UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA DE TELEINFORMÁTICA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE TELEINFORMÁTICA

MESTRADO ACADÊMICO EM ENGENHARIA DE TELEINFORMÁTICA

JACQUES HENRIQUE BESSA ARAÚJO

ATENUAÇÃO DE RUÍDO EM SINAIS DE VOZ UTILIZANDO REDES NEURAIS PROFUNDAS

FORTALEZA

2020

JACQUES HENRIQUE BESSA ARAÚJO

ATENUAÇÃO DE RUÍDO EM SINAIS DE VOZ UTILIZANDO REDES NEURAIS PROFUNDAS

Dissertação apresentada ao Curso de Mestrado Acadêmico em Engenharia de Teleinformática do Programa de Pós-Graduação em Engenharia de Teleinformática do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Engenharia de Teleinformática. Área de Concentração: Sinais e Sistemas

Orientador: Prof. Dr. Paulo César Cortez

FORTALEZA

2020

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

A689a Araújo, Jacques Henrique Bessa.
Atenuação de Ruído em Sinais de Voz utilizando Redes Neurais Profundas / Jacques Henrique Bessa Araújo. – 2020.
118 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Tecnologia, Programa de Pós-Graduação em Engenharia de Teleinformática, Fortaleza, 2020.
Orientação: Prof. Dr. Paulo César Cortez.

1. Processamento de sinais de voz. 2. Redes neurais profundas. 3. Redução de ruído aditivo. 4. Mapeamento espectral não linear. I. Título.

CDD 621.38

JACQUES HENRIQUE BESSA ARAÚJO

ATENUAÇÃO DE RUÍDO EM SINAIS DE VOZ UTILIZANDO REDES NEURAIS PROFUNDAS

Dissertação apresentada ao Curso de Mestrado Acadêmico em Engenharia de Teleinformática do Programa de Pós-Graduação em Engenharia de Teleinformática do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Engenharia de Teleinformática. Área de Concentração: Sinais e Sistemas

Aprovada em: 27 de fevereiro de 2020

BANCA EXAMINADORA

Prof. Dr. Paulo César Cortez (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. Francisco Madeiro Bernardino Junior
Universidade Católica de Pernambuco (UNICAP)

Prof. Dr. João Paulo do Vale Madeiro
Universidade Federal do Ceará (UFC)

Prof. Dr. Yuri Carvalho Barbosa Silva
Universidade Federal do Ceará (UFC)

Dedico este trabalho a Deus, à minha família e a todos que contribuíram direta ou indiretamente para a realização deste trabalho.

AGRADECIMENTOS

Agradeço primeiramente a Deus pelo dom da vida, pela proteção, pelas bênçãos e por me guiar por bons caminhos.

Minha eterna gratidão aos meus pais, Joaquim e Grasças, pelo amor incondicional, pela proteção, pelas orações, pelas bençãos concedidas e pelos anos de dedicação à minha educação e ao meu crescimento.

Meu agradecimento especial à minha irmã Heline, por ser exemplo e fonte de inspiração para o meu desenvolvimento pessoal e profissional.

Ao professor Dr. Paulo César Cortez, por me orientar e contribuir com a realização deste trabalho, por abrir as portas do Laboratório de Engenharia de Sistemas de Computação (LESC), por compartilhar conhecimento e experiências, pelo apoio e paciência e pelas lições na hora do café. Herdo com muita honra, dos nossos anos de convivência, a sua humildade e o seu exemplo de dedicação ao serviço público e à educação e produção científica no Brasil.

Ao professor Dr. Alexandre Augusto da Penha Coelho, pelo apoio, pelos conselhos, pelas oportunidades abertas nos projetos de pesquisa durante minha graduação junto ao LESC, por incentivar meu crescimento profissional para que pudesse me tornar mestre e professor.

Aos meus amigos de longa data André Lopes, Leandro Lopes, Hermeson Costa, Tiago Oliveira, Amanda Veras, Hortência Siebra e Jeissy Oliveira, pela valiosa amizade, pelas palavras de apoio, por me dar forças e coragem nos momentos difíceis dessa caminhada.

Aos meus amigos professores do IFCE, André Camelo, Alexandro Damasceno, David Silveira, Edmílson Moreira e Fabiano Rocha.

Ao corpo docente do Departamento de Engenharia de Teleinformática, pelos ensinamentos ao longo da minha trajetória na graduação e pós-graduação em Engenharia de Teleinformática.

Aos professores participantes da banca examinadora, Prof. Dr. Francisco Madeiro Bernardino Júnior, Prof. Dr. João Paulo do Vale Madeiro e Prof. Dr. Yuri Carvalho Barbosa Silva, pelo tempo e atenção dedicado à revisão deste trabalho, pelas críticas e sugestões de melhoria. Estou certo de que tais contribuições subsidiam meu crescimento e um trabalho de excelência.

Agradeço ainda aos servidores da Universidade Federal do Ceará, em especial aos servidores do Departamento de Engenharia de Teleinformática, pela cordialidade, profissionalismo e suporte às atividades de ensino e aprendizagem.

"I believe that this work is worthwhile simply because any progress toward the solution of a hard scientific problem is always worthwhile for its own sake and because it will add to our bag of tricks for solving other hard problems."

(Arthur Lee Samuel)

RESUMO

A presença de SRAV tem se tornado cada vez mais comum no dia a dia das pessoas e mudado a forma como os seres humanos interagem com os dispositivos. Tais sistemas têm sido utilizados em aplicativos de pesquisa por voz, em assistentes virtuais e eletrônicos, em sistemas de automação residencial e veicular, em jogos e aplicativos de entretenimento, em transcritores de texto, em tradutores instantâneos, em sistemas de reconhecimento de pessoas e de emoções, entre outros. O advento e o crescimento da internet das coisas e a evolução das comunicações móveis potencializam a utilização de tais sistemas, uma vez que dispositivos de entrada e saída comumente utilizados em computadores, como mouse e teclado, se tornam inadequados na interação do ser humano com, por exemplo, eletrodomésticos e sistemas de automação veicular. SRAV proporcionam uma interação homem-máquina natural e conveniente à natureza humana. Nesse sentido, é imprescindível que tais sistemas sejam confiáveis e precisos. No entanto, em condições adversas, altos níveis de ruído, reverberação e sinais interferentes de múltiplas fontes interferem no reconhecimento correto da fala. A taxa de erro de palavra, métrica comumente utilizada em sistemas de reconhecimento automático de voz, está intrinsecamente relacionada a algoritmos de processamento de sinais que minimizam os efeitos ocasionados por tais fatores. Diante dessa problemática, esta dissertação se propõe a investigar a eficiência e as limitações de uma estrutura de rede neural profunda aplicada à atenuação de ruído em sinais de voz em ambientes ruidosos simulados do mundo físico. Para a avaliação de desempenho, utilizam-se quatro métricas objetivas e os resultados são comparados com os algoritmos subtração espectral e filtro de Wiener. Os resultados apontam que a rede neural profunda obteve desempenho superior em todos os cenários de ruído aditivo e relação sinal-ruído quando comparado aos demais algoritmos. Em relação à métrica LSD, o resultado médio da RNP é 36% menor comparado ao filtro de Wiener e 25% menor com relação à subtração espectral. Na métrica STOI, o resultado médio da RNP é 12% superior ao obtido pelo filtro de Wiener e 8% superior ao resultado obtido pela subtração espectral. Na métrica PESQ, a RNP é superior em 17% comparado ao filtro de Wiener e 13% em relação à subtração espectral. Já na métrica WER, a RNP apresenta um resultado 29% menor que o filtro de Wiener e 13% menor que a subtração espectral.

Palavras-chave: Processamento de Sinais de Voz. Redes Neurais Profundas. Redução de Ruído. Mapeamento Espectral.

ABSTRACT

ASR have become increasingly common in people's daily lives and have changed the way humans interact with devices. Voice search applications, virtual and electronic assistants, home and vehicular automation systems, games and entertainment applications, text transcribers, voice translators, people, and emotion recognition systems, among others, are some ASR applications. The advent and growth of the Internet of things and the evolution of mobile communications enhance the use of such systems, since input and output devices commonly used in computers, such as mouse and keyboard, become inadequate in the interaction of the human being with, for example, household appliances and vehicular automation systems. ASR provide natural and convenient human-machine interaction. In this sense, such systems must be reliable and accurate. However, in adverse conditions, high levels of noise, reverberation, and interfering signals from multiple sources interfere with correctly recognizing speech. The word error rate, a metric commonly used in automatic speech recognition systems, is intrinsically related to signal processing algorithms to minimize the effects caused by such factors. This dissertation aims to investigate the efficiency and limitations of deep neural networks applied to noise attenuation in noisy environments. Four objective metrics are used for performance evaluation and the results are compared with those obtained by spectral subtraction algorithms and Wiener filter. The results show that the DNN algorithm obtained the best results in all SNR scenarios when compared to other algorithms. About the LSD metric, the average result of RNP is 36% lower compared to the Wiener filter and 25% lower than the spectral subtraction algorithm. For the STOI metric, the average result of DNN is 12% higher than that obtained by the Wiener filter and 8% higher than that obtained by spectral subtraction. For the PESQ metric, DNN is 17% higher than the Wiener filter and 13% higher than the spectral subtraction. For the WER metric, DNN is 29% lower than the Wiener filter and 13% lower than the spectral subtraction.

Keywords: Speech Signal Processing. Deep Neural Network. Noise Reduction. Spectral Mapping.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ALGORITMOS

LISTA DE ABREVIATURAS E SIGLAS

SRAV	Sistema de Reconhecimento Automático de Voz
NPS	Nível de Pressão Sonora
MRB	Máquina Restrita de Boltzman
RNP	Rede Neural Profunda
STOI	*Short-Time Objective Intelligibility*
TEP	Taxa de Erro de Palavra
SNR	Relação Sinal-Ruído
PESQ	*Perceptual Evaluation of Speech Quality*
HMM	*Hidden Markov Model*
GMM	*Gaussian Mixture Model*
TDF	Transformada Discreta de Fourier
EQM	Erro Quadrático Médio
LIT	linear invariante no tempo
NLMV	negativo do logaritmo da máximo-verossimilhança
GDE	gradiente descendente estocástico
segSNR	*segmental SNR*
DWT	Transformada *Wavelet* Discreta
WPD	Decomposição de Pacotes *Wavelet*
NAT	*noise aware training*
TDFI	Transformada Discreta de Fourier Inversa
LSD	distorção log-espectral
VAD	detector de atividade de voz
PSQM	*Perceputal Speech Quality Measure*
MOS	opinião média subjetiva
WSS	*Weighted Spectral Slope*
LLR	*Log-Likelihood Ratio*
IS	Distância *Itakura-Saito*
API	*application programming interface*
MATLAB	*MATrix LABoratory*
SSD	*Solid-State Drive*

SUMÁRIO

1 INTRODUÇÃO

O reconhecimento automático de voz é uma tecnologia no qual os dispositivos são capazes de extrair informações linguísticas e converter o sinal de voz em uma sequência de palavras correspondente por meio de um conjunto de algoritmos implementados em computadores. Uma vez que a voz é o principal meio de comunicação do ser humano, é extremamente útil que sistemas de computação possam compreender e interpretar a voz e estabelecer uma interface de comunicação tão natural quanto possível (LI *et al.*, 2016).

Este Capítulo apresenta uma introdução sobre tais sistemas. A Seção 1.1 descreve algumas aplicações, uma perspectiva histórica de pesquisa e desenvolvimento, as principais tendências do mercado para os próximos anos e alguns dos desafios desses sistemas. A Seção 1.2 apresenta a delimitação do escopo deste trabalho em relação a SRAVs e a descrição dos desafios no contexto do aprimoramento de sinais de voz.A Seção 1.3 apresenta a formulação do problema do ruído aditivo. A Seção 1.4 esclarece as motivações para o desenvolvimento desta dissertação. Os objetivos são apresentados na Seção 1.5 e as principais contribuições são especificadas na Seção 1.6. Por fim, a Seção 1.7 descreve a organização e estrutura desta dissertação.

1.1 Sistemas de reconhecimento automático de voz

O reconhecimento de voz está presente em diversos filmes de ficção científica, como Star Trek (1966), 2001: Uma Odisseia no Espaço (1968), Star Wars: Uma Nova Esperança (1977), Eu, Robô (2004) e WALL-E (2008), ao mostrar a interação entre homens e máquinas. Até alguns anos atrás, SRAVs eram incomuns, limitados e pouco populares devido às restrições de tecnologia.

Cada vez mais presente no dia a dia do ser humano, o reconhecimento automático de voz vem sendo incorporado gradualmente em uma vasta quantidade de aplicações e dispositivos. Assistentes virtuais como o Microsoft Cortana, o Amazon Alexa, o Apple Siri, o Hound, o Nuance Dragon, o IBM Watson e o Google Assistente, presentes em celulares, tablets, relógios, alto-falantes, óculos, jogos, aplicativos de entretenimento, eletrodomésticos, carros e outras dezenas de dispositivos, utilizam SRAVs. Tais sistemas são usados para reconhecer, interpretar elocuções humanas e instruir os sistemas de computação a executar comandos úteis como: realizar pesquisas e ligações; agendar reuniões; tocar músicas; ligar lâmpadas; informar sobre as condições relacionadas ao trânsito; entre outras centenas de comandos. Também estão presentes

em: centrais telefônicas, reconhecendo, colhendo informações e interpretando comandos de voz para agilizar o atendimento aos usuários; jogos eletrônicos, entendendo comandos de voz e instruindo personagens a realizar determinadas ações; e aplicativos de tradução simultânea.

Outras aplicações em que a digitação, a transcrição ou a interface de comunicação sejam morosas podem se valer dos benefícios desta tecnologia. Em consultórios, por exemplo, os médicos empregam boa parte do tempo das suas consultas digitando ou escrevendo informações sobre o histórico, diagnóstico e tratamento do paciente. Nesse caso, um assistente de voz pode auxiliar no processo de transcrição das informações, tanto do paciente quanto do próprio médico, permitindo que ele possa dar mais atenção e cuidado ao paciente. Com isso, a eficiência do diagnóstico pode aumentar concomitantemente à redução do tempo de atendimento. SRAVs também permitem que portadores de deficiência visual utilizem os dispositivos móveis com autonomia. Assistentes de voz em carros podem minimizar os riscos relacionados à segurança ao permitir a transcrição de voz, ao procurar uma lista de músicas e ao orientar o sistema a determinar o melhor caminho de navegação para um certo destino.

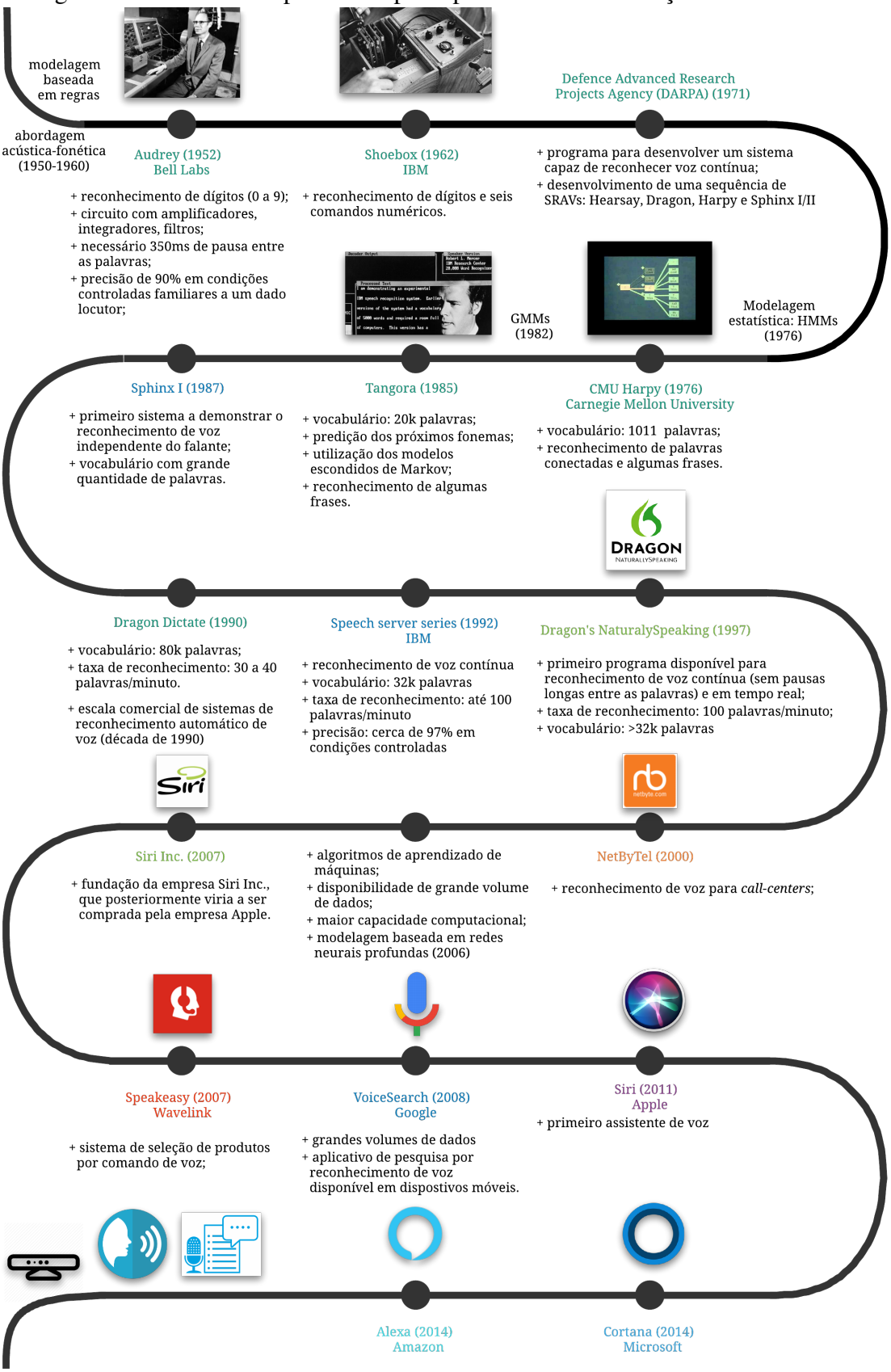
1.1.1 Perspectiva histórica

O reconhecimento automático de voz possui cerca de 60 anos de pesquisa e desenvolvimento. Do Audrey (DAVIS *et al.*, 1952), considerado o primeiro sistema de reconhecimento de voz, ao SIRI[1], atualmente um dos mais avançados assistentes de voz, inúmeros problemas foram solucionados e/ou minimizados para alcançar o nível de precisão e interatividade disponíveis atualmente. A Figura 1 apresenta uma linha do tempo com uma série de fatos relacionados ao desenvolvimento de SRAVs nos últimos 60 anos.

