}$$

$$|X[l,k]|^2 = |S[l,k]|^2 + |D[l,k]|^2 + 2\,\mathrm{Re}(S[l,k] \cdot D^*[l,k]) \tag{4.6}$$

A densidade espectral de potência na saída é composta pela soma das densidades espectrais de potência do sinal de fala e do ruído, acrescidos de um termo cruzado. O operador $\mathrm{Re}(\cdot)$ descreve a parte real de um número complexo, enquanto o operador $*$ descreve o conjugado complexo. Uma vez que $|D[l,k]|^2$, $S[l,k] \cdot D^*[l,k]$ e $S^*[l,k] \cdot D[l,k]$ não são conhecidos *a priori*, esses termos são aproximados pelos valores esperados $\mathbb{E}\{|D[l,k]|^2\}$, $\mathbb{E}\{S[l,k] \cdot D^*[l,k]\}$ e $\mathbb{E}\{S^*[l,k] \cdot D[l,k]\}$, sendo $\mathbb{E}\{\cdot\}$ o operador para valor esperado. Ao assumir que as parcelas que compõem o sinal são estacionárias, que o ruído tem média igual a zero e é descorrelacionado com o sinal de fala, temos que:

$$\mathbb{E}\{S[l,k] \cdot D^*[l,k]\} = \mathbb{E}\{S[l,k]\} \cdot \mathbb{E}\{^*D[l,k]\} = 0 \tag{4.7}$$

$$\mathbb{E}\{S^*[l,k] \cdot D[l,k]\} = \mathbb{E}\{S^*[l,k]\} \cdot \mathbb{E}\{D[l,k]\} = 0 \tag{4.8}$$

$$|X[l,k]|^2 = |S[l,k]|^2 + |D[l,k]|^2 \tag{4.9}$$

Finalmente, ao aplicar a estimativa do ruído na equação (4.9) e truncando valores negativos tal como na equação (4.4), a estimativa do sinal original é descrita por:

$$|\hat{S}[l,k]|^2 = \begin{cases} |X[l,k]|^2 - |\hat{D}[l,k]|^2, & \text{se } |X[l,k]|^2 > |\hat{D}[l,k]|^2 \\ 0, & \text{caso contrário.} \end{cases} \tag{4.10}$$

Na prática, é importante salientar que o sinal de fala não é estacionário. Mesmo em janelas curtas de tempo da ordem de milissegundos, assume-se que o sinal é quasi-estacionário. Além disso, dependendo da aplicação, o ruído pode ter correlação com o sinal de fala desejado. Esses dois fatores contribuem para que os termos cruzados permaneçam presentes em $|\hat{S}[l,k]|^2$. Apesar disso, por questão de simplicidade,

assume-se que o valor dos termos seja igual a zero.

Ao retificar parte dos valores obtidos, também ocorre um segundo problema: são gerados picos espectrais na vizinhança dos valores truncados. Esses picos geram tons na estimativa obtida do sinal de fala, que não estão presentes no sinal original.

Um método para a subtração espectral no domínio acústico é dado por [40] :

$$|\hat{S}[l,k]|^\gamma = \begin{cases} |X[l,k]|^\gamma - \alpha|\hat{D}[l,k]|^\gamma, & \text{caso } |X[l,k]|^\gamma > (\alpha+\beta)|\hat{D}[l,k]|^\gamma \\ \beta|\hat{D}[l,k]|^\gamma & \text{caso contrário} \end{cases} \tag{4.11}$$

No qual $\alpha(k) \geq 1$ é o fator de subtração, $0 < \beta \ll 1$ é o parâmetro de piso espectral[41] e γ é o fator de potência.

A partir da escolha de α, é possível reduzir a intensidade dos picos tonais característicos da subtração espectral. O parâmetro β suaviza os vales na vizinhança dos picos ao aumentar a intensidade nas vizinhanças, evitando a permanência de picos estreitos. $\gamma = 1$ corresponde à subtração espectral de módulo, enquanto que $\gamma = 2$ corresponde à subtração espectral de potência. Note que, para $\alpha = 1$, $\beta = 0$ e $\gamma = 2$, a equação (4.11) é idêntica à equação (4.10).