Em 1952, Bell Laboratories lançou o Audrey, considerado o primeiro reconhecedor de voz. Este era capaz de reconhecer dígitos, em condições controladas proferidos por seu inventor, com uma precisão de 90%. Em 1962, a IBM desenvolveu um dispositivo capaz de reconhecer e diferenciar 16 palavras e realizar cálculos com dígitos de 0 a 9. Na década seguinte, cientistas da Universidade Carnegie Mellon criaram o sistema Harpy, um SRAV capaz de reconhecer 1011 palavras, quantidade que representa o vocabulário de uma criança de três anos (JUANG; RABINER, 2005).

[1] Abreviatura do inglês *Speech Interpretation and Recognition Interface*

Figura 1 – Linha do tempo com os principais marcos da evolução dos SRAVs.



Fonte: elaborado pelo autor, baseado em informações de (IBM, 2015; HUANG *et al.*, 2014; BENESTY *et al.*, 2007; FURUI, 2005, tradução nossa).

Segundo Reddy (1976), sistemas de reconhecimento de voz rudimentares como o Harpy possuíam diversas limitações: reconhecimento somente de palavras isoladas, dificuldade de estabelecer conexão entre as palavras, vocabulário pequeno, tempo de resposta elevado e reconhecimento apenas em cenários controlados. Tais dificuldades representavam uma frustração de experiência para o usuário.

De 1976 a 2005, a maioria dos sistemas de reconhecimento automático de voz se baseavam na regra de Bayes, no modelo oculto de Markov[2], e no modelo de misturas gaussianas[3]. Em 1997, o sistema *Dragon Naturally Speaking* abriu um precedente ao ser capaz de transcrever algumas frases, ainda que em condições controladas, da voz de um usuário comum com 95% de precisão.

A partir de 2005, o reconhecimento de voz teve um progresso significativo com os avanços no campo do aprendizado de máquinas. Em 1969, John Pierce, da Bell Labs, profetizou os avanços dos SRAVs em um artigo no Jornal da Sociedade Acústica da América:

> O reconhecimento da fala não será possível até que a inteligência e a competência linguística de um falante humano possam ser incorporadas à máquina. (YOUNG, 2010, tradução nossa).

De fato, a utilização de redes neurais artificiais impulsionou o estado da arte em reconhecimento de voz e acelerou o crescimento da tecnologia graças à sua capacidade de abstrair e armazenar informações das características dos sinais de fala. Algoritmos de aprendizado de máquina estatísticos possuem a capacidade de processar e armazenar, de forma abstrata, as características do sinal de voz como timbre, a velocidade da fala, o sotaque e a forma como as palavras são conectadas. Outras informações complementares acerca do histórico de desenvolvimento de SRAVs podem ser encontradas em (IBM, 2015; HUANG *et al.*, 2014; BENESTY *et al.*, 2007).

1.1.2 Tendências de mercado

Diversas empresas vem adotando inovações em seus produtos baseado em SRAVs, o que favorece o crescimento desse mercado. Há uma tendência de expansão nos próximos anos dado o potencial que esta tecnologia oferece e os avanços que têm sido realizados. De acordo com o site Global Industry Global Industry AnalystsInc (2018), o mercado global de aplicativos de reconhecimento de voz deverá atingir cerca de US$ 19,6 bilhões em 2024, impulsionado

[2] Expressão traduzida do inglês *Hidden Markov Model* (HMM).
[3] Expressão traduzida do inglês *Gaussian Mixture Model* (GMM).

pelo desenvolvimento de tecnologias relacionadas à computação que aumentam a precisão e a versatilidade do uso em aplicações industriais, empresariais e do dia a dia.

Dados disponibilizados pelo site Pew Research Center (2017) afirmam que quase metade dos americanos utilizavam assistentes de voz, principalmente em celulares. As principais razões para o uso dessa tecnologia foram: (1) praticidade de utilizar o dispositivo sem o uso das mãos; (2) diversão; (3) sensação de liberdade ao utilizar a voz como um meio natural de comunicação; (4) e facilidade de utilização pelas crianças. Além disso, a estimativa é de que 300 milhões de carros estejam equipados com SRAVs até 2024, segundo dados do site IHS Markit (2019). Em 2018, esse número era de apenas 40 milhões.

1.1.3 Desafios

A interface de reconhecimento de voz tem se tornado atraente, em detrimento de métodos convencionais de interface de comunicação, à medida que os SRAVs oferecem respostas mais rápidas, precisas e intuitivas. Contudo, mesmo diante dos avanços dos últimos anos, a interface de comunicação entre o homem e o computador possui inúmeros desafios até alcançar o aspecto de linguagem natural, a ponto de o ser humano utilizar um dispositivo com sistema de reconhecimento de voz de forma instintiva. Por conta disso, SRAVs permanecem como área ativa de pesquisa e desenvolvimento devido às inúmeras limitações práticas para oferecer uma capacidade de entendimento e comunicação homem-máquina semelhante ou até superior à do ser humano.

Atualmente as principais pesquisas se concentram no entendimento das estratégias do sistema auditivo humano para o processamento do sinal de voz em cenários multi-adversos, no desenvolvimento de algoritmos de processamento de sinais eficientes para aprimorar o sinal de voz, na construção de modelos capazes de aumentar a capacidade dos dispositivos de reconhecer uma ampla variedade de palavras, idiomas, sotaques, entonações, dialetos, figuras de linguagem e a variabilidade dos falantes. De acordo com Huang *et al.* (2014), outro desafio diz respeito à capacidade do sistema detectar, de forma confiável, uma palavra desconhecida ou confusa. Outras pesquisas se concentram ainda nos estudos da fonética-acústica, da percepção de voz, da linguística, da psicoacústica, da semântica e do contexto.

1.2 Melhoria da qualidade de sinais de voz

No escopo do ambiente acústico, as principais limitações práticas estão relacionadas ao ruído, ao eco, à reverberação, à distância do sensor para a fonte e aos sinais interferentes. Tais condições são inerentes à maioria dos ambientes de utilização desses sistemas e acarretam uma alta variabilidade nos sinais de voz.

O ruído aditivo se refere a quaisquer perturbações indesejadas que se sobrepõem ao sinal de voz. O ambiente acústico pode variar de um local quieto (ex. sala com isolamento acústico em silêncio) a um ambiente extremamente ruidoso (ex. uma avenida). Nesta última situação, o ser humano apresenta um desempenho muito superior aos SRAVs, pois é capaz de tolerar o ruído do ambiente e distinguir as palavras (mesmo os fonemas não vozeados) do ruído de fundo.

O eco é decorrente das reflexões sonoras do ambiente e/ou do acoplamento entre microfones e alto-faltantes. Quando os sinais refletidos atingem um observador com atrasos (tempo total de ida e volta) maiores do que sua duração, é possível distinguir entre os sinais emitido e os refletidos. Por esta razão, os ecos são perceptíveis somente em grandes recintos (MCLOUGHLIN, 2016, p.101).

Diferentemente do eco, na reverberação, o intervalo de tempo entre o sinal original e o sinal refletido decorrente dos múltiplos percursos refletidos no ambiente não é suficiente para diferenciar os dois sinais. Os sinais interferentes, por sua vez, são oriundos de outras fontes de som (BENESTY, 2018).

Já a distância entre cada fonte e o sensor (microfone) degrada a relação sinal-ruído à medida que a onda sonora perde a intensidade ao se propagar pelo ar e se dispersar por todas as direções. Consequentemente, a precisão do reconhecimento de voz à distância (acima de 1 metro) é consideravelmente inferior ao de campo próximo (menor que 1 metro). Devido a isso, atualmente é necessário aproximar os dispositivos para um reconhecimento mais preciso.

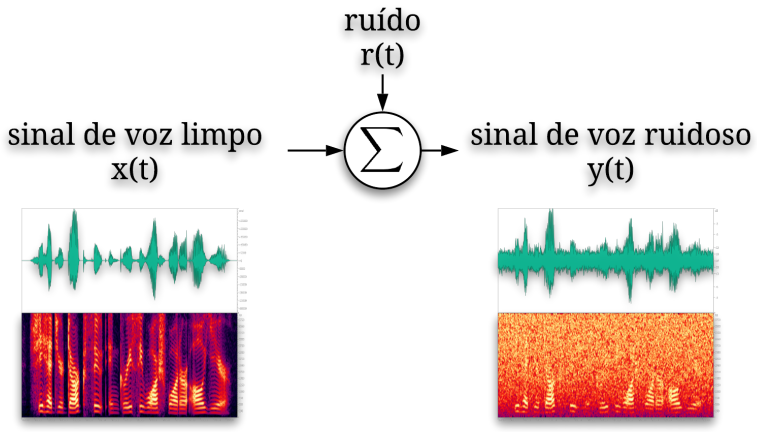
Solucionar esses desafios é tema de pesquisa e desenvolvimento de técnicas de melhoria da qualidade de sinais de voz[4]. Esta subárea de processamento de sinais é uma das mais atrativas e desafiadoras. Nela estão inclusas categorias como: atenuação de ruído aditivo, cancelamento de eco, de-reverberação e separação de fontes. Esta dissertação se concentra na solução do problema do ruído aditivo.

[4] Expressão traduzida do inglês *speech enhancement* que significa melhorar algum aspecto perceptivo da voz que tenha sido degradado. Segundo Xu *et al.* (2014), o objetivo do *speech enhancement* é melhorar a qualidade e a inteligibilidade dos sinais de voz degradados por condições adversas.

1.3 Formulação do problema

O ruído aditivo é um dos principais fatores que afetam a precisão dos SRAVs em ambientes do dia a dia. Ruído aditivos são aqueles que se sobrepõem ao sinal de interesse e estão presentes na nossa vida diária: nas ruas (ex. trânsito), nos carros (ex. motor, vento, ar condicionado, atrito do pneu com o asfalto), nos restaurantes e centros de compras (ex. pessoas conversando, celulares tocando, música, objetos caindo), entre outros lugares. Nesse contexto, o problema pode ser formulado conforme ilustrado na Figura 2.

Figura 2 – Modelo do sinal de voz composto por ruído aditivo.



Fonte: elaborado pelo autor.

Seja $x(t)$ o sinal de voz limpo e $r(t)$ o ruído aditivo. Matematicamente, o sinal de voz ruidoso pode ser modelado através da seguinte equação

$$y(t) = x(t) + r(t). \tag{1.1}$$

A forma de onda do sinal de voz limpo no domínio do tempo, correspondente à frase em inglês "*She had your dark suit in greasy wash water all year*", e seu respectivo espectrograma[5] são ilustrados nas Figuras 3a e 3b, respectivamente. Observa-se que a maior energia do sinal ("curvas" vermelhas e amarelas dispostas majoritariamente na horizontal) se concentra nos períodos em que as palavras são proferidas pelo locutor.

[5] Recurso visual utilizado para analisar a densidade espectral de potência. A variação de cores do azul para o vermelho representa a variação da quantidade de energia (menor para maior), em dB/Hz, de um determinado intervalo de tempo e frequência.

Em geral, SRAV apresentam taxas de erros de palavras baixas para elocuções com baixos de níveis de ruído. Contudo, para um sinal ruidoso[6] como o ilustrado na Figura 4, o reconhecimento correto das palavras se torna mais complexo e desafiador. O envelope do sinal no domínio do tempo e o espectrograma do sinal ruidoso são bem diferentes do sinal limpo. A amplitude do sinal ruidoso possui alta variabilidade, tanto em períodos de silêncio, como nos períodos em que as palavras são proferidas. No gráfico da Figura 4b, o espectrograma apresenta concentração de alta energia (-80 a -40 dB), principalmente na faixa de frequência entre 100 Hz e 1 kHz. Tais diferenças afetam a percepção de qualidade e inteligibilidade de um ouvinte e dificultam a transcrição correta de um SRAV.

1.4 Motivação

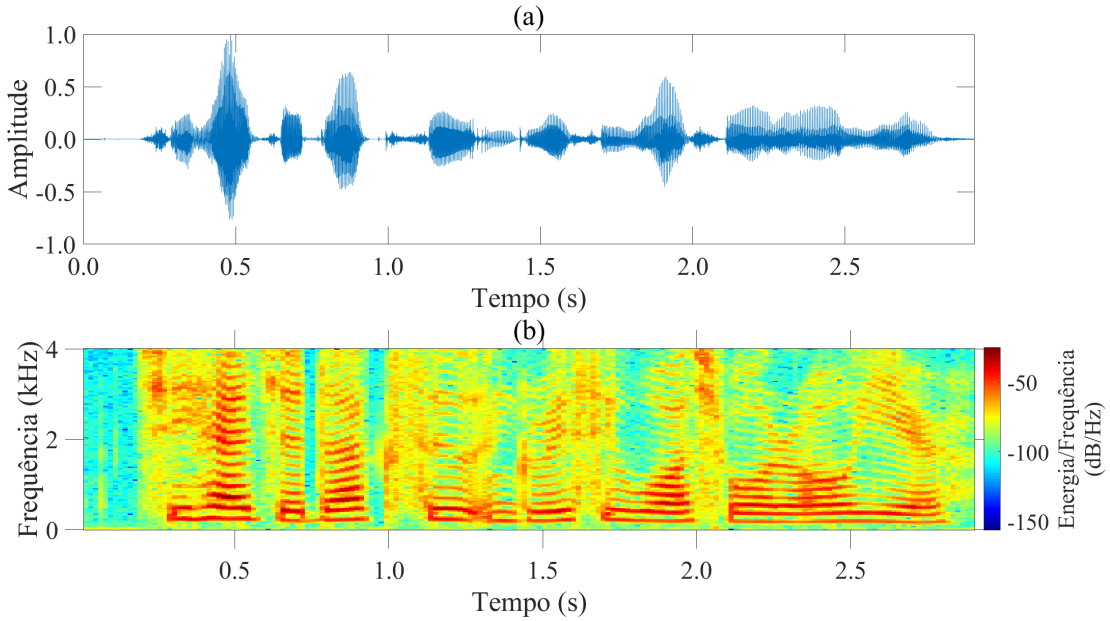
Diante da difusão de utilização de sistemas de reconhecimento automático de voz e o crescimento do mercado, é essencial que a taxa de erro de palavra seja a menor possível, mesmo em ambientes desfavoráveis ao seu uso. A utilização de técnicas de processamento digital de sinais tem o potencial de reduzir os níveis de ruído aditivo, estimar os sinais de voz limpos[7] a partir dos ruidosos e aumentar os níveis de qualidade e inteligibilidade.

Diante desta problemática, a principal motivação para o desenvolvimento desta dissertação consiste em utilizar técnicas de processamento de sinais que minimizem o efeito provocado pelo ruído em diferentes cenários e ambientes acústicos simulados do mundo físico típicos de utilização de SRAVs e que reflitam na redução da taxa de erro de palavras (TEP).

[6] Os tipos de ruído aditivos serão abordados na Seção 2.2.

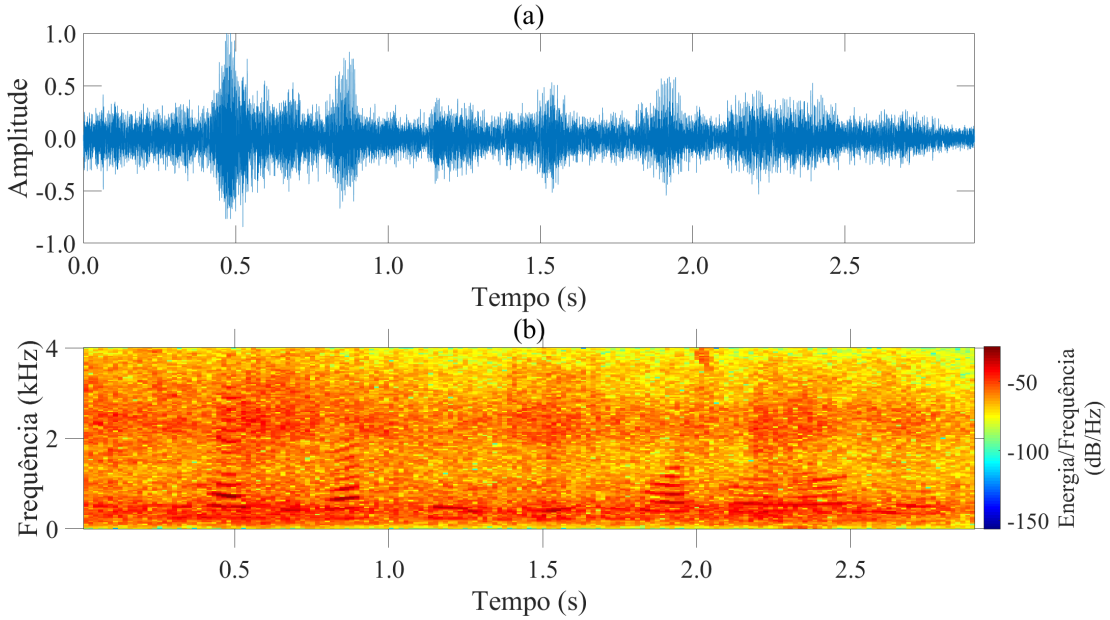
[7] Por sinal de voz limpo, entende-se sinais com baixos níveis de ruído e baixos níveis de sinais interferentes que não comprometam a qualidade e a inteligibilidade das elocuções.

Figura 3 – (a) Forma de onda de um sinal de voz limpo no domínio do tempo e (b) seu espectrograma correspondente.



Fonte: elaborado pelo autor.

Figura 4 – (a) Forma de onda do mesmo sinal da Figura 3 corrompido com ruído balbucios (SNR = 0 dB) no domínio do tempo e (b) seu espectrograma correspondente.



Fonte: elaborado pelo autor.

1.5 Objetivos

Geral

O principal objetivo desta dissertação é aplicar algoritmos de processamento digital de sinais e aprendizado de máquina para redução dos níveis de ruído aditivo e, consequentemente, aumentar a qualidade e a inteligibilidade de sinais de voz ruidosos.

Específicos

Para alcançar o objetivo principal, estabelecem-se os seguintes objetivos específicos:

1. banco de dados de treinamento e teste com sinais de voz degradados com diferentes tipos de ruído e diferentes níveis de SNR;
2. técnica de extração de características do sinal de voz baseada na magnitude do espectro de frequência;
3. técnica de rede neural profunda (RNP) para realizar o mapeamento espectral não linear entre o sinal de voz ruidoso e o sinal de voz limpo;
4. aplicação de técnicas de regularização da RNP;
5. aplicação de técnicas de otimização para o treinamento da RNP;
6. filtragem do ruído aditivo em sinais de voz utilizando a técnica de filtro de Wiener;
7. filtragem do ruído aditivo em sinais de voz utilizando a técnica de subtração espectral.

1.6 Contribuições

Este trabalho produziu um artigo, intitulado *Exploratory Analysis of Deep Neural Networks Applied to Speech Enhancement*, aceito no evento científico *2020 3rd International Conference on Communication System, Computing and IT Applications* (CSCITA), Mumbai, India.

As principais contribuições técnicas são:

- técnica de extração de características de sinais de voz;
- caracterização do uso de redes neurais profundas aplicada à redução de ruídos em sinais de voz em cenários práticos simulados do mundo físico;
- utilização de SNRs aleatórias para o conjunto de dados de treinamento e conjunto de dados de teste da RNP;
- uso de um sistema de reconhecimento automático de voz (SRAV) para avaliar a melhoria na qualidade e na inteligibilidade dos sinais de voz limpos estimados através da métrica taxa de erro de palavra (TEP).

1.7 Organização

Esta dissertação está organizada em seis capítulos. Este Capítulo apresentou uma visão sistemática sobre o assunto, a formulação do problema, a motivação para resolvê-lo, os objetivos e as principais contribuições desta pesquisa. O Capítulo 2 apresenta a fundamentação teórica e os principais conceitos relacionados ao sinais de voz e ruído, ao processo de extração de características e às técnicas utilizadas para a redução de ruído em sinais de voz. São apresentados os modelos de subtração espectral, filtro de Wiener e a rede neural profunda. O Capítulo 3 apresenta uma revisão da literatura e dos principais trabalhos relacionados à esta dissertação. O Capítulo 4 descreve a metodologia utilizada para o desenvolvimento deste trabalho. A visão geral do sistema, o banco de dados, os parâmetros utilizados na rede neural profunda, as técnicas de regularização e otimização, a reconstrução do sinal e as métricas de desempenho, são apresentados neste Capítulo. Os resultados dos algoritmos subtração espectral, filtro de Wiener e rede neural profunda são apresentados e discutidos no Capítulo 5. Finalmente, no Capítulo 6 estão contidas as deduções lógicas fundamentais, as conclusões acerca do que foi desenvolvido e os encaminhamentos para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Neste Capítulo é apresentada a fundamentação teórica utilizada para o desenvolvimento desta dissertação. A Seção 2.1 descreve o processo de produção de voz e as características dos sinais de voz. Os tipos de ruído aditivo e os níveis de amplitude do sinal de voz e ruído em vários ambientes são apresentados na Seção 2.2. As técnicas utilizadas para extrair as características dos sinais (segmentação, sobreposição, janelamento, soma e sobreposição de quadros, transformada discreta de Fourier) são detalhadas na Seção 2.3. A técnica subtração espectral é apresentada na Seção 2.4. Já a Seção 2.5 apresenta os conceitos relacionados ao filtro de Wiener. Finalmente, a fundamentação da rede neural profunda é apresentada na Seção 2.6.

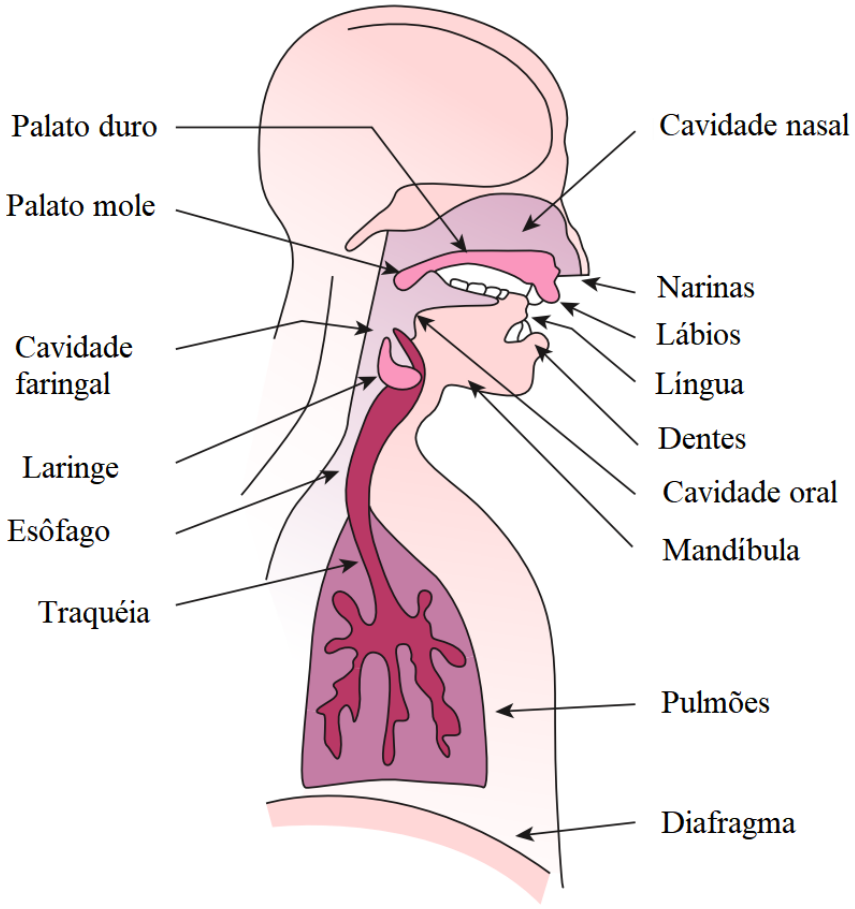
2.1 Caracterização do sinal de voz

O processo de produção vocal (fonação) se dá a partir da vibração das pregas vocais produzida (localizadas na laringe) pela saída do fluxo de ar dos pulmões. Quando não há produção de voz, as pregas vocais ficam abertas e relaxadas e o fluxo de ar passa sem produzir nenhuma vibração. Já para a produção de voz, a pressão respiratória exercida pelo diafragma direciona o fluxo de ar dos pulmões para a traqueia em direção à laringe. Em seguida, as pregas vocais se unem e a passagem do fluxo de ar proveniente provoca uma vibração nas pregas. Quando estão esticadas e unidas, as pregas vocais produzem sons mais agudos. Quando estão mais afastadas e distendidas, produzem sons mais graves. O som básico produzido pelas pregas vocais na laringe (fonte glótica) passa por uma série de cavidades ressonantes (laringe, faringe, boca e nariz) que amplificam o som e ganha forma quando articulado através de movimentos da língua, lábios, mandíbula, dentes e palato (D'AVILA, 2005; BOONE *et al.*, 2014). Uma visão geral das partes do corpo humano envolvidas na produção da voz é apresentada na Figura 5.

Sobre a produção da voz, Pignatari e Anselmo-Lima (2018) fornecem os seguintes conceitos:

> A produção dos sons vocais ocorre por sons gerados pelo sistema fonador, quando as pregas vocais estiverem em vibração ou não, ou mais precisamente pelo fluxo de ar pulmonar que é modificado pelas pregas vocais em vibração e depois pelo trato vocal e por vezes também pela cavidade nasal. O conceito de voz também está associado às pregas vocais em vibração e ao efeito do trato vocal sobre o som produzido pela vibração das pregas vocais [...] Para produzir a voz, dependemos do ar dos pulmões, da vibração das prega vocais e do trato vocal que moldará o som, composto pela laringe, faringe, boca, lábios e língua. (PIGNATARI; ANSELMO-LIMA, 2018).

Figura 5 – Diagrama esquemático do mecanismo vocal humano.



Fonte: (ABHANG *et al.*, 2016, p. 12, tradução nossa).

Segundo Hill (2019), a teoria da produção de voz Fonte-Filtro fornece os fundamentos para os sistemas de codificação e reconhecimento de voz. Esta teoria foi produzida por Johannes Müller, em 1848, através de um conjunto de experimentos com cadáveres humanos. Fant (1970) produziu uma versão moderna e que tem sido amplamente aceita. De acordo com esta teoria, a fonte produz sons vozeados, a partir da vibração das pregas vocais. O espectro de som vozeado contém energia na frequencial fundamental f_0 da vibração das pregas vocais e nas harmônicas associadas (i.e., frequências que são múltiplos inteiros ou frações racionais da frequência fundamental). Por outro lado, as vibrações de sons surdos (não vozeados) são causados pela vibração da passagem de ar através das partes estreitas do trato vocal. Tais sons são caracterizados por ocuparem uma ampla faixa do espectro de frequência (HILL, 2019). O filtro, por sua vez, é o efeito da variação temporal do trato vocal (laringe, faringe, boca, lábios e língua) sobre o sinal de saída da fonte glótica e uma combinação de efeitos ressonantes. Portanto, segundo essa teoria, o formato da onda de saída do trato vocal pode ser modelado como a

convolução da fonte e do filtro, isto é, o som de saída é produzido pelos sinais da fonte vozeados e não vozeados convolvidos com o efeito do filtro do trato vocal variante no tempo. Assim, o espectro da voz tem características tanto da fonte (pregas vocais) quanto do filtro (trato vocal).

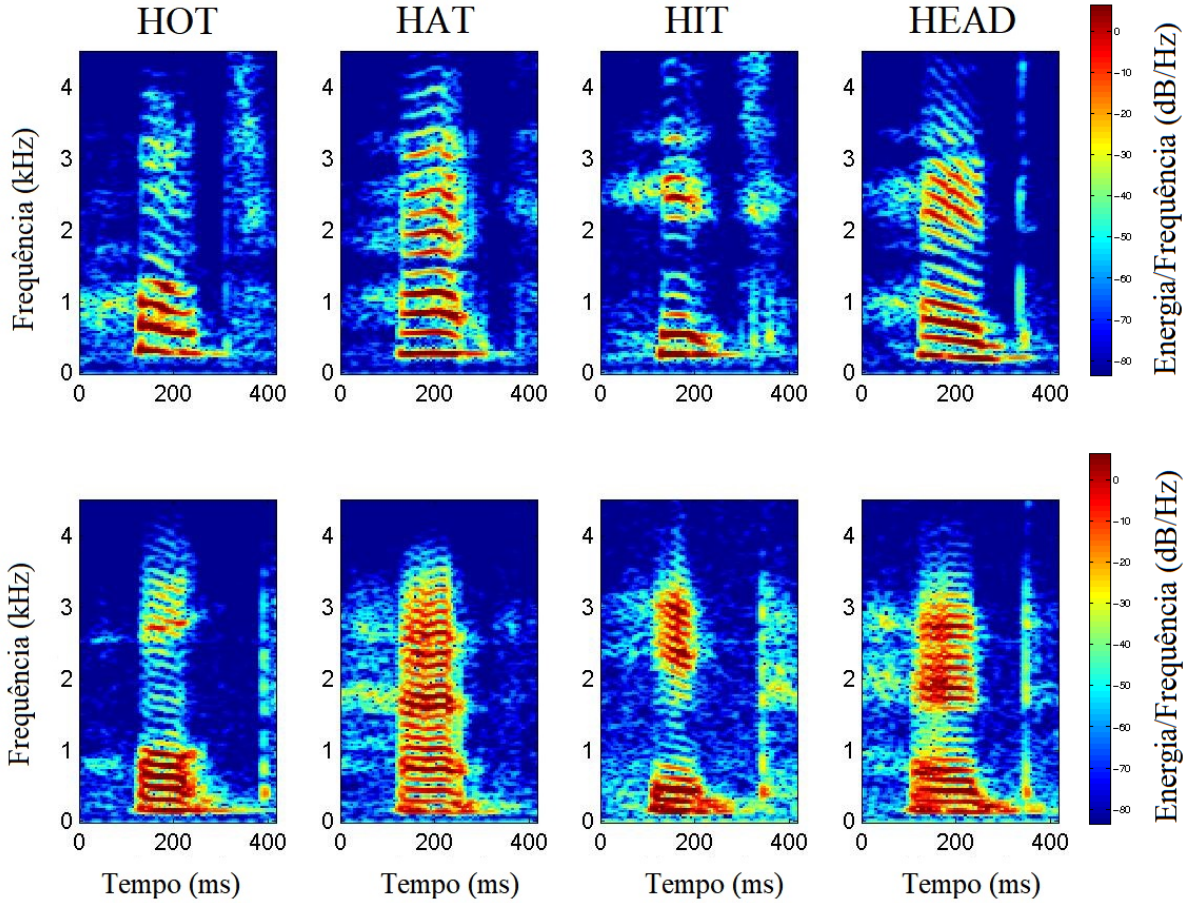
Algumas das componentes harmônicas provenientes das pregas vocais, ao chegar às cavidades de ressonância, possuem compatibilidade com a frequência do trato vocal, cujas frequências naturais de ressonância do trato vocal recebem a denominação de formante. Quando a componente harmônica proveniente das pregas vocais coincide com a frequência formante do trato vocal, esta componente é amplificada. Em outras palavras, o alinhamento dos formantes com as componentes harmônicas intensifica a intensidade do som. O formante com a frequência mais baixa é denominado f_1, o segundo f_2 e assim por diante. Em análise acústica, os primeiros cinco formantes são os de maior interesse, sendo que os três primeiros são responsáveis pela identidade das vogais (GUSMÃO *et al.*, 2010).

Os espectrogramas das pronúncias das palavras *hot*, *hat*, *hit* e *head*, no idioma inglês, faladas por uma mulher (os quatro espectrogramas da linha superior) e por um homem (os quatro espectrogramas da linha inferior) são ilustrados na Figura 6. Observa-se que cada vogal possui um conjunto de harmônicas com intensidades diferentes (filamentos horizontais). Por exemplo, a vogal */a/* tem mais energia na frequência 1,8 kHz do que as vogais */o/* e */i/*. Sobre os aspectos relacionados à Figura 6, Schnupp *et al.* (2011) comentam:

> As vogais são facilmente visíveis nas 'pilhas harmônicas' que marcam a chegada do trem de pulso glótico e o espaçamento dos harmônicos é claramente maior nas vogais pronunciadas pela mulher (frequências altas). Quando em alguns sons os harmônicos não estão dispostos exatamente na horizontal, diz-se que o espaçamento destas vogais não é perfeitamente estável. A localização exata dos formadores nos espectrogramas das vogais não é aparente. Contudo, está claro que, por exemplo, no fonema /i/, as componentes harmônicos em torno de 500 Hz são expressivas (destacadas em vermelho). Há pouca energia sonora entre 800 e 2000 Hz e outros picos em cerca de 2,3 kHz e 3 kHz. Já no fonema /a/, a energia é mais uniformemente distribuída pela faixa de frequência, com picos em torno de 800 Hz, 1,8 kHz e 2,5 kHz. Já para o fonema /o/, o som tem muita energia até cerca de 1,1 kHz e muito menos no restante da faixa espectral até atingir outro pico menor em torno de 2,5 kHz. (SCHNUPP *et al.*, 2011, p.37, tradução nossa).

Ainda de acordo com Schnupp *et al.* (2011), o ser humano controla as frequências formantes através da movimentação dos articuladores do trato vocal, incluindo lábios, língua, mandíbula e palato mole. Mudar o tamanho e o formato das cavidades ressonantes no trato vocal modifica as frequências ressonantes, i.e., os formantes. É através do controle das frequências ressonantes que o ser humano controla qual vogal produz. Já para alterar o tom, alteram-se as frequências harmônicas a partir das movimentações das pregas vocais.

Figura 6 – Espectrogramas das palavras *hot*, *hat*, *hit* e *head* (em inglês), pronunciadas por uma mulher (predominância de altas frequência, i.e., alto tom) (linha superior) e por um homem (predominância de baixas frequências, i.e., baixo tom) (linha inferior).



Fonte: (SCHNUPP *et al.*, 2011, p. 37, tradução nossa).

A respeito da movimentação dos músculos articulatórios do sistema vocal humano, McLoughlin (2016) afirma que vários músculos que produzem a voz são relativamente lentos em comparação à capacidade de processamento dos computadores, resultando em variações lentas nas características espectrais do sinal de voz. Contudo, um sinal de voz é não estacionário, pois suas características espectrais se modificam conforme as palavras/frases são proferidas. Uma regra utilizada em processamento digital de sinais de voz é considerar que um segmento de voz (quadro) com um intervalo entre 20 e 40 ms é quase estacionário, isto é, as propriedades do sinal nesse intervalo de tempo apresentam pouca variação.

2.2 Ruído aditivo

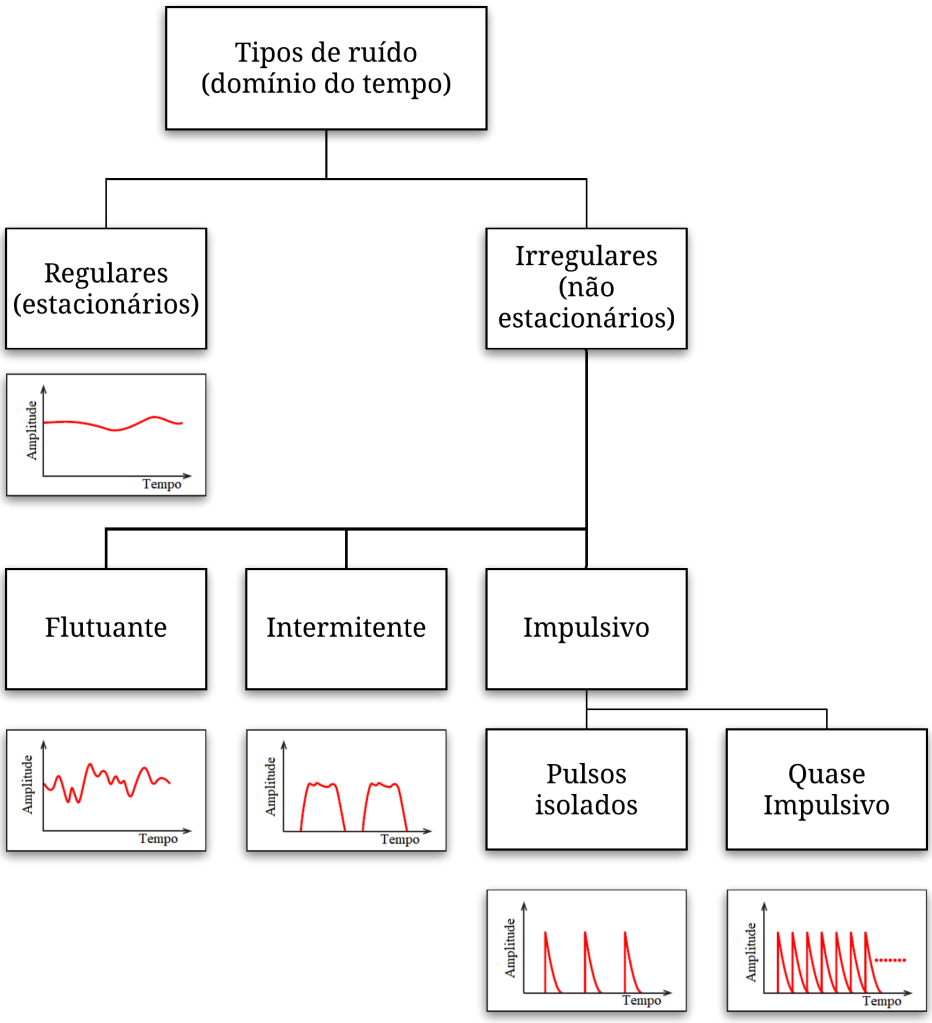
O ruído aditivo é um sinal aleatório e diz respeito a quaisquer perturbações indesejadas que se somam ao sinal de voz. Há vários tipos de ruído aditivo em termos de características temporais e espectrais, níveis de amplitude do sinal de voz e do ruído em alguns ambientes habituais de utilização de SRAVs.