Uma vez que α deve ser pequeno para *frames* com razão sinal-ruído baixas e vice-versa, temos que:

$$\alpha = \begin{cases} 5, & \text{SNR} < -5\,\text{dB} \\ \alpha_0 - \frac{\text{SNR}}{s}, & -5\,\text{dB} \leq \text{SNR} \leq 20\,\text{dB} \\ 1, & \text{SNR} \geq 20\,\text{dB} \end{cases} \tag{4.12}$$

no qual α_0 é o valor de α para $\text{SNR} = 0\,\text{dB}$, $\frac{1}{s}$ é a inclinação da reta entre α_0 em $0\,\text{dB}$ e $\alpha = 1$ em $20\,\text{dB}$. . Para a estimativa da parcela de ruído do sinal, é necessário identificar em quais instantes há ausência de fala a partir de um detector de atividade vocal (VAD, do inglês *voice activity detector*).

adicionar efeito da fase

citar quem fala das solucoes

falar o q acontece com β alto ou baixo, α alto ou baixo

descrever como escolher parametros

4.2.2 Filtro de Wiener no Domínio Acústico

O filtro de Wiener é o filtro ótimo e linear com erro mínimo entre o sinal desejado e sua estimativa. Seu projeto pode ser tanto de resposta ao impulso de duração finita (FIR, do inglês *finite-duration impulse response*) quanto de resposta ao impulso de duração infinita (IIR, do inglês *infinite-duration impulse response*)[39].

Para um projeto de filtro FIR no domínio do tempo discreto, a estimativa do sinal desejado é obtida ao multiplicar cada coeficiente à amostra correspondente na saída do sinal:

$$\hat{d}[n] = \sum_{k=0}^{M-1} h_k x[n-k] \tag{4.13}$$

$$\hat{d}[n] = \mathbf{h}^\top \mathbf{x} \tag{4.14}$$

em que M é a quantidade de coeficientes, h_k é o k-ésimo coeficiente do filtro FIR, $\mathbf{h}$ é o vetor de coeficientes, $(\cdot)^\top$ é o operador de matriz transposta e $\mathbf{x}$ é o vetor com últimas M amostras na saída. O erro ao estimar o sinal desejado é descrito por:

$$e[n] = d[n] - \hat{d}[n] \tag{4.15}$$

$$e[n] = d[n] - \mathbf{h}^\top \mathbf{y} \tag{4.16}$$

Por sua vez, no projeto do filtro de Wiener IIR no domínio do tempo discreto, a saída $\hat{d}[n]$ depende tanto de amostras passadas quanto futuras de $x[n]$, no qual $(*)$ denota a operação de convolução linear:

$$\hat{d}[n] = \sum_{k=-\infty}^{\infty} h_k x[n-k], \quad -\infty < n < \infty \tag{4.17}$$

$$\hat{d}[n] = h[n] * x[n] \tag{4.18}$$

Ao descrever a equação 4.18 no domínio da frequência discreta, obtemos a equação 4.19, em que $\hat{D}[k]$, $H[k]$ e $X[k]$ são as DFTs de $\hat{d}[n]$, $h[n]$ e $x[n]$, respectivamente:

$$\hat{D}[k] = H[k]X[k] \tag{4.19}$$

Em aplicações de *Speech Enhancement*, partirmos das equações 4.1 e 4.15, que descrevem o sinal impregnado pela parcela de ruído e a parcela de erro. No domínio da frequência

$$E[k] = D[k] - \hat{D}[k] \tag{4.20}$$

$$E[k] = D[k] - H[k]X[k] \tag{4.21}$$

Para minimizar o erro quadrático médio que torna o filtro ótimo, obtém-se a expressão do erro:

$$\mathbb{E}\left[\,|E^2[k]|\,\right] = \mathbb{E}\left[\,(D[k] - H[k]X[k]\,)^*\,(\,D[k] - H[k]X[k]\,)\,\right] \tag{4.22}$$

$$\mathbb{E}\left[\,|E^2[k]|\,\right] = \mathbb{E}\left[\,|D^2[k]|\,\right] - H[k]S_{xd}[k] - H^*[k]S_{dx}[k] + |H[k]|^2 S_{xx}[k] \tag{4.23}$$

em que $S_{xx}[k]$, $S_{dx}[k]$ $S_{xd}[k]$ são o espectro de potência de $x[n]$ e os espectros de

potência cruzados de $x[n]$ e $d[n]$, respectivamente:

$$S_{xx}[k] = \mathbb{E}\left[\,|X[k]|^2\,\right] \tag{4.24}$$
$$S_{dx}[k] = \mathbb{E}\left[\,X^*[k]D[k]\,\right] \tag{4.25}$$
$$S_{xd}[k] = \mathbb{E}\left[\,D^*[k]X[k]\,\right] \tag{4.26}$$