2.2.1 Tipos de ruído aditivo

Segundo a ISO (2010), que dispõe sobre os ruídos emitidos por máquinas e equipamentos, pode-se classificar os ruídos no domínio do tempo em duas categorias: ruído regular (ou quase estacionário) e irregular (não estacionário). O ruído regular possui pequenas flutuações negligenciáveis na amplitude no período de observação, enquanto o ruído irregular apresenta variações significativas de amplitude. Este último pode ser classificado em três categorias: ruído irregular flutuante, ruído irregular intermitente e ruído irregular impulsivo. O ruído irregular flutuante é aquele cuja amplitude varia de forma contínua, de forma regular ou irregular, em um período considerável da observação (ex. tráfego de veículos). O ruído irregular intermitente é aquele cujo nível cai abruptamente para o ruído de fundo várias vezes durante o período de observação (ex. passagem de um trem ou avião). A intermitência do nível varia de ambiente para ambiente e é da ordem de 1 s ou mais. Já o ruído impulsivo consiste numa série de pulsos de energia e é classificado em duas categorias: ruído irregular com pulsos de energia isolados e ruído irregular quase impulsivo. No ruído irregular com pulsos de energia isolados, os intervalos entre os pulsos individuais são maiores do que 0,2 s. Já no ruído irregular quase impulsivo, os intervalos entre os pulsos são menores do que 0,2 s (ISO, 2010). A Figura 7 apresenta a hierarquia de classificação no domínio do tempo segundo a ISO 12.001.

Pode-se estender tal hierarquia de classificação de sinais de ruído para demais aplicações. Para fins de simplificação do problema, trabalha-se com a classificação de alto nível e que é comumente utilizada na literatura (LOIZOU, 2013; MIHOV *et al.*, 2009): estacionário e não estacionário. Sinais estacionários são aqueles que mantêm a variação de amplitude constante (e possivelmente as características espectrais como o tom e o timbre) ao longo do tempo. Uma nota musical, balbucios de pessoas conversando e o ruído provocado pelo vento de um ventilador ou de um ar-condicionado são alguns exemplos de ruídos estacionários. Já os sinais não estacionários (isto é, sinais com resposta impulsiva) são aqueles que apresentam

Figura 7 – Tipos de ruído de acordo com o comportamento dos sinais no domínio do tempo.



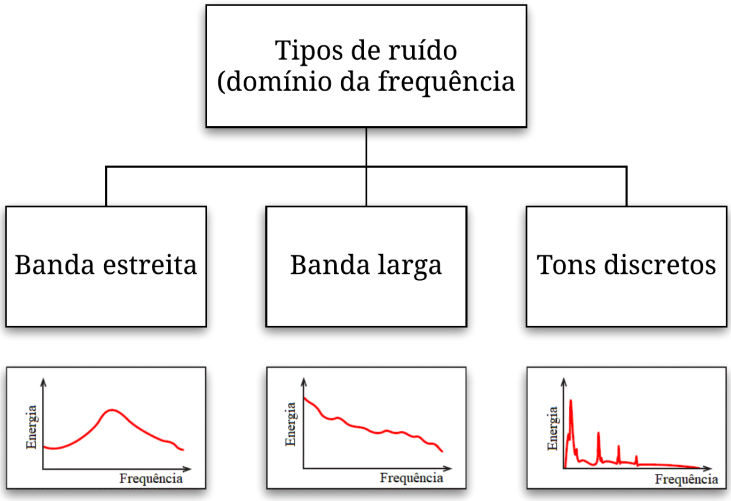
Fonte: elaborado pelo autor, baseado em informações de (ISO, 2010, tradução nossa).

oscilações abruptas de amplitude (e possivelmente mudanças nas características espectrais) do sinal ao longo do tempo. Um objeto caindo no chão, a mudança de nota musical, a aceleração do motor e buzina de um carro (curta duração), uma batida de palmas e um grito (curta duração) são exemplos de ruídos não estacionários. Devido à imprevisibilidade do comportamento nas mudanças de amplitude ao longo do tempo (e possivelmente nas características espectrais), o desafio de reduzir o ruído não estacionário é maior.

Outra classificação diz respeito ao espectro de frequência: ruídos de banda estreita, ruídos de banda larga e tons discretos. Os ruídos de banda estreita possuem energia acústica concentrada em um pequeno intervalo de frequências [ex. um trovão distante (baixa frequência), saída de ar de um pneu (alta frequência)], enquanto nos ruídos banda larga, a energia é distribuída por uma ampla faixa de frequências (ex. vento do ar condicionado, barulho em estradas). Já o

ruído tons discretos possui flutuações de energia de tons discretos periódicos (ex. zumbido de um ventilador, bipe de dispositivo digital, nota de um instrumento musical). A hierarquia de classificação no domínio da frequência é apresentada na Figura 8.

Figura 8 – Tipos de ruído de acordo com as características dos sinais no domínio da frequência.



Fonte: elaborado pelo autor, baseado em informações de (ISO, 2010).

Um sumário de ruídos gravados em diversos tipos de ambiente provenientes da base de dados AURORA-2 (HIRSCH; PEARCE, 2000) é apresentado na Tabela 1. Os áudios possuem duração de 10 s. Hirsch e Pearce (2000) realizam a classificação de alto nível (estacionário e não estacionário) no domínio do tempo e consideram que os ruídos balbucios, restaurante, aeroporto e trem contêm segmentos não estacionários, enquanto os ruídos carro e salão de exposições são praticamente estacionários. Do ponto de vista espectral, consideram que a maioria dos ruídos são similares, embora tenham sido gravados em ambientes totalmente diferentes. Já Loizou (2017) considera o ruído carro quase estacionário e os ruídos restaurante e trem não estacionários. A análise das formas de onda no domínio do tempo e frequência permite realizar a classificação. Com exceção do ruído rua, todos são classificados como quase estacionários por conterem quadros com pulsos isolados ao longo da sua duração e todos são classificados como banda larga.

Tabela 1 – Ruídos de diversos ambientes e sua respectiva classificação nos domínios do tempo e frequência.

Ambiente	Ruídos	Classificação (domínio do tempo)	Classificação (domínio da frequência)
aeroporto	balbucios	quase estacionário	banda larga
balbucios	pessoas conversando ao fundo	quase estacionário	banda larga
carro	motor do carro	quase estacionário	banda larga
salão de exposições	balbucios e gritos ao fundo	quase estacionário	banda larga
restaurante	balbucios	quase estacionário	banda larga
ruas	motores de carros, buzinas, construções	não estacionário flutuante	banda larga
trem	ruído da via férrea	quase estacionário	banda larga

Fonte: elaborado pelo autor, baseado na base de dados AURORA-2 (HIRSCH; PEARCE, 2000).

2.2.2 Níveis de amplitude do sinal de voz e ruído em vários ambientes

Como critério de desenvolvimento de algoritmos de atenuação de ruído, é importante conhecer a estimativa da relação sinal-ruído de diversos ambientes. Os níveis da voz e do ruído variam de acordo com o ambiente e a distância do sensor para a fonte. Pearsons *et al.* (1977) realizaram um amplo estudo dos níveis de voz e ruído em ambientes como salas de aula, casas urbanas e suburbanas, hospitais, lojas, trens e aviões. Os sensores foram posicionados em diversas distâncias. Foram utilizados medidores de pressão sonora (sonômetro), registradas em NPS em dB. NPS é a pressão relativa do som em relação a $P_0 = 2 \cdot 10^{-5}$ Pa (ou $P_0 = 2 \cdot 10^{-5}$ N/m^2, que corresponde ao limiar de pressão sonora audível 0 dB na escala logarítmica). A faixa de pressão sonora audível varia de $2 \cdot 10^{-5}$ N/m^2 a 20 N/m^2 (limiar da dor), o que equivale a 0 dB e 120 dB na escala logarítmica, respectivamente. O nível de pressão sonora medido, em decibéis, é dado por (HAVELOCK *et al.*, 2008)

$$NPS_{dB} = 10 \cdot \log \left(\frac{P}{P_0} \right)^2, \tag{2.1}$$

em que P é a pressão sonora da fonte.

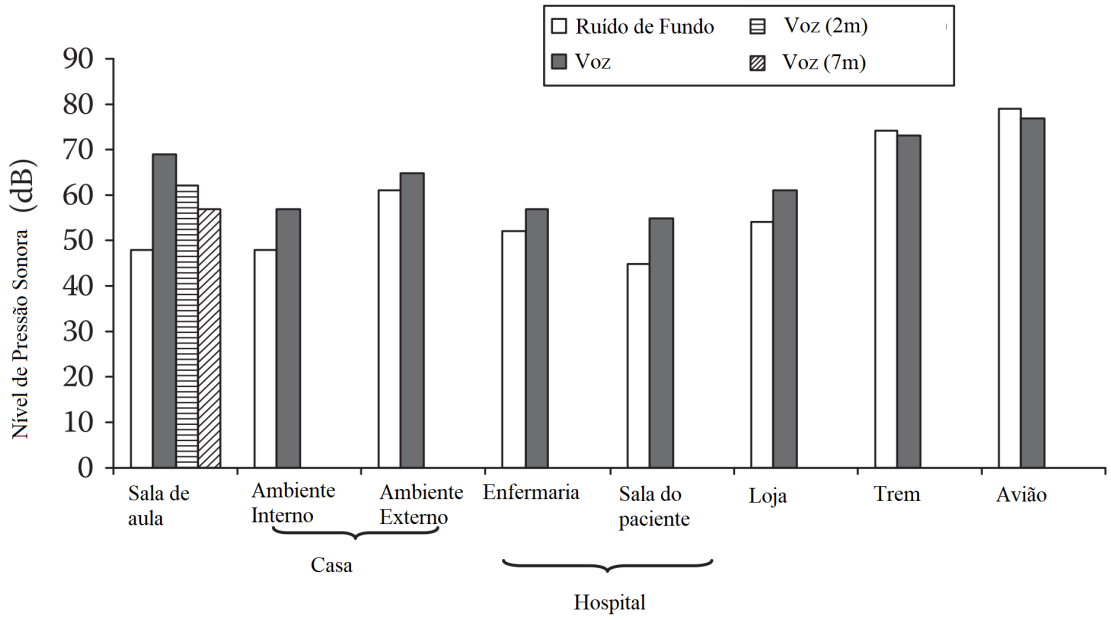
De acordo com Weisser e Buchholz (2019), a distância média estimada de uma comunicação entre duas pessoas é de um metro. A onda sonora perde a sua intensidade (nível

de pressão sonora) à medida que se propaga pelo ar e se dispersa por todas as direções. Esta propriedade do som é denominada atenuação com a distância e tem influência direta nos níveis de relação sinal-ruído do sinal de voz. Se a distância entre o sensor e a fonte sonora dobrar (i.e., $r_1 = 2r_0$), o nível de pressão sonora é reduzido em 6,02 dB (LOIZOU, 2013).

Para que as elocuções possam ser inteligíveis, as pessoas costumam reduzir a distância de comunicação ou aumentar o potência sonora da voz conforme os níveis de ruído de fundo aumentam. Em geral, quando o ruído ultrapassa o valor de 45 NPS dB, as pessoas tendem a aumentar o nível de voz, um fenômeno conhecido como efeito Lombard (LOMBARD, 1911). Segundo Loizou (2013), a intensidade da voz aumenta 0.5 dB para cada 1 dB de aumento no nível do ruído, porém, quando este alcança 70 NPS dB, as pessoas costumam cessar esse aumento.

As médias dos níveis de voz e ruído medidos em alguns ambientes são apresentados na Figura 9. Níveis de ruído mais baixos (entre 60 e 70 NPS dB) são encontrados em salas de aula silenciosas, hospitais, no ambiente interno de uma casa e em lojas. O nível de voz nesses ambientes varia entre 60 e 70 NPS dB, o que sugere que a SNR nesses ambientes varia entre 5 e 15 dB (casos extremos). Os níveis de ruido e de voz estimados em trens e aviões estão praticamente no mesmo nível (entre 70 e 75 NPS dB), o que sugere níveis de SNR em torno de 0 dB. Loizou (2013) considera que, para que algoritmos de atenuação de ruído sejam utilizados em aplicações práticas, é necessário que operem em um intervalo de SNR entre -5 e 15 dB.

Figura 9 – Média dos níveis de voz e ruído (medidos em NPS dB) em diversos ambientes.



Fonte: (LOIZOU, 2013, p. 5, tradução nossa).

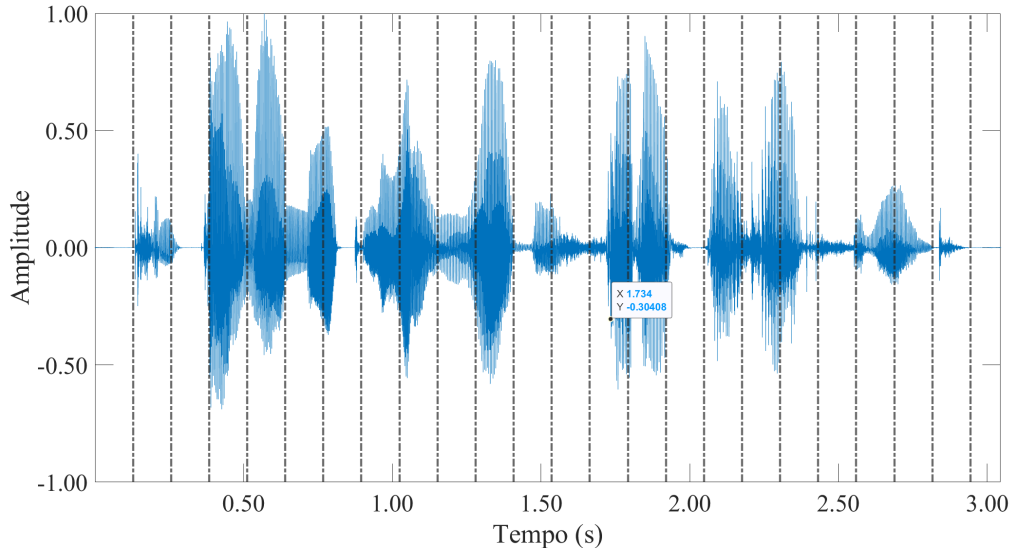
2.3 Extração de Características

Nesta Seção, descrevem-se os processos de segmentação, sobreposição, janelamento e a técnica de soma e sobreposição de quadros. Apresenta-se também a utilização da Transformada Discreta de Fourier (TDF) para os nossos experimentos, bem como os parâmetros relacionados.

2.3.1 Segmentação

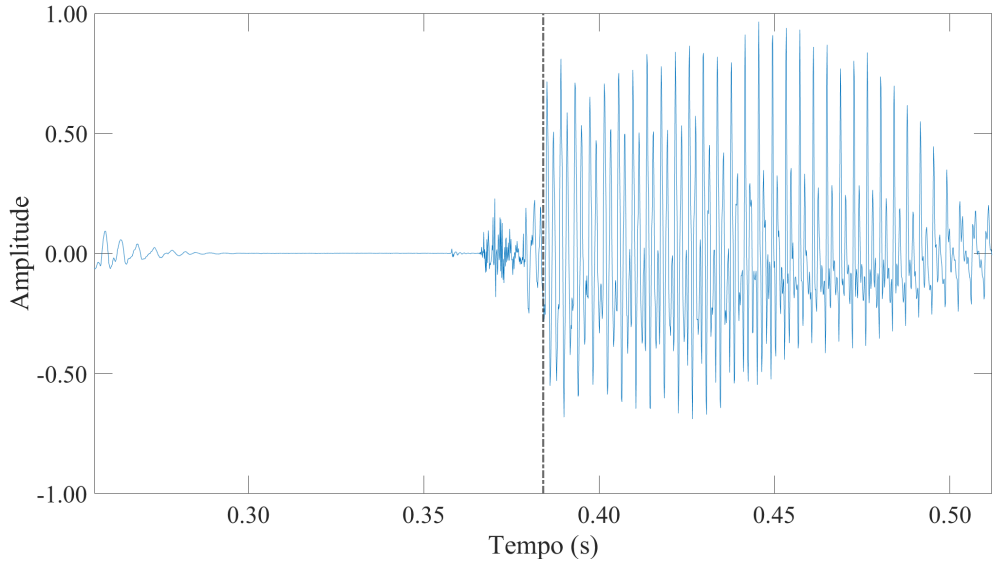
Quando o sinal de voz é longo ou a análise e processamento é feita em tempo real, há a necessidade de dividi-lo em um conjunto de quadros de tal forma que estes possam ser analisados e processados de forma independente. Tal processo é conhecido na literatura como segmentação (MCLOUGHLIN, 2016, p. 24). A Figura 10 ilustra um sinal de voz no domínio do tempo, de duração de aproximadamente 3 segundos, com linhas verticais tracejadas indicando a divisão da elocução em quadros de 1024 amostras (o que corresponde a 128 ms utilizando uma frequência de amostragem de 8 kHz). A Figura 11 apresenta especificamente os quadros 3 e 4 do sinal da Figura 10. Enquanto o quadro 3 é caracterizado majoritariamente por um período de silêncio, o quadro 4 é caracterizado pela pronúncia de parte de um fonema.

Figura 10 – Sinal de voz dividido em uma sequência de quadros de 1024 amostras.



Fonte: elaborado pelo autor.

Figura 11 – Quadros 3 e 4 do sinal de voz da Figura 10 em uma escala maior.



Fonte: elaborado pelo autor.

De acordo com McLoughlin (2016), a escolha do comprimento do quadro deve considerar os seguintes aspectos:

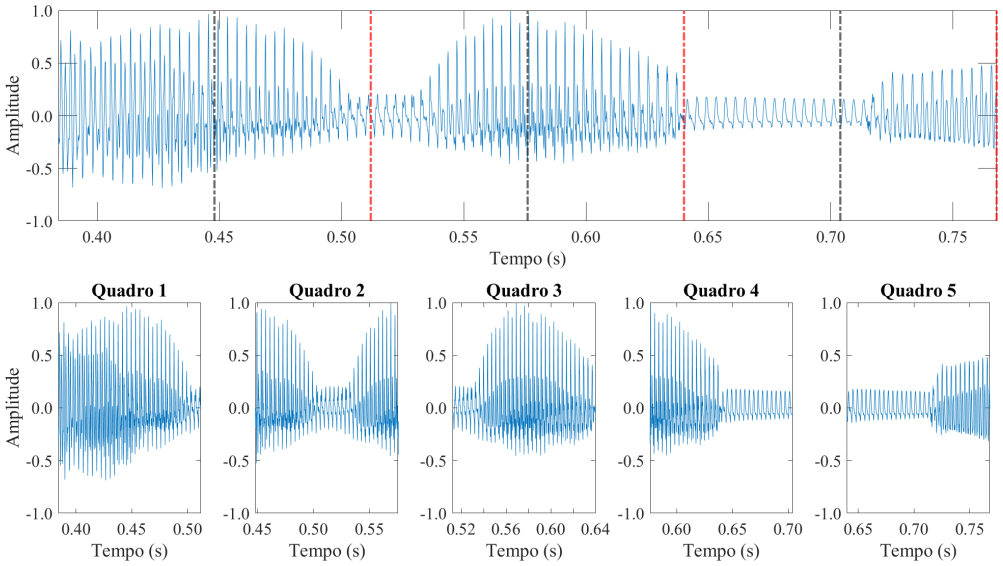
- as características do sinal de áudio no domínio do tempo;
- o algoritmo que está sendo usado para processar o sinal de áudio;
- a resolução de frequência;
- e os recursos computacionais disponíveis.

É prática utilizar uma potência de dois para o comprimento de quadro N_s, tais como 128, 256, 512, 1024, pois o processamento dos algoritmos se torna mais rápido (MCLOUGHLIN, 2016). Para um comprimento de quadro N_s e frequência de amostragem f_s (em Hz), a duração do quadro é definida como $T = \frac{N_s}{f_s}$.

2.3.2 Sobreposição

O processo de segmentação do áudio tem como consequência a perca de informações espectrais das regiões de fronteira entre um quadro e outro devido às descontinuidades. Para solucionar esse problema, utiliza-se a sobreposição de quadros. Isso significa que um determinado quadro do sinal contém um percentual dos quadros anterior e posterior. O processo de segmentação do sinal da Figura 10, especificamente no intervalo entre 0.38 e 0.76 segundos, com sobreposição de 50% das amostras, é apresentado na Figura 12. O quadro 2 é composto por metade das amostras do quadro 1 e metade das amostras do quadro 3, e assim por diante.

Figura 12 – Processo de segmentação de quadros com 1024 amostras com sobreposição de 50%.



Fonte: elaborado pelo autor.

As linhas tracejadas vermelha e preta no gráfico superior indicam as fronteiras entre quadros adjacentes utilizando sobreposição.

Se por um lado a sobreposição de quadros minimiza o problema da perda de informações espectrais, por outro, aumenta a quantidade de quadros a serem processados e, consequentemente, a necessidade de um tempo de processamento maior. Além disso, os processos de segmentação e sobreposição de quadros ocasionam problemas no processo de reconstrução do sinal. McLoughlin (2016) comenta que a simples adição dos quadros resulta numa amplitude maior no sinal reconstruído em relação ao original e haveria um número maior de amostras (o dobro, caso a sobreposição seja de 50% das amostras). Ademais, a descontinuidade abrupta entre quadros vizinhos no sinal reconstruído pode provocar cliques indesejados. Este pode ser minimizado utilizando a técnica denominada janelamento. Já o problema do rearranjo de quadros pode ser solucionado utilizando a técnica denominada soma e sobreposição[1].

2.3.3 Janelamento

A técnica de janelamento consiste em multiplicar os quadros por uma determinada função, também denominada de janela, de tal forma que as amplitudes do início e do fim do quadro sejam aproximadas para zero. Uma escolha adequada da janela minimiza os efeitos da descontinuidade no processo de rearranjo de quadros (MCLOUGHLIN, 2016). As janelas

[1] Tradução da expressão em inglês *overlap-add*.

de Hanning e Hamming são as mais comumente utilizadas na literatura e são definidas por Oppenheim e Schafer (2013, p. 317).

Cada uma dessas janelas possui características diferentes nos domínios do tempo e da frequência. Matematicamente, o quadro l após a aplicação do janelamento pode ser expresso por

$$x_w^l[n] = x^l[n] \cdot w[n], \tag{2.2}$$

em que $w[n]$ representa a função da janela e $x^l[n]$ representa o quadro antes de aplicar a função.

2.3.4 Soma e sobreposição de quadros

A sequência de processos de segmentação com sobreposição, janelamento e reconstrução do sinal utilizando a técnica de soma e sobreposição é ilustrada na Figura 13. Uma vez que as amplitudes do início e do fim do quadro são minimizadas pela função (janela), os quadros podem ser sobrepostos e somados para reconstruir o sinal no domínio do tempo, conforme apresentado por Allen (1977).

2.3.5 Transformada Discreta de Fourier (TDF)

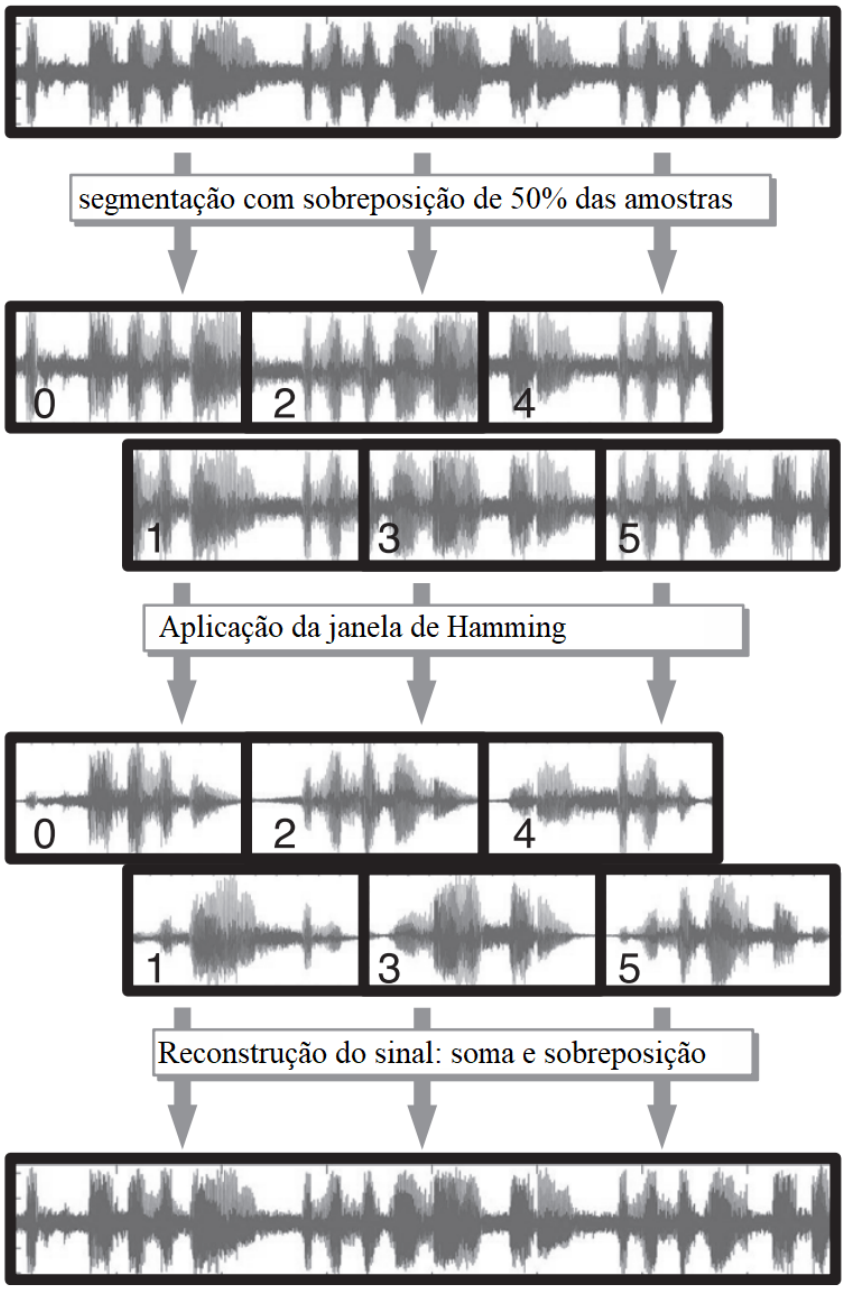
A transformada discreta de Fourier (TDF), definida por diversos autores na literatura (OPPENHEIM; SCHAFER, 2013; INGLE, 2016; LATHI; GREEN, 2005), assume a condição que o sinal discreto $x[n]$ é aperiódico e possui comprimento N_s finito. Da definição, segue que o sinal no domínio da frequência $X[k]$ possui comprimento finito e de mesmo tamanho do sinal original.

A TDF é utilizada para obter a representação dos quadros no domínio da frequência. A partir dessa representação, calcula-se a magnitude do espectro de potência do sinal para que essa informação possa ser usada como característica para o processamento dos sinais.

O número de pontos de frequência[2], isto é, o comprimento N_s da TDF, determina a resolução e contempla a energia de um determinado intervalo de frequências do sinal. A resolução R_{freq} é definida como a relação entre a frequência de amostragem F_s e o número de amostras N_s do quadro através da equação $R_{freq} = \frac{F_s}{N_s}$.

[2] Tradução da expressão em inglês *frequency bins*.

Figura 13 – Divisão de uma elocução em quadros com sobreposição de 50% das amostras. Sobre cada quadro, é aplicada a janela de Hamming antes do sinal ser reconstruído.



Fonte: (MCLOUGHLIN, 2016, p. 28, tradução nossa).

Quanto maior o tamanho do quadro, menor será a resolução em frequência e, consequentemente, mais preciso (cada ponto de frequência representa uma faixa menor do espectro). Contudo, aumentar o tamanho do quadro resulta em mais amostras e, consequentemente, custo computacional maior. Reduzir a frequência de amostragem também reduz a resolução, porém, o espectro a ser analisado também será menor, o que pode resultar em perca de componentes de frequência importantes. Pode-se concluir que há um compromisso entre o tamanho do quadro e

a frequência de amostragem e, portanto, um compromisso entre a resolução (precisão) e o custo computacional.

A comparação entre a resolução no tempo e frequência de várias combinações de comprimentos de quadros e taxas de amostragem comumente utilizadas são apresentados na Tabela 2. Para uma frequência de amostragem de 44,1 kHz, a faixa de frequência a ser analisada vai até 22,05 kHz. Se for utilizado um tamanho de quadro de 512 amostras, a resolução de frequência será 86,13 Hz, que é mais impreciso que no exemplo anterior. Além disso, a quantidade de dados a ser processado (i.e., o custo computacional) será bem maior.

Tabela 2 – Resoluções de frequência, durações e comprimentos do quadros para algumas taxas de amostragens comumente utilizadas.

Taxa de amostragem (kHz)	Frequência de Nyquist (kHz)	Duração do quadro (ms)	Comprimento do quadro (Nº de amostras)	Resolução de frequência (Hz)
8,00	4,00	16,00	0,016*8000=128	8000/128=62,5
8,00	4,00	32,00	256	31,25
8,00	4,00	64,00	512	15,62
16,00	8,00	8,00	128	125,00
16,00	8,00	16,00	256	62,50
16,00	8,00	32,00	512	31,25
44,10	22,05	11,61	512	86,13
44,10	22,05	23,22	1024	43,07
48,00	24,00	10,67	512	93,75
48,00	24,00	21,33	1024	46,87

Fonte: (MCLOUGHLIN, 2016, p.36, tradução nossa).

2.3.6 Expansão de quadros

A expansão de quadros consiste em concatenar os vetores de características log-espectrais de três, cinco ou mais quadros adjacentes, de tal forma que a RNP seja capaz de aprender a informação contextual entre os quadros. Após o processo de extração de características, é formado um vetor maior a partir da concatenação das características espectrais dos quadros anteriores ao l-ésimo quadro x^l, do próprio l-ésimo quadro e dos quadros posteriores. Para uma expansão de quadros k, o vetor de características resultante é definido como

$$\mathbf{X^l} = [X^{l-k} \ldots X^{l-2}\, X^{l-1}\, X^l\, X^{l+1}\, X^{l+2} \ldots X^{l+k}]^T \tag{2.3}$$

2.4 Subtração Espectral

A subtração espectral é um dos algoritmos clássicos da área de melhoria da qualidade do sinal de voz. Segundo Loizou (2013), a técnica se baseia no seguinte princípio:

> Assumindo que o ruído é aditivo, a estimativa da magnitude do espectro do sinal limpo pode ser obtida subtraindo a estimativa da magnitude do espectro do ruído da magnitude do espectro do sinal ruidoso. O espectro do ruído pode ser estimado e atualizado durante os períodos em que o sinal de voz está ausente. Assume-se que o ruído é estacionário ou que o processo varia lentamente e que o espectro do ruído não varia significativamente entre os períodos de atualização. A melhoria da qualidade do sinal é obtida através do cálculo da transformada discreta de Fourier do espectro do sinal estimado utilizando a fase do sinal ruidoso. O algoritmo é computacionalmente simples, uma vez que envolve apenas uma transformada de Fourier e sua operação inversa. (LOIZOU, 2013, p.93, tradução nossa).

Assume-se que o sinal de voz e o sinal do ruído são descorrelacionados. Calculando a transformada de Fourier em ambos os lados da equação (1.1), obtém-se

$$Y(w) = X(w) + R(w), \tag{2.4}$$

em que $Y(w)$, $X(w)$ e $R(w)$ são as representações do sinal de voz ruidoso, do sinal de voz limpo e do ruído, respectivamente, no domínio da frequência. Escrevendo $Y(w)$ e $R(w)$ na forma polar

$$\begin{aligned} R(w) &= |R(w)| \exp^{j\phi_r(w)}, \\ Y(w) &= |Y(w)| \exp^{j\phi_y(w)}, \end{aligned} \tag{2.5}$$

em que $|Y(w)|$ é a magnitude e $\phi_y(w)$ a fase do espectro do sinal ruidoso, respectivamente, e $|R(w)|$ é a magnitude e $\phi_r(w)$ a fase do espectro do ruído, respectivamente. Como a magnitude e a fase do espectro do ruído são desconhecidas, podem ser estimadas a partir do valor médio da magnitude e da fase do sinal ruidoso durante os períodos de silêncio. Assim, a estimativa do sinal limpo a partir do algoritmo básico da subtração espectral é obtida através da seguinte equação (LOIZOU, 2013, p.94)

$$\hat{X}(w) = [|Y(w)| - |\hat{R}(w)|] \exp^{j\phi_y(w)}, \tag{2.6}$$

em que $|\hat{R}(w)|$ é a estimativa da magnitude do espectro do ruído durante os períodos de silêncio. A estimativa do sinal limpo no domínio do tempo é obtida através da transformada discreta de Fourier inversa.

Há inúmeras publicações de variações do algoritmo básico da subtração espectral (LOIZOU, 2013, cap. 5). Nesta dissertação, utiliza-se o modelo proposto por Lockwood e Boudy (1992), à época, utilizado como uma etapa de pré-processamento para aumentar o desempenho de SRAVs na presença de ruído. A subtração espectral nesta proposta possui a seguinte equação (LOIZOU, 2013, p.114, tradução e adaptação nossa)

$$|\hat{X}(w)| = \begin{cases} |\overline{Y}(w)| - \alpha(w)N(w), \text{ se } |\overline{Y}(w)| > a(w)N(w) + \beta|\overline{R}(w)|, \\ \beta|\overline{Y}(w)|, \text{ caso contrário,} \end{cases} \tag{2.7}$$

em que β é o limiar inferior espectral, $|\overline{Y}(w)|$ e $|\overline{R}(w)|$ são as estimativas atenuadas do sinal de voz ruidoso e do ruído, respectivamente, e $\alpha(w)$ é fator de subtração dependente da frequência. $N(w)$ é a função não linear do espectro do ruído e é obtida a partir do cálculo da magnitude máxima do espectro do ruído nos últimos 40 quadros. $|\overline{Y}(w)|$ e $|\overline{R}(w)|$ são obtidos a partir das seguintes equações (LOIZOU, 2013, p.114)

$$|\overline{Y}_i(w)| = \mu_y|\overline{Y}_{i-1}(w)| + (1-\mu_y)|\hat{Y}_i(w)|, \tag{2.8}$$

$$|\overline{R}_i(w)| = \mu_r|\overline{R}_{i-1}(w)| + (1-\mu_r)|\hat{R}_i(w)|, \tag{2.9}$$

em que $|\overline{Y}_i(w)|$ é a magnitude do espectro de frequência do sinal ruidoso do i-ésimo quadro, $|\overline{R}_i(w)|$ é a estimativa da magnitude do espectro de frequência do ruído do i-ésimo quadro e as constantes μ_y e μ_r assumem valores no intervalo $0{,}1 \leq \mu_y \leq 0{,}5$ e $0{,}5 \leq \mu_r \leq 0{,}9$, respectivamente. O fator de subtração dependente da frequência $\alpha(w)$ é obtido através da relação entre o fator de escala γ e a raiz quadrada da estimativa da SNR a posteriori $\rho(w)$ (LOIZOU, 2013, p.114)

$$\alpha(w) = \frac{1}{1+\gamma\rho(w)} \tag{2.10}$$

em que a SNR a posteriori $\rho(w)$ é obtida através da seguinte equação

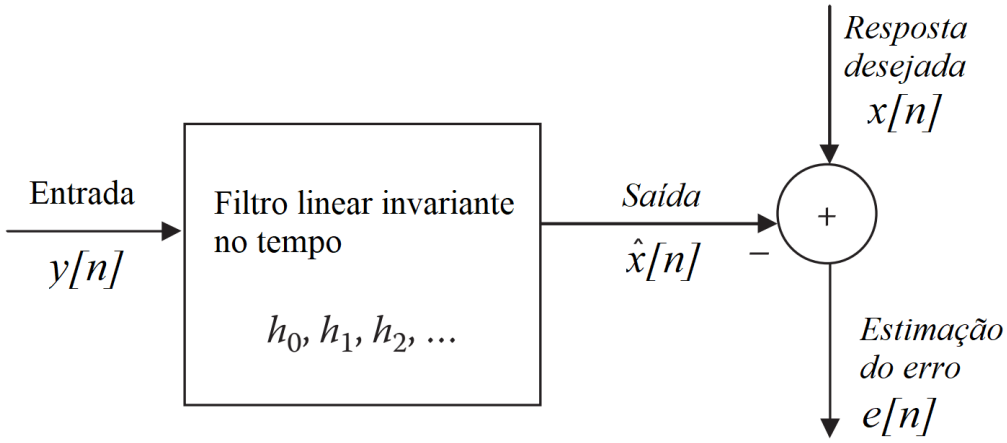
$$\rho(w) = \frac{|\overline{Y}(w)|}{|\overline{R}(w)|}. \tag{2.11}$$

2.5 Filtro de Wiener

O filtro de Wiener é uma classe de filtros lineares ótimos discretos no tempo. Diferentemente da subtração espectral, a estimativa do sinal limpo é obtida através da minimização de uma função custo. O Erro Quadrático Médio (EQM) é comumente utilizado como função custo para o projeto de otimização do filtro. Minimizar o EQM envolve estatísticas de segunda ordem (correlações) e conduz a uma teoria de filtragem útil em diversas aplicações práticas. A técnica possui esse nome em homenagem ao matemático Norbert Wiener, que formulou e solucionou o problema da filtragem no domínio do tempo contínuo (LOIZOU, 2013, p.137).