No domínio da STFT, obtém-se os ganhos do filtro ao minimizar a equação equivalente à (4.21) [4]:

$$\hat{W}[k,l] = \underset{W}{\operatorname{argmin}}\,\mathbb{E}[\,|D[k,l] - H[k,l]X[k,l|^2\,] \tag{4.27}$$

4.3 *Speech Enhancement* no Domínio da Modulação

4.3.1 Subtração Espectral no Domínio da Modulação

A subtração espectral de modulação é dada por [10]:

$$|\hat{\mathcal{S}}[\ell,k,\mu]|^\gamma = \begin{cases} \Delta[\ell,k,\mu]^{\frac{1}{\gamma}}, & \text{caso } \Delta[\ell,k,\mu] \geq \beta|\hat{D}[l,k,\mu]|^\gamma \\ (\beta|\hat{D}[\ell,k,\mu]]|^\gamma)^{\frac{1}{\gamma}}, & \text{caso contrário} \end{cases} \tag{4.28}$$

tal que

$$\Delta[\ell,k,\mu] = |\mathcal{X}[\ell,k,\mu]|^\gamma - \alpha|\hat{D}[\ell,k,\mu]|^\gamma. \tag{4.29}$$

No domínio da modulação, um VAD pode ser descrito de forma binária para cada *frame*:

$$\Phi[\ell,k] = \begin{cases} 1, & \text{se } \phi[\ell,k] \geq \theta \\ 0, & \text{caso contrário} \end{cases} \tag{4.30}$$

em que l é o índice do *frame* no domínio acústico, k é o índice do *bin* da frequência acústica e θ é o limiar de decisão. Por sua vez, $\phi[\ell,k]$ descreve a razão sinal-ruído do sinal:

$$\phi[\ell,k] = 10\log_{10}\left(\frac{\sum_l |\mathcal{X}[\ell,k,\mu]|^2}{\sum_l |\hat{D}[l-1,k,\mu]|^2}\right) \tag{4.31}$$

em que $\hat{D}[l-1,k,\mu]$ é a estimativa do ruído para o $frame$ anterior. A estimativa do ruído para determinado $frame$ é atualizado durante a ausência de fala, a partir de seu valor para o $frame$ anterior, do espectro de modulação para aquele instante e de um fator de esquecimento λ:

a eq precisa ser no dominio da modulacao? parseval n resolve?

$$|\hat{D}[l,k,\mu]|^{\gamma} = |\hat{D}[l-1,k,\mu]|^{\gamma} + (1-\lambda)|\mathcal{X}[\ell,k,\mu]|^{\lambda}. \tag{4.32}$$

4.3.2 Filtro de Wiener no Domínio da Modulação

4.4 Comparação e Resultados

Capítulo 5

Conclusões

5.1 Avaliação dos Resultados

5.2 Trabalhos Futuros

A bibliografia existente de espectrograma de modulação, até o presente momento, aplica uma segunda STFT sobre o domínio acústico, com a finalidade de medir oscilações de intensidade no eixo da frequência. Dessa forma, o novo eixo obtido pelo espectrograma de modulação representa modulações de amplitude.

Considerando que o espectrograma acústico reside em um espaço vetorial $\mathbb{R}_3$, uma proposta de trabalho futuro é investigar a possibilidade de representar modulações em frequência, bastando que a segunda STFT meça as oscilações no eixo da frequência para uma dada intensidade constante.

Referências Bibliográficas

[1] GAROFOLO, J. S., LAMEL, L. F., FISHER, W. M., et al. "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1", *NASA STI/Recon technical report n*, v. 93, pp. 27403, 1993.