Considere o problema da filtragem estatística ilustrado em diagrama de blocos na Figura 14. O sinal de entrada segue por um sistema linear invariante no tempo para produzir um sinal de saída $\hat{x}[n]$. Projeta-se o filtro de tal maneira que a saída do sistema $\hat{x}[n]$ seja aproximada do sinal desejado $x[n]$ tanto quanto possível.

Figura 14 – Diagrama de blocos do problema da filtragem estatística.



Fonte: (LOIZOU, 2013)[p.138, tradução nossa].

Em outras palavras, deseja-se encontrar uma combinação linear dos dados de $x[n]$ de tal forma que a função custo (LOIZOU, 2013, p.137)

$$J[n] = E\{|x[n] - \hat{x}[n]|^2\} = E\{|e^2[n]|\} \tag{2.12}$$

seja minimizada. Os coeficientes ótimos do filtro $h_0, h_1, h_2 ...$ podem ser obtidos tanto no domínio do tempo quanto no domínio da frequência. A estimativa do sinal desejado $\hat{x}[n]$ é modelada como a convolução entre o sinal de entrada $y[n]$ e a resposta ao impulso $h[n]$ de um sistema linear

invariante no tempo (LIT)

$$\hat{x}[n] = h[n] * y[n]. \tag{2.13}$$

No domínio da frequência, tem-se que

$$\hat{X}(w_k) = H(w_k) \cdot Y(w_k). \tag{2.14}$$

em que $H(w_k)$ e $Y(w_k)$ são as TDF de $h[n]$ e $y[n]$, respectivamente. O cálculo da TDF de M pontos de frequência é realizado em um quadro de duração finita nas frequências $w_k = 2k\pi/M$ rad.

Definindo a estimação do erro na frequência w, tem-se que

$$\hat{E}(w_k) = X(w_k) - \hat{X}(w_k) = X(w_k) - H(w_k)Y(w_k). \tag{2.15}$$

Calculando $H(w_k)$ que minimiza o erro quadrático médio, obtém-se a forma geral do filtro de Wiener no domínio da frequência, dada por

$$H(w_k) = \frac{P_{xy}(w_k)}{P_{yy}(w_k)}, \tag{2.16}$$

em que $P_{yy}(w_k) = E\{|Y(w_k)|^2\}$ é o espectro de potência de $y[n]$ e $P_{xy}(w_k) = E[Y(w_k)D^*(w_k)]$ é o espectro de potência cruzado entre o sinal desejado $d[n]$ e o de entrada $y[n]$. Assumindo que o sinal de voz e o ruído são descorrelacionados, pode-se obter que $P_{xy}(w_k) = P_{xx}(w_k)$. Assim, a estimativa do sinal limpo é obtida através da equação (2.14).

No contexto da área de melhoria da qualidade do sinal de voz, o objetivo do filtro de Wiener é realizar uma estimativa do sinal limpo $x[n]$. Calculando a transformada de Fourier na equação (1.1), temos que

$$Y(w_k) = X(w_k) + R(w_k). \tag{2.17}$$

De acordo com Loizou (2013, p.146), o filtro de Wiener é expresso como

$$H(w_k) = \frac{\xi_k}{\xi_k + 1}, \text{ em que } 0 \leq H(w) \leq 1. \tag{2.18}$$

ξ_k é a relação sinal-ruído a priori na frequência w_k definida como

$$\xi(w_k) \triangleq \frac{P_{xx}(w_k)}{P_{rr}(w_k)}, \tag{2.19}$$

em que $P_{xx}(w_k)$ é a transformada de Fourier da autocorrelação do sinal desejado, r_{xx}. De acordo com a equação (2.19), o filtro de Wiener enfatiza as porções do espectro onde a SNR é alta e atenua as porções do espectro onde a SNR é baixa. Além disso, a equação (2.19) assume que r_{xx} é conhecida, porém, essa hipótese não é verdadeira na prática (LOIZOU, 2013, p.150). Há diversas técnicas para se estimar o filtro de Wiener a partir do sinal ruidoso. Contudo, a abordagem de cada uma delas foge do escopo desta dissertação. Para mais detalhes, sugere-se a leitura de Loizou (2013, cap. 6) e Poularikas (2015, cap. 4).

Nesta dissertação, utiliza-se a abordagem de Scalart *et al.* (1996) para se estimar a SNR a priori $\hat{\xi}_k(m)$. A função ganho é definida como

$$g(k) = \frac{\xi_k}{\xi_k + \mu}, \tag{2.20}$$

em que μ é um parâmetro de ajuste. A SNR a priori $\hat{\xi}_k(m)$ no quadro m é estimada como (LOIZOU, 2013, p.179)

$$\hat{\xi}_k(m) = \alpha \frac{|\hat{X}_k(m-1)|^2}{|N_k(m-1)|^2} + (1-\alpha)\max\left(\frac{|\hat{Y}_k(m)|^2}{|N_k(m)|^2} - 1\right), \tag{2.21}$$

em que α é um constante de suavização (0,98 em (SCALART *et al.*, 1996)), $|\hat{X}_k(m-1)|$ é a magnitude do espectro de frequência da estimação do sinal desejado no quadro $m-1$, e $|\hat{Y}_k(m)|$ e $|\hat{N}_k(m)|$ são as magnitudes dos espectros de frequências do sinal ruidoso e do ruído, respectivamente (LOIZOU, 2013, p.179).

A estimação do espectro de potência do ruído é crítica no algoritmo de Wiener. Diversas derivações do filtro de Wiener e algoritmos para estimação do ruído são apresentados por Loizou (2013). No estudo comparativo realizado por este autor, destaca-se a técnica da média recursiva dependente da SNR, proposta por Ephraim e Malah (1985), utilizada para implementação do algoritmo do filtro de Wiener nesta dissertação.

2.6 Rede Neural Profunda

Segundo Yu e Deng (2015, p.57), uma rede neural profunda (RNP) é definida como uma estrutura com duas ou mais camadas escondidas de perceptrons[3]. Com a RNP, é possível realizar o mapeamento espectral não linear entre os sinais de voz ruidosos e os sinais de voz limpos. A Figura 15 ilustra um exemplo de uma RNP com uma camada visível de entrada (identificada como camada $l = 0$), $L-1$ camadas escondidas e uma camada visível de saída (identificada como camada L). Cada camada é uma representação de alto nível das características da camada hierarquicamente inferior. Haykin (2009, p.126) afirma que os neurônios nas camadas ocultas atuam como detectores de características, realizando uma transformação não linear para um novo espaço de características. Na Figura 15, os neurônios das camadas imediatamente adjacentes estão todos conectados. A estimativa da saída desejada $\hat{d}_{n,m}$ no n-ésimo neurônio da camada de saída da m-ésima observação é obtida a partir do vetor de características de entrada $\mathbf{v^0} = \mathbf{x} \in \mathbb{R}^n$ e do conjunto de parâmetros $\{\mathbf{W}^l, \mathbf{b}^l\}$ que minimizam a função custo.

A saída do l-ésimo neurônio nas $L-1$ primeiras camadas é dado por (YU; DENG, 2015, p. 57, tradução nossa)

$$\mathbf{v}^l = f(\mathbf{z}^l) = f(\mathbf{W}^l \mathbf{v}^{l-1} + \mathbf{b}^l) \text{ para } 0 < l < L, \tag{2.22}$$

em que $f(\cdot)$ representa a função de ativação, $\mathbf{z}^l \in \mathbb{R}^{N_l \times 1}$ é o vetor de excitação, $\mathbf{v}^l \in \mathbb{R}^{N_l \times 1}$ é o vetor de ativação, $\mathbf{W}^l \in \mathbb{R}^{N_l \times N_{l-1}}$ é a a matriz de pesos, $\mathbf{b}^l \in \mathbb{R}^{N_l \times 1}$ é o vetor bias e $N_l \in \mathbb{Z}$ é o número de neurônios na l-ésima camada, respectivamente. A função de ativação deve ser diferenciável e a escolha depende da aplicação. Em aplicações de regressão, a função sigmoide, definida como,

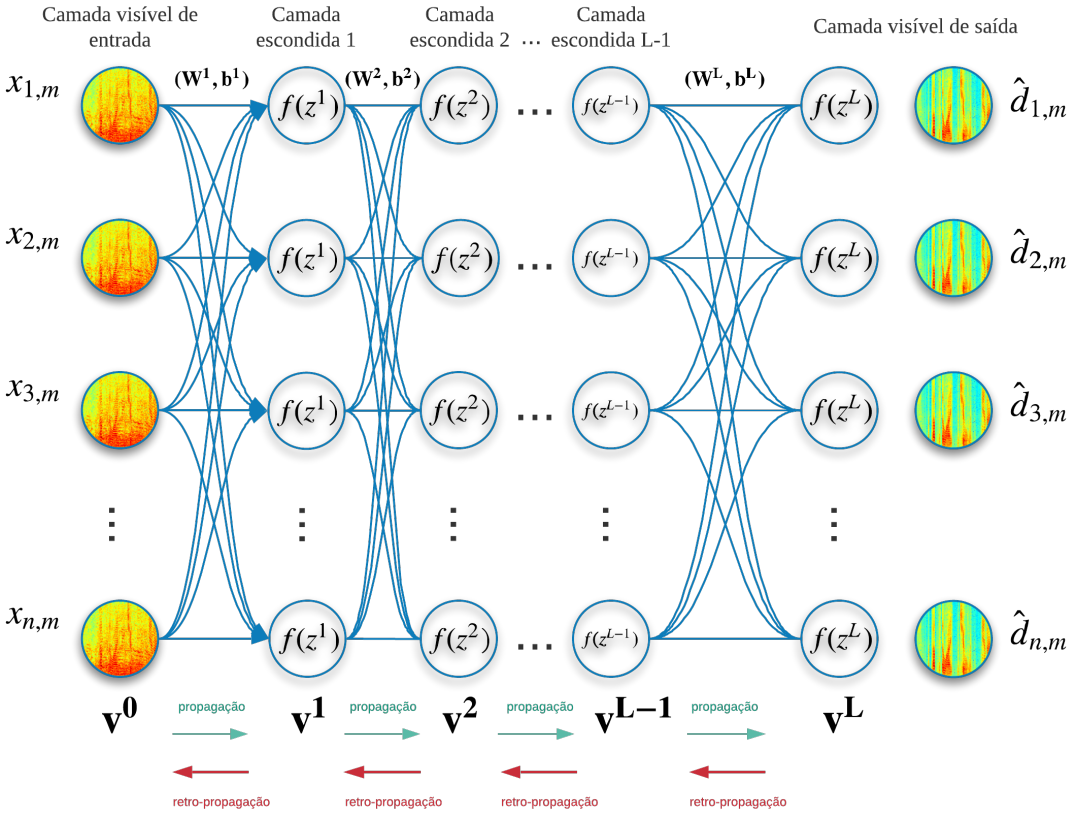
$$\rho(z) = \frac{1}{1+\exp(-z)}, \tag{2.23}$$

é comumente utilizada como função de ativação nos perceptrons das camadas ocultas, enquanto a função linear, definida como

$$\mathbf{v}^L = \mathbf{z}^L = \mathbf{W}^L \mathbf{v}^{L-1} + \mathbf{b}^L, \tag{2.24}$$

[3] A forma mais simples de uma rede neural usada para classificação de padrões linearmente separáveis. Consiste basicamente de um simples neurônio com pesos e *bias* ajustáveis (HAYKIN, 2009; ROSENBLATT, 1958, p.48).

Figura 15 – Exemplo de uma arquitetura de rede neural profunda típica com $L-1$ camadas escondidas.



Fonte: elaborado pelo autor, baseado em (YU; DENG, 2015, p.58, tradução nossa).

é utilizada como função de ativação nos neurônios da camada de saída L.

Goodfellow *et al.* (2016) definem o conceito de aprendizagem em sistemas computacionais:

> Um programa de computador aprende com a experiência E (supervisionada ou não supervisionada) com relação a uma classe de tarefas T (classificação, regressão, transcrição, tradução, detecção de anomalias, síntese, redução de ruído, estimação da densidade de probabilidade, entre outras), se seu desempenho nas tarefas T aumenta com a experiência E para uma determinada medida de desempenho P. (GOODFELLOW *et al.*, 2016, p.95, tradução nossa).

O processo de aprendizagem pode se dar de forma supervisionada ou não supervisionada. Haykin (2009) define o aprendizado supervisionado:

> No processo de aprendizagem supervisionada, o ajuste dos parâmetros da rede é realizado através da influência combinada do vetor de entrada de treinamento e a função custo (erro). A função custo é definida como a diferença entre a saída desejada e a resposta atual da rede. Nesse sentido, o conhecimento do ambiente apresentado através dos vetores de entrada, combinado com o conhecimento da saída desejada, é transferido e armazenado nos pesos sinápticos. Tais pesos representam uma memória de longo prazo. (HAYKIN, 2009, p.35, tradução nossa).

O sistema realiza o processo de aprendizado supervisionado baseado no critério de correção do erro. Uma função custo comumente utilizada é o EQM. O aprendizado não supervisionado é definido da seguinte maneira:

> O processo de aprendizado não supervisionado utiliza uma métrica independente de qualidade da representação da rede de tal forma que ela possa aprender a se auto-organizar. (HAYKIN, 2009, p.35, adaptação e tradução nossa).

De forma complementar, Goodfellow *et al.* (2016, p.103) explica que algoritmos de aprendizagem não supervisionada realizam um conjunto de observações dos dados e, de forma implícita ou explícita, aprendem a distribuição de probabilidade $p(x)$ ou alguma estrutura (ou padrão) importante do conjunto de dados de treinamento. As Subseções 2.6.1 e 2.6.2 apresentam os conceitos fundamentais das técnicas utilizadas para pré-treinamento e treinamento, respectivamente, na RNP. Na primeira, utiliza-se a técnica da máquina restrita de Boltzmann, que realiza treinamento não supervisionado. Já no treinamento, utiliza-se o algoritmo de retropropagação, que realiza treinamento supervisionado.

2.6.1 Pré-treinamento

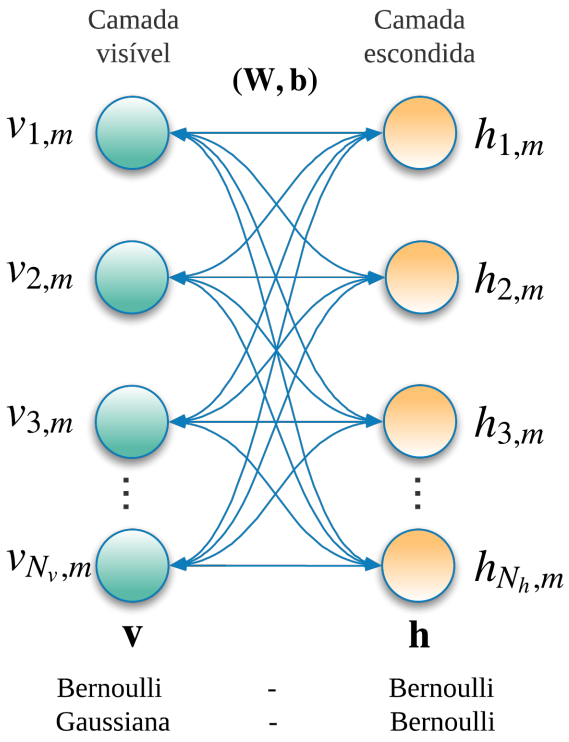
O pré-treinamento desempenha um papel fundamental em RNPs. Nesta etapa, os parâmetros do modelo ($\mathbf{W}^l$, $\mathbf{b}^l$) são atualizados de forma não supervisionada a fim de evitar mínimos locais na etapa de treinamento. Um dos algoritmos comumente utilizados nesta etapa na área da melhoria da qualidade de voz é a MRB[4], desenvolvida por Salakhutdinov *et al.* (2007).

A máquina restrita de Boltmann é uma versão simplificada da máquina de Boltzmann. O termo restrito se refere à ausência de interações diretas entre quaisquer unidades visíveis ou quaisquer unidades ocultas, como observado no gráfico do modelo na Figura 16. Na máquina de Boltzmann (geral, sem restrições), é possível haver conexões entre quaisquer unidades. Yu e Deng (2015) definem a MRB como uma rede neural estocástica generativa que consiste basicamente de modelo gráfico bidirecional[5] construído a partir de uma camada visível de neurônios e uma camada oculta estocástica de neurônios. A MRB não é um modelo profundo (i.e., de múltiplas camadas ocultas), porém, várias MRBs podem ser empilhadas para formar redes neurais profundas.

[4] Tradução da expressão em inglês *restricted Boltzmann machines* (RBM).

[5] Modelos não direcionados utilizam gráficos com conexões bidirecionais [...] (GOODFELLOW *et al.*, 2016, p.75, tradução nossa).

Figura 16 – Estrutura de uma MRB.



Fonte: elaborado pelo autor.