[2] HAYKIN, S., MOHER, M. *Communication Systems*. 5 ed. Nova Iorque, Estados Unidos, John Wiley & Sons, 2008.

[3] LI, Q., ATLAS, L. "Properties for modulation spectral filtering". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*, v. 4, pp. iv–521. IEEE, 2005.

[4] PARCHAMI, M., ZHU, W.-P., CHAMPAGNE, B., et al. "Recent developments in speech enhancement in the short-time Fourier transform domain", *IEEE Circuits and Systems Magazine*, v. 16, n. 3, pp. 45–77, 2016.

[5] POEPPEL, D., ASSANEO, M. F. "Speech rhythms and their neural foundations", *Nature Reviews Neuroscience*, v. 21, n. 6, pp. 322–334, 2020.

[6] VARNET, L., ORTIZ-BARAJAS, M. C., ERRA, R. G., et al. "A cross-linguistic study of speech modulation spectra", *The Journal of the Acoustical Society of America*, v. 142, n. 4, pp. 1976–1989, 2017.

[7] SCHWERIN, B., SO, S. "A comparative study on acoustic and modulation domain speech enhancement algorithms for improving noise robustness in speech recognition". In: *17th Australasian International Conference on Speech Science and Technology (ASSTA)*, pp. 113–116, Sydney, Australia, 2018.

[8] SCHIMMEL, S., ATLAS, L. "Coherent envelope detection for modulation filtering of speech". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*, v. 1, pp. I–221. IEEE, 2005.

[9] DRULLMAN, R., FESTEN, J. M., PLOMP, R. "Effect of temporal envelope smearing on speech reception", *The Journal of the Acoustical Society of America*, v. 95, n. 2, pp. 1053–1064, 1994.

[10] PALIWAL, K., SCHWERIN, B. "Modulation Processing for Speech Enhancement". In: *Speech and Audio Processing for Coding, Enhancement and Recognition*, Springer, pp. 319–345, Nova Iorque, Estados Unidos, 2015.

[11] HOUTGAST, T., STEENEKEN, H. J. "The modulation transfer function in room acoustics as a predictor of speech intelligibility", *Acta Acustica united with Acustica*, v. 28, n. 1, pp. 66–73, 1973.

[12] STEENEKEN, H. J., HOUTGAST, T. "A physical method for measuring speech-transmission quality", *The Journal of the Acoustical Society of America*, v. 67, n. 1, pp. 318–326, 1980.

[13] SCHIMMEL, S. M. *Theory of Modulation Frequency Analysis and Modulation Filtering, with Applications to Hearing Devices*. Tese de Doutorado, University of Washington, Seattle, Estados Unidos, 2007.

[14] ATLAS, L., SHAMMA, S. A. "Joint acoustic and modulation frequency", *Journal on Advances in Signal Processing, European Association for Signal and Image Processing (EURASIP)*, v. 2003, n. 7, pp. 1–8, 2003.

[15] CLARK, C. P. *Coherent Demodulation of Nonstationary Random Processes*. Tese de Doutorado, University of Washington, Seattle, Estados Unidos, 2012.

[16] MARKAKI, M., STYLIANOU, Y. "Using modulation spectra for voice pathology detection and classification". In: *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 2514–2517, Minneapolis, Estados Unidos, 2009.

[17] HERMANSKY, H., MORGAN, N. "RASTA processing of speech", *IEEE transactions on speech and audio processing*, v. 2, n. 4, pp. 578–589, 1994.

[18] KANEDERA, N., ARAI, T., HERMANSKY, H., et al. "On the relative importance of various components of the modulation spectrum for automatic speech recognition", *Speech Communication*, v. 28, n. 1, pp. 43–55, 1999.

[19] KINGSBURY, B. E., MORGAN, N., GREENBERG, S. "Robust speech recognition using the modulation spectrogram", *Speech communication*, v. 25, n. 1-3, pp. 117–132, 1998.

[20] LU, X., MATSUDA, S., UNOKI, M., et al. "Temporal contrast normalization and edge-preserved smoothing of temporal modulation structures of speech for robust speech recognition", *Speech Communication*, v. 52, n. 1, pp. 1–11, 2010.