Os neurônios na camada oculta normalmente assumem valores binários e seguem a distribuição de probabilidade de Bernoulli. Já os neurônios na camada visível assumem valores binários (distribuição de Bernoulli) ou reais (distribuição Gaussiana). A MRB atribui uma energia para cada configuração de vetor visível $\mathbf{v}$ e vetor oculto $\mathbf{h}$. Para uma MRB Bernoulli-Bernoulli, (i.e., tanto a camada visível quanto a camada oculta seguem uma distribuição de probabilidade de Bernoulli) em que $\mathbf{v} \in \{0,1\}^{N_v \times 1}$ e $\mathbf{h} \in \{0,1\}^{N_h \times 1}$, a energia é (YU; DENG, 2015, p.79-82, tradução nossa)

$$E(\mathbf{v},\mathbf{h}) = \mathbf{a}^T\mathbf{v} - \mathbf{b}^T\mathbf{h} - \mathbf{h}^T\mathbf{W}\mathbf{v}, \tag{2.25}$$

em que N_v e N_h representam a quantidade de neurônios na camada visível e oculta, respectivamente, $\mathbf{W} \in \mathbb{R}^{N_h \times N_v}$ é a matriz de pesos e representa a conectividade entre os neurônios da camada visível e oculta, e $\mathbf{a} \in \mathbb{R}^{N_v \times 1}$ e $\mathbf{b} \in \mathbb{R}^{N_h \times 1}$ são os vetores *bias* da camada visível e oculta, respectivamente. Para uma MRB Gaussiana-Bernoulli (i.e., $\mathbf{v} \in \mathbb{R}^{N_v \times 1}$), a MRB atribui a energia

$$E(\mathbf{v},\mathbf{h}) = \frac{1}{2}(\mathbf{v}-\mathbf{a})^T(\mathbf{v}-\mathbf{a}) - \mathbf{b}^T\mathbf{h} - \mathbf{h}^T\mathbf{W}\mathbf{v} \tag{2.26}$$

para cada configuração (**v**, **h**). A rede atribui uma probabilidade, que é função da energia $E(\mathbf{v},\mathbf{h})$, para cada possível par de vetores visível e oculto $(\mathbf{v},\mathbf{h})$

$$P(\mathbf{v},\mathbf{h}) = \frac{e^{-E(\mathbf{v},\mathbf{h})}}{Z}. \tag{2.27}$$

em que $Z = \sum_{\mathbf{v}}\sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}$ é o fator de normalização conhecido como função de partição.

Na MRB, as probabilidades a posteriori $P(\mathbf{v}|\mathbf{h})$ e $P(\mathbf{h}|\mathbf{v})$ podem ser calculadas devido à ausência de conexões diretas entre neurônios na camada visível e entre os neurônios na camada oculta. Para uma MRB Bernoulli-Bernoulli, temos que

$$P(\mathbf{h} = 1|\mathbf{v}) = \sigma(\mathbf{W}\mathbf{v} + \mathbf{b}), \tag{2.28}$$

em que $\sigma(\cdot)$ representa a função sigmoide.

Para os neurônios da camada visível de uma MRB Bernoulli-Bernoulli, temos

$$P(\mathbf{v} = 1|\mathbf{h}) = \sigma\left(\mathbf{W^T}\mathbf{h} + \mathbf{a}\right), \tag{2.29}$$

e para uma MRB Gaussiana-Bernoulli, a $P(\mathbf{v}|\mathbf{h})$ é estimada como

$$P(\mathbf{v}|\mathbf{h}) = \mathcal{N}\left(\mathbf{v}; \mathbf{W^T}\mathbf{h} + \mathbf{a}, I\right), \tag{2.30}$$

em que I é a matriz identidade e $\mathcal{N}(\mathbf{x};\mu,\sum)$ é uma distribuição de probabilidade gaussiana com média $E(x) = \mu$ e matriz de covariância $E[(\mathbf{x} - \overline{\mathbf{x}})(\mathbf{x} - \overline{\mathbf{x}})^T] = \sum$.

Para treinar uma MRB, utiliza-se o algoritmo gradiente descendente estocástico para minimizar o valor negativo do logaritmo da máximo-verossimilhança (NLMV)[6] J_{LMV} e atualizar os parâmetros $(\mathbf{W},\mathbf{a},\mathbf{b})$

$$\mathbf{W}_{t+1} \leftarrow \mathbf{W}_t - \lambda\Delta\mathbf{W}_t, \tag{2.31}$$

$$\mathbf{a}_{t+1} \leftarrow \mathbf{a}_t - \lambda\Delta\mathbf{a}_t, \tag{2.32}$$

[6] Expressão traduzida do inglês *negative log likehood* (NLL).

$$\mathbf{b}_{t+1} \leftarrow \mathbf{b}_t - \lambda \Delta \mathbf{b}_t, \tag{2.33}$$

em que λ é a taxa de aprendizagem e M_b é o tamanho do mini-grupo (número de amostras processadas de forma paralela em cada iteração). $\Delta \mathbf{W}_t$, $\Delta \mathbf{a}_t$ e $\Delta \mathbf{b}_t$ são definidos por

$$\Delta \mathbf{W}_t \leftarrow \rho \Delta \mathbf{W}_{t-1} + (1-\rho)\frac{1}{M_b}\sum_{m=1}^{M_b} \nabla_{\mathbf{W_t}} J_{NLMV}(\mathbf{W}, \mathbf{a}, \mathbf{b}; \mathbf{v}^m), \tag{2.34}$$

$$\Delta \mathbf{a}_t \leftarrow \rho \Delta \mathbf{a}_{t-1} + (1-\rho)\frac{1}{M_b}\sum_{m=1}^{M_b} \nabla_{\mathbf{a_t}} J_{NLMV}(\mathbf{W}, \mathbf{a}, \mathbf{b}; \mathbf{v}^m), \tag{2.35}$$

$$\Delta \mathbf{b}_t \leftarrow \rho \Delta \mathbf{b}_{t-1} + (1-\rho)\frac{1}{M_b}\sum_{m=1}^{M_b} \nabla_{\mathbf{b_t}} J_{NLMV}(\mathbf{W}, \mathbf{a}, \mathbf{b}; \mathbf{v}^m), \tag{2.36}$$

em que ρ é o parâmetro momento e $\nabla_{\mathbf{W_t}} J_{NLMV}(\mathbf{W}, \mathbf{a}, \mathbf{b}; \mathbf{v}^m)$, $\nabla_{\mathbf{a_t}} J_{LMV}(\mathbf{W}, \mathbf{a}, \mathbf{b}; \mathbf{v}^m)$ e $\nabla_{\mathbf{b_t}} J_{NLMV}(\mathbf{W}, \mathbf{a}, \mathbf{b}; \mathbf{v}^m)$ são os gradientes do critério NLMV em relação aos parâmetros $\mathbf{W}$, $\mathbf{a}$ e $\mathbf{b}$.

A estimação completa dos gradientes na forma matricial, obtida através do amostrador Gibbs e do algoritmo divergência contrastiva, é dada por (YU; DENG, 2015, p.86)

$$\nabla_{\mathbf{W_t}} J_{NLMV}(\mathbf{W}, \mathbf{a}, \mathbf{b}; \mathbf{v}^m) = -\left[\left\langle \mathbf{h}\mathbf{v}^{\mathbf{T}} \right\rangle_{dados} - \left\langle \mathbf{h}\mathbf{v}^{\mathbf{T}} \right\rangle_{modelo}\right], \tag{2.37}$$

$$\nabla_{\mathbf{a_t}} J_{NLMV}(\mathbf{W}, \mathbf{a}, \mathbf{b}; \mathbf{v}^m) = -\left[\langle \mathbf{v} \rangle_{dados} - \langle \mathbf{v} \rangle_{modelo}\right], \tag{2.38}$$

$$\nabla_{\mathbf{b_t}} J_{NLMV}(\mathbf{W}, \mathbf{a}, \mathbf{b}; \mathbf{v}^m) = -\left[\langle \mathbf{h} \rangle_{dados} - \langle \mathbf{h} \rangle_{modelo}\right]. \tag{2.39}$$

em que $\langle x \rangle_{dados}$ e $\langle x \rangle_{modelo}$ são as esperanças estimadas de x a partir dos dados de treinamento e do modelo, respectivamente.

O algoritmo divergência contrastiva também pode ser aplicado para treinar uma MRB Gaussiana-Bernoulli. A única diferença é que, neste caso, usa-se a equação (2.30) para estimar o valor esperado da probabilidade a posteriori $\mathbb{E}(\mathbf{v}|\mathbf{h})$. Um guia prático completo do treinamento de MRB é encontrado em (HINTON, 2012).

2.6.2 Treinamento (ajuste fino)

O treinamento consiste em encontrar a solução ótima para os parâmetros $\{\mathbf{W}, \mathbf{b}\}$ de tal forma que a função custo entre as amostras de treinamento e a saída desejada seja minimizada. O algoritmo de retropropagação[7], desenvolvido por Rumelhart *et al.* (1986), baseado no gradiente descendente estocástico (GDE)[8] e mini-grupos de observações de entrada, é um dos métodos comumente utilizados para treinamento de uma rede neural com múltiplas camadas. O treinamento possui dois estágios: propagação e retropropagação.

No estágio de propagação, dada uma observação de entrada $\mathbf{v}^0 = \mathbf{x}^m = [x_{1,m}, x_{2,m}..., x_{n,m}]$, a saída da RNP com parâmetros $\{\mathbf{W}, \mathbf{b}\}$ é calculada camada a camada, utilizando a equação (2.22) para as $L-1$ primeiras camadas, e a equação (2.24) para a camada de saída L. Nesta fase, não há modificações nos parâmetros do modelo $\{\mathbf{W}, \mathbf{b}\}$.

Já no estágio de retropropagação, o erro é calculado a partir da diferença entre a saída $\mathbf{v}^L$ da rede e a resposta desejada $\mathbf{d}$ (aprendizado supervisionado). O erro resultante $\mathbf{e}$ é propagado camada a camada na direção oposta à direção do estágio da propagação.

Na sua forma mais simples, os parâmetros $\{\mathbf{W}, \mathbf{b}\}$ podem ser atualizados de acordo com a regra de aprendizagem (YU; DENG, 2015, p.61, tradução nossa)

$$\begin{aligned} \mathbf{W}_{t+1}^l &\leftarrow \mathbf{W}_t^l - \lambda^l \Delta \mathbf{W}_t^l, \\ \mathbf{W}_{t+1}^l &\leftarrow \mathbf{W}_t^l - \lambda^l \frac{1}{M_b} \sum_{m=1}^{M_b} \nabla_{\mathbf{W_t^l}} J(\mathbf{W}, \mathbf{b}; \mathbf{x}, \mathbf{d}), \end{aligned} \tag{2.40}$$

$$\begin{aligned} \mathbf{b}_{t+1}^l &\leftarrow \mathbf{b}_t^l - \lambda^l \Delta \mathbf{b}_t^l, \\ \mathbf{b}_{t+1}^l &\leftarrow \mathbf{b}_t^l - \lambda^l \frac{1}{M_b} \sum_{m=1}^{M_b} \nabla_{\mathbf{b_t^l}} J(\mathbf{W}, \mathbf{b}; \mathbf{x}, \mathbf{d}), \end{aligned} \tag{2.41}$$

em que λ^l é a taxa de aprendizagem na l-ésima camada, $\mathbf{W}_t^l$ e $\mathbf{b}_t^l$ são a matriz de pesos e o vetor *bias* na l-ésima camada depois da t-ésima atualização, $\Delta \mathbf{W}_t^l$ e $\Delta \mathbf{b}_t^l$ são os gradientes médios da função custo em relação à matriz de pesos e ao vetor *bias*, respectivamente, estimados a partir do treinamento por mini-grupos de M_b amostras. $\nabla_x J$ é o gradiente da função custo em relação a $\mathbf{x}$.

[7] Expressão traduzida do inglês *backpropagation*.
[8] Tradução da expressão em inglês *stochastic gradient descent* (SGD).

Na forma completa, os sucessivos ajustes são realizados nos parâmetros $\{\mathbf{W}, \mathbf{b}\}$ da RNP de acordo com as regras de aprendizagem

$$\mathbf{W}_{t+1}^{l} \leftarrow \rho \mathbf{W}_{t}^{l} - (1-\rho)\frac{\lambda^{l}}{M_b}\sum_{m=1}^{M_b}\left(\nabla_{\mathbf{W_t^l}} J(\mathbf{W}, \mathbf{b}; \mathbf{x}, \mathbf{d}) + \gamma \nabla_{\mathbf{W_t^l}} \mathbf{R}(\mathbf{W_t^l})\right), \tag{2.42}$$

$$\mathbf{b}_{t+1}^{l} \leftarrow \rho \mathbf{b}_{t}^{l} - (1-\rho)\frac{\lambda^{l}}{M_b}\sum_{m=1}^{M_b}\left(\nabla_{\mathbf{b}_t^l} J(\mathbf{W}, \mathbf{b}; \mathbf{x}, \mathbf{d})\right), \tag{2.43}$$

em que $\mathbf{W}_t^l$ é a matriz de pesos na l-ésima camada na t-ésima iteração, $\mathbf{b}_t^l$ é o vetor bias, $\nabla_{\mathbf{W_t^l}} J(\mathbf{W}, \mathbf{b}; \mathbf{x}, \mathbf{d})$ é o gradiente da função custo (erro médio quadrático) relativo à matriz de pesos da camada l, $\nabla_{\mathbf{b_t^l}} J(\mathbf{W}, \mathbf{b}; \mathbf{x}, \mathbf{d})$ é o gradiente da função custo relativo ao vetor *bias* da camada l, λ^l é a taxa de aprendizagem na camada l, M_b é o tamanho do mini-grupo de observações de entrada, ρ é o fator de momento, γ é o fator de interpolação (regularização dos pesos) e $\mathbf{R}$ é o termo de regularização (decaimento ponderado), $\nabla_{\mathbf{W_t^l}} \mathbf{R_1}(\mathbf{W}) = sign(\mathbf{W_t^l})$ para norma L_1 e $\nabla_{\mathbf{W_t^l}} \mathbf{R_2}(\mathbf{W}) = 2\mathbf{W_t^l}$ para norma L_2.

Para regressão, o critério do erro médio quadrado é comumente utilizado como função custo e definido como (YU; DENG, 2015, p.60, tradução nossa)

$$J_{EQM}(\mathbf{W}, \mathbf{b}; \mathbb{S}) = \frac{1}{M}\sum_{m=1}^{M} J_{EQM}(\mathbf{W}, \mathbf{b}; \mathbf{x}^m, \mathbf{d}^m) \tag{2.44}$$

em que $\mathbb{S} = \{(\mathbf{x}^m, \mathbf{d}^m) | 0 \leq m < M\}$ é o conjunto de treinamento e

$$J_{EQM}(\mathbf{W}, \mathbf{b}; \mathbf{x}, \mathbf{d}) = \frac{1}{2}||\mathbf{v}^L - \mathbf{d}||^2 = \frac{1}{2}\left(\mathbf{v}^L - \mathbf{d}\right)^T \left(\mathbf{v}^L - \mathbf{d}\right). \tag{2.45}$$

Para realizar o estágio de retropropagação, é necessário definir os gradientes da função custo com relação às matrizes de pesos e aos vetores *bias*, tanto em relação à camada de saída, quanto em relação às $L-1$ camadas ocultas. O gradiente da função custo com relação à matriz de pesos da camada de saída L é definido por

$$\nabla_{\mathbf{W_t^l}} J_{EQM}(\mathbf{W}, \mathbf{b}; \mathbf{x}, \mathbf{d}) = \left(\mathbf{v}^L - \mathbf{d}\right)^T \left(\mathbf{v}^{L-1}\right)^T, \tag{2.46}$$

em que

$$e_t^L = \left(\mathbf{v}^L - \mathbf{d}\right) \tag{2.47}$$

é definido como o erro de estimação do sinal desejado na camada de saída L. Já o gradiente da função custo com relação ao vetor *bias* da camada de saída L é definido por

$$\nabla_{\mathbf{b_t^l}} J_{EQM}(\mathbf{W},\mathbf{b};\mathbf{x},\mathbf{d}) = \left(\mathbf{v}^L - \mathbf{d}\right)^T. \tag{2.48}$$

Para as l camadas ocultas ($0 < l < L$), os gradientes da função custo com relação à matriz de pesos e ao vetor *bias* são definidos, respectivamente, por

$$\nabla_{\mathbf{W_t^l}} J_{EQM}(\mathbf{W},\mathbf{b};\mathbf{x},\mathbf{d}) = \left[f'\left(\mathbf{z}_t^l\right) \bullet \mathbf{e}_t^l\right]\left(\mathbf{v}_t^{l-1}\right)^T, \tag{2.49}$$

e

$$\nabla_{\mathbf{b_t^l}} J_{EQM}(\mathbf{W},\mathbf{b};\mathbf{x},\mathbf{d}) = f'\left(\mathbf{z}_t^l\right) \bullet \mathbf{e}_t^l, \tag{2.50}$$

em que $\mathbf{e}_t^l$ é o erro na l-ésima camada, $\bullet$ é produto elemento a elemento e $f'(\cdot)$ é a derivada elemento a elemento da função de ativação. Para a função sigmoide, a derivada é obtida como (YU; DENG, 2015, p.64, tradução nossa)

$$\sigma'\left(\mathbf{z}_t^l\right) = \left(1 - \sigma(\mathbf{z}_t^l)\right) \bullet \sigma\left(\mathbf{z}_t^l\right) = \left(1 - \mathbf{v}_t^l\right) \bullet \mathbf{v}_t^l, \tag{2.51}$$

O erro do sinal pode ser retropropagado por

$$e_t^{L-1} = \left(\mathbf{W}^l\right)^T e_t^L. \tag{2.52}$$

para a camada de saída e, para $l < L$, por

$$e_t^{l-1} = \left(\mathbf{W}^l\right)^T \left[f'\left(\mathbf{z}_t^l\right) \bullet \mathbf{e}_t^l\right]. \tag{2.53}$$

Considerações práticas de desenvolvimento do algoritmo de retropropagação em relação ao pré-processamento de dados, à inicialização do modelo, ao decaimento ponderado, ao *dropout*, à seleção do tamanho dos mini-grupos, à técnica momento, às taxas de aprendizagem, ao critério de parada, à capacidade de generalização da rede, aos mínimos locais e à interrupção preventiva podem ser encontrados nas referências (YU; DENG, 2015) e (HAYKIN, 2009).

3 TRABALHOS RELACIONADOS

Neste Capítulo, é feito um levantamento de trabalhos existentes na área da melhoria da qualidade de sinais, desenvolvidos e publicados nos últimos anos. Os métodos encontrados na pesquisa se baseiam na subtração espectral, no filtro de Wiener, no filtro de Kalman, no estimador de minimização do EQM, na transformada *wavelet* e em redes neurais.

Para revisão da literatura, foram utilizadas palavras-chave como melhoria da qualidade de sinais de voz, redução de ruído, redes neurais profundas e mapeamento espectral não linear. Para seleção das referências, foram utilizados os seguintes critérios: o grau de pertinência à esta dissertação, a quantidade de citações, os resultados obtidos, o ano de publicação e a relevância das publicações dos autores nesta área. O portal de catálogo de teses e dissertações da CAPES (CAPES, 2016), a plataforma Google Acadêmico (GOOGLE, 2004), o portal de pesquisa do IEEE (IEEE, 2004), o portal ResearchGate (MADISCH *et al.*, 2008) e o portal ScienceDirect (ELSEVIER, 1997) foram as principais ferramentas de pesquisa utilizadas.