[21] NADEU, C., PACHÈS-LEAL, P., JUANG, B.-H. "Filtering the time sequences of spectral parameters for speech recognition", *Speech communication*, v. 22, n. 4, pp. 315–332, 1997.

[22] TYAGI, V., MCCOWAN, I., MISRA, H., et al. "Mel-cepstrum modulation spectrum (MCMS) features for robust ASR". In: *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721)*, pp. 399–404. IEEE, 2003.

[23] XIAO, X., CHNG, E. S., LI, H. "Normalization of the speech modulation spectra for robust speech recognition", *IEEE transactions on audio, speech, and language processing*, v. 16, n. 8, pp. 1662–1674, 2008.

[24] ATLAS, L., LI, Q., THOMPSON, J. "Homomorphic modulation spectra". In: *International Conference on Acoustics, Speech, and Signal Processing*, v. 2, pp. ii–761. IEEE, 2004.

[25] GREENBERG, S., KINGSBURY, B. E. "The modulation spectrogram: In pursuit of an invariant representation of speech". In: *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, v. 3, pp. 1647–1650. IEEE, 1997.

[26] HOUTGAST, T., STEENEKEN, H. J. "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria", *The Journal of the Acoustical Society of America*, v. 77, n. 3, pp. 1069–1077, 1985.

[27] PALIWAL, K., WÓJCICKI, K., SCHWERIN, B. "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain", *Speech communication*, v. 52, n. 5, pp. 450–475, 2010.

[28] SO, S., PALIWAL, K. K. "Modulation-domain Kalman filtering for single-channel speech enhancement", *Speech Communication*, v. 53, n. 6, pp. 818–829, 2011.

[29] PALIWAL, K., SCHWERIN, B., WÓJCICKI, K. "Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator", *Speech Communication*, v. 54, n. 2, pp. 282–305, 2012.

[30] MÜLLER, M. *Fundamentals of music processing: Audio, analysis, algorithms, applications*. 1 ed. Nova Iorque, Estados Unidos, Springer, 2015.

[31] LES ATLAS, P. C., SCHIMMEL, S. *Modulation Toolbox Version 2.1 for MATLAB*. Pacote de software, University of Washington, sep 2010. Disponível em: <`https://sites.google.com/a/uw.edu/isdl/projects/modulation-toolbox`>.

[32] DA COSTA, M. D. V. M. *Sistema de Consulta Cantarolada com Geração Automática de um Banco de Músicas Adaptativo*. Tese de Mestrado, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil, 2015.

[33] CLARK, P., ATLAS, L. "A sum-of-products model for effective coherent modulation filtering". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'09)*, pp. 4485–4488. IEEE, 2009.

[34] LATHI, B. P., GREEN, R. A. *Linear systems and signals*. 3 ed. Nova Iorque, Estados Unidos, Oxford University Press, 2018.

[35] CLARK, P., ATLAS, L. "Time-frequency coherent modulation filtering of nonstationary signals", *IEEE Transactions on Signal Processing*, v. 57, n. 11, pp. 4323–4332, 2009.

[36] HAYKIN, S. *Digital communication systems*. 1 ed. Nova Jersey, EUA, Nova Jersey: John Wiley & Sons, 2014.

[37] GOULART, A. J. H. *Efeitos de áudio baseados em decomposição AM/FM*. Tese de Doutorado, Universidade de São Paulo, São Paulo, Brasil, 2017.

[38] COHEN, L., LOUGHLIN, P., VAKMAN, D. "On an ambiguity in the definition of the amplitude and phase of a signal", *Signal Processing*, v. 79, n. 3, pp. 301–307, 1999.

[39] LOIZOU, P. C. *Speech enhancement: theory and practice*. 2 ed. Flórida, Estados Unidos, CRC press, 2013.

[40] ZHANG, Y., ZHAO, Y. "Real and imaginary modulation spectral subtraction for speech enhancement", *Speech Communication*, v. 55, n. 4, pp. 509–522, 2013.

[41] BEROUTI, M., SCHWARTZ, R., MAKHOUL, J. "Enhancement of speech corrupted by acoustic noise". In: *ICASSP'79. IEEE International Conference on Acoustics, Speech, and Signal Processing*, v. 4, pp. 208–211. IEEE, 1979.