Taha e Hussain (2018) realizam um amplo levantamento de técnicas dedicadas à melhoria da qualidade de sinais de voz. O artigo aborda diversos métodos como a subtração espectral, o filtro de Wiener, métodos estatísticos, métodos de subespaço e métodos de aprendizagem de máquina (como redes neurais superficiais, redes neurais profundas e redes neurais convolucionais). Os autores classificam os métodos em quatro categorias: métodos convencionais, métodos de filtragem adaptativa, métodos multimodais e métodos de aprendizado de máquina. As principais vantagens e desvantagens de cada um dos métodos abordados no levantamento também são descritas. Esse trabalho apresenta uma visão sistemática do estado da arte de métodos que objetivam a melhoria da qualidade de sinais de voz, permitindo que pesquisadores explorem os desafios existentes.

A subtração espectral e o filtro de Wiener são dois algoritmos clássicos apresentados na pesquisa de Taha e Hussain (2018) e bastante explorados na literatura. Tais algoritmos são utilizados como referência para avaliação de desempenho e comparação com os resultados obtidos pelo método da rede neural profunda proposta nesta dissertação. De acordo com Xu *et al.* (2014), um problema normalmente encontrado nesses métodos convencionais é que o sinal de voz limpo estimado sofre de um problema denominado ruído musical. Tais métodos também falham na presença de ruídos não estacionários cujas condições acústicas são inesperadas (maioria dos cenários do mundo físico).

Em um trabalho recente, Abreu (2017) apresenta uma análise dos principais métodos

para melhorar a qualidade de sinais de voz, baseados tanto na análise de Fourier quanto em *wavelets*. São analisados os modelos subtração espectral, filtro de Wiener, estimador da minimização do EQM, estimador máximo a posteriori, limiarização *wavelet* tradicional, limiarização *wavelet* adaptativa e o método *wavelet* não limiar. O autor também apresenta um modelo para identificar e classificar ruídos, pois todos esses métodos necessitam de uma estimação precisa do perfil do ruído para realizar a redução de ruído de forma eficiente. Para avaliação de desempenho, são utilizadas as métricas SNR Global, PESQ e o coeficiente de correlação de Pearson. Dentre os métodos citados, o filtro de Wiener obteve o melhor desempenho geral.

LIMA (2014) realiza uma avaliação comparativa da aplicação dos filtros de Wiener ótimo e sub-ótimo (CHEN *et al.*, 2006) para reduzir o ruído aditivo branco Gaussiano com níveis de SNR de 0, 3, 5, 10, 15 e 20 dB. Foram utilizados 20 arquivos contendo elocuções no idioma português e os resultados são avaliados em um SRAV. Contudo, a origem dos arquivos não é descrita e os mesmos também não são disponibilizados para fins de replicação. Além disso, a metodologia de avaliação das palavras reconhecidas não é descrita. Uma palavra excluída ou inserida, por exemplo, como na métrica TEP, é considerada como um erro de reconhecimento? Outra apreciação está relacionada à utilização de somente um tipo de ruído, o que torna o cenário para melhoria do sinal de voz menos complexo e desafiador e, portanto, distante de cenários práticos. Por necessitar de uma estimação do perfil do ruído precisa, o desempenho dos modelos de Abreu (2017) e LIMA (2014) tende a ser baixo com ruídos não estacionários. Além disso, o nível de distorção inserido em detrimento da redução do ruído é alto em níveis baixos de SNR.

Um estudo comparativo de técnicas de redução de ruído baseado nos filtros de Kalman e na subtração espectral é apresentado por Silva (2007). Sinais de voz contendo uma única palavra são degradados com ruído branco e ruído colorido em níveis de SNR (0, 3 e 6 dB). Os resultados são avaliados com as métricas *segmental SNR* (segSNR) e a distância de Itakura-Sato. Nesse trabalho, não são avaliados ruídos em cenários do mundo físico e o vocabulário de arquivos de sinais de voz é limitado (apenas palavras). O autor afirma que o método proposto aumenta a qualidade e a inteligibilidade, porém, esta última não é avaliada por nenhuma das métricas de avaliação de desempenho. Segundo Loizou (2017), as métricas segSNR e Itakura-Sato possuem baixas correlações com a distorção espectral e a inteligibilidade de um modo geral, enquanto as métricas PESQ e STOI possuem alta correlação.

Santos (2015) combina as técnicas filtro de Kalman e a transformada *wavelet* para reduzir o ruído branco Gaussiano, em sinais de voz, degradados com os níveis de SNR 0, 3, 5,

10, 15 e 20 dB. O autor utiliza um algoritmo genético para determinar o conjunto de coeficientes ótimo para a Transformada *Wavelet* Discreta (DWT) e para a transformada Decomposição de Pacotes *Wavelet* (WPD). Utiliza-se um banco de dados com 20 arquivos contendo elocuções (10 vozes masculinas e 10 vozes femininas), contendo elocuções com uma única palavra. Os resultados são avaliados com as métricas objetivas segSNR, a distância de Itakura-Saito e de maneira subjetiva por um grupo de 10 voluntários com idades variando entre 9 e 65 anos. Como referência, o autor comparou o método proposto com o método da subtração espectral. A dissertação não avalia cenários com ruídos do mundo físico e não caracteriza os ruídos utilizados no domínio do tempo e no domínio da frequência. Além disso, a técnica proposta tende a espalhar a energia pelo espectro de frequência, o que afeta a inteligibilidade e aumenta a distorção espectral, não avaliadas pelas métricas objetivas utilizadas.

Rede neural profunda é uma abordagem relativamente recente que tem sido desenvolvida e aplicada para a melhoria da qualidade dos sinais de voz (XU *et al.*, 2014; XU *et al.*, 2015; HAN *et al.*, 2015; LIU *et al.*, 2014; PARK; LEE, 2016). Xu *et al.* (2015) apresentam uma estrutura de processamento de sinais de voz baseado em RNP escondidas, no qual as características espectrais dos sinais são usadas para realizar um mapeamento não linear entre os sinais de voz ruidosos e os sinais de voz limpos. Os autores construíram um banco de dados de treinamento (idioma inglês) contendo os ruídos aditivo branco Gaussiano, balbucios, restaurante e rua e os níveis de SNRs -5, 0, 5, 10, 15 e 20 dB. Foram utilizados 4.620 arquivos contendo sinais de voz para cada combinação de ruído e nível de SNR, totalizando cerca de 100 horas (incluindo sinais de voz limpos) para o banco de dados de treinamento. Já para o banco de dados de teste, são utilizados 200 arquivos contendo elocuções para cada combinação de ruído e SNR. Além disso, os ruídos carro e salão de exposições e a SNR 7 dB são acrescentadas aos cenários de teste.

Xu *et al.* (2015) avaliou o desempenho utilizando as métricas objetivas SNR segmental (segSNR, em dB), o logaritmo da distância espectral[1] e a PESQ. Outro teste realizado foi o uso 200 elocuções no idioma mandarim corrompidos com o ruído balbucios. Foram realizados testes subjetivos com 5 homens e 5 mulheres. A preferência média nos testes subjetivos e os melhores resultados são obtidos pela RNP em comparação com os métodos L-MSE e rede neural superficial. Entretanto, técnicas de regularização (*dropout*, decaimento ponderado, otimizador Adam, entre outros) não utilizadas nesse trabalho têm o potencial de aumentar a capacidade de

[1] Expressão traduzida do inglês, *log-spectral distance* (LSD).

generalização da rede e o seu desempenho geral. Vale observar que Xu *et al.* (2015) apresentam somente o comparativo da métrica PESQ com os demais algoritmos sob a justificativa da limitação de espaço do artigo. Por conta disso, não fica claro o quantitativo de redução ou aumento da distorção espectral e da SNR segmentada provocado pelos algoritmos avaliados (inclusive a RNP). Os autores também afirmam que foi utilizada uma taxa de aprendizagem de 0,1 e um tamanho de mini-grupo de 128 em todas as camadas de aprendizagem, contudo, não foi possível replicar os resultados demonstrados neste artigo com os valores apresentados, pois o EQM nesta configuração não converge.

LeCun (1993, p.20, tradução nossa) afirma que alguns parâmetros necessitam de uma taxa de aprendizagem pequena para evitar divergência, enquanto outros necessitam de uma taxa de aprendizagem alta para convergir em uma velocidade (quantidade de épocas de treinamento) razoável. Haykin (2009, p.150, tradução nossa) e LeCun (1993) afirmam que, idealmente, todos os neurônios deveriam aprender à mesma taxa, porém, as últimas camadas normalmente possuem um gradiente local maior do que nas camadas iniciais da rede. Por conta disso, a taxa de aprendizagem λ^l nas últimas camadas deve ser menor do que a taxa de aprendizagem para as camadas iniciais. Além disso, o número de conexões sinápticas influencia na taxa de aprendizagem do neurônio. Os testes realizados nesta dissertação indicam a necessidade de uma taxa de aprendizagem de pelo menos 0,001 e um tamanho de mini-grupo de 128, para aprendizagem dos parâmetros da última camada da RNP, na etapa de treinamento, da proposta de Xu *et al.* (2015). As demais taxas de aprendizagem de 0,1 e um tamanho de mini-grupo de 128 podem ser mantidas.

Xu *et al.* (2015) desenvolveram uma extensão do trabalho anterior. Para aumentar a capacidade de generalização da rede, são acrescentados 100 tipos diferentes de ruído no banco de dados de treinamento, especialmente ruídos não estacionários. Isso permite aumentar a capacidade de generalização da rede e avaliação de desempenho de sinais ruidosos não utilizados no banco de dados de treinamento. A equalização com variância global, o *dropout*[2] e o treinamento sensível ao ruído[3] são técnicas acrescentadas para aumentar a capacidade de generalização da rede, suavizar os picos dos sinais de fala reconhecidos, reduzir o ruído residual e aumentar o grau de precisão do sinal limpo. Para atualização dos pesos, os autores utilizaram a técnica momento e a técnica decaimento ponderado. Graças ao rico banco de dados heterogêneo

[2] Técnica utilizada para desativar, de forma aleatória, um percentual de neurônios na camada de entrada e nas camadas ocultas em cada época da etapa de treinamento.

[3] Tradução da expressão em inglês *noise aware training* (NAT).

de treinamento (mais de 625 horas), com diversas categorias de ruído, e as técnicas introduzidas, o método proposto é capaz de suprimir o ruído não estacionário, que, em geral, é difícil de suprimir.

Em relação aos artigos desenvolvidos por Xu *et al.* (2014), Xu *et al.* (2015), propõe-se nesta dissertação a utilização de sinais de voz gerados com SNRs aleatórias com distribuição uniforme, no intervalo de 0 a 15 dB. Espera-se com essa medida aumentar a capacidade de generalização da rede neural profunda para diferentes níveis de SNR nesse intervalo[4]. A partir do levantamento bibliográfico realizado, não há nenhum outro trabalho prévio que utilize essa abordagem. Além disso, outra contribuição consiste na utilização de um sistema de reconhecimento automático de voz (SRAV), em substituição aos testes subjetivos, para avaliar a melhoria na qualidade e na inteligibilidade dos sinais de voz. Em relação ao treinamento da RNP, são utilizadas taxas de aprendizagem diferentes para cada uma das camadas, a fim de obter uma convergência do EQM numa quantidade menor de épocas. A técnica *dropout* é utilizada para aumentar a capacidade de generalização da rede. Utiliza-se também uma quantidade maior de arquivos de sinais de voz (840) para o conjunto de dados de teste, para cada um dos 24 cenários[5].

[4] Restringe-se o intervalo de SNR utilizado nesses artigos devido ao custo computacional elevado para treinar uma grande quantidade de arquivos de áudio

[5] Cada um dos cenários é a combinação de um ruído (balbucios, carro, trem ou aeroporto) com um nível de SNR(0, 5, 10, 15 dB e uma categoria de SNRs aleatórias).

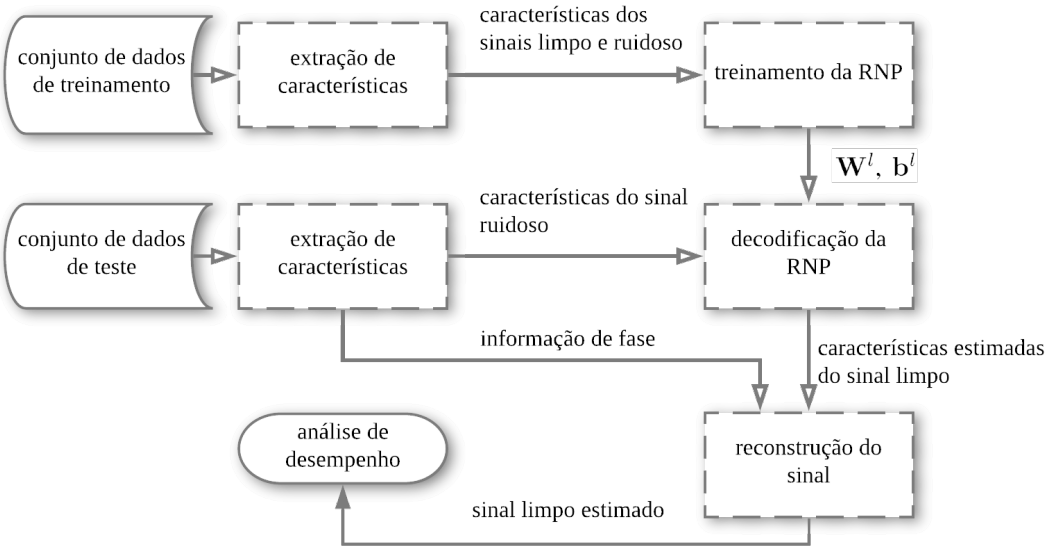
4 METODOLOGIA

Este Capítulo apresenta a metodologia de desenvolvimento do trabalho. A Seção 4.1 apresenta uma visão geral do sistema utilizado para redução de ruído utilizando a RNP. A Seção 4.2 descreve o banco de dados. Os parâmetros da extração de características são descritos na Seção 4.3. Detalhes acerca da RNP são apresentados na Seção 4.4 e a reconstrução do sinal é apresentada na Seção 4.5. Finalmente, a Seção 4.6 apresenta as métricas de desempenho utilizadas para avaliação dos resultados.

4.1 Visão geral

A estrutura para atenuação de ruído utilizando a RNP é fundamentada na estrutura proposta por Xu *et al.* (2014). Estes autores se basearam na estrutura apresentada em (DU; HUO, 2008), que utilizava GMM e uma aproximação linear por partes para estimar o sinal limpo, a partir do sinal ruidoso. O diagrama da estrutura é ilustrado na Figura 17 e é composto pelos seguintes blocos: extração de características, treinamento da RNP, decodificação da RNP e reconstrução do sinal. As seções a seguir apresentam a descrição detalhada de cada um dos blocos.

Figura 17 – Diagrama de blocos da estrutura de atenuação de ruído.



Fonte: elaborado pelo autor, baseado na estrutura proposta por Xu *et al.* (2014).

4.2 Banco de dados

O sucesso da aplicação de uma RNP com aprendizado supervisionado, no contexto de aprimoramento do sinal de voz, depende da disponibilidade de uma base de dados composta por dezenas de horas de áudio mono-canal com locutores homens e mulheres de diferentes idades, sotaques e dialetos, bem como suas respectivas transcrições. Utiliza-se a base de dados TIMIT (GAROFOLO *et al.*, 1988) no idioma inglês[1] que contém um total de 6.300 áudios (sinais limpos). Esse número é resultado da gravação de 10 frases de um total de 630 pessoas (438 homens e 192 mulheres) de 8 regiões dialéticas dos Estados Unidos.

São usados 4.620 arquivos de áudios (73,3% do total, o que equivale a 3 horas, 56 minutos e 33 segundos) para o conjunto de dados de treinamento, 840 arquivos de áudio (13,33% do total, o que equivale a 43 minutos e 13 segundos) para o conjunto de validação e 840 arquivos de áudio para o conjunto de teste (os mesmos 13,33% do total, o que equivale a 43 minutos e 13 segundos).

Para cada sinal de voz do conjunto de treinamento, adicionam-se três tipos de ruídos (balbucios, carro, trem) e são estabelecidos quatro níveis principais de SNR (0 dB, 5 dB, 10 dB e 15 dB). Para cada sinal de voz do conjunto de validação e teste, somam-se quatro tipos de ruídos (balbucios, carro, trem e aeroporto) e são estabelecidos cinco níveis principais de SNR (0 dB, 5 dB, 7 dB, 10 dB e 15 dB). O ruído aeroporto e a SNR 7 dB são utilizados no banco de testes para avaliar a capacidade de generalização da RNP, uma vez que tais cenários não fazem parte do conjunto de dados de treinamento.

Adicionam-se também 4.620 arquivos de áudio ruidoso com valores aleatórios de SNR gerados com uma distribuição uniforme em um intervalo entre 0 e 15 dB, para cada tipo de ruído do conjunto de treinamento. De forma similar, adicionam-se 1.640 arquivos de áudio ruidoso com valores de SNR aleatórios também gerados com distribuição uniforme entre 0 e 15 dB para cada tipo de ruído do conjunto de validação e teste. Com isso, objetiva-se ampliar a capacidade de generalização da RNP. Um sumário da duração total utilizada para cada tipo de ruído e valor de SNR para o conjunto de dados de treinamento e para o conjunto de dados de teste são apresentados nas Tabelas 3 e 4, respectivamente.

Os áudios dos ruídos são provenientes da base de dados AURORA-2 (HIRSCH;

[1] Em nossas pesquisas, não encontramos uma base de dados no idioma português tão rica e heterogênea quanto a TIMIT com sinais de voz limpo. Além disso, a maioria dos bancos de dados encontrados no idioma português já possuem algum tipo de ruído e/ou reverberação. Por conta desses fatores, este trabalho foi baseado no idioma inglês.

Tabela 3 – Sumário da duração do conjunto de arquivos de áudio ruidoso gerado para o conjunto de dados de treinamento referente aos diferentes tipos de ruído e valores de SNR.

SNR/ tipo de ruído	0 dB	5 dB	10 dB	15 dB	SNR aleatória (0 a 15 dB)	duração total
balbucios	3:56:33	3:56:33	3:56:33	3:56:33	3:56:33	19:42:45
carro	3:56:33	3:56:33	3:56:33	3:56:33	3:56:33	19:42:45
trem	3:56:33	3:56:33	3:56:33	3:56:33	3:56:33	19:42:45
duração total	11:49:39	11:49:39	11:49:39	11:49:39	11:49:39	59:08:15

Fonte: elaborado pelo autor.

Tabela 4 – Sumário da duração de áudio ruidoso gerado para o conjunto de dados de validação e teste referente aos diferentes tipos de ruído e valores de SNR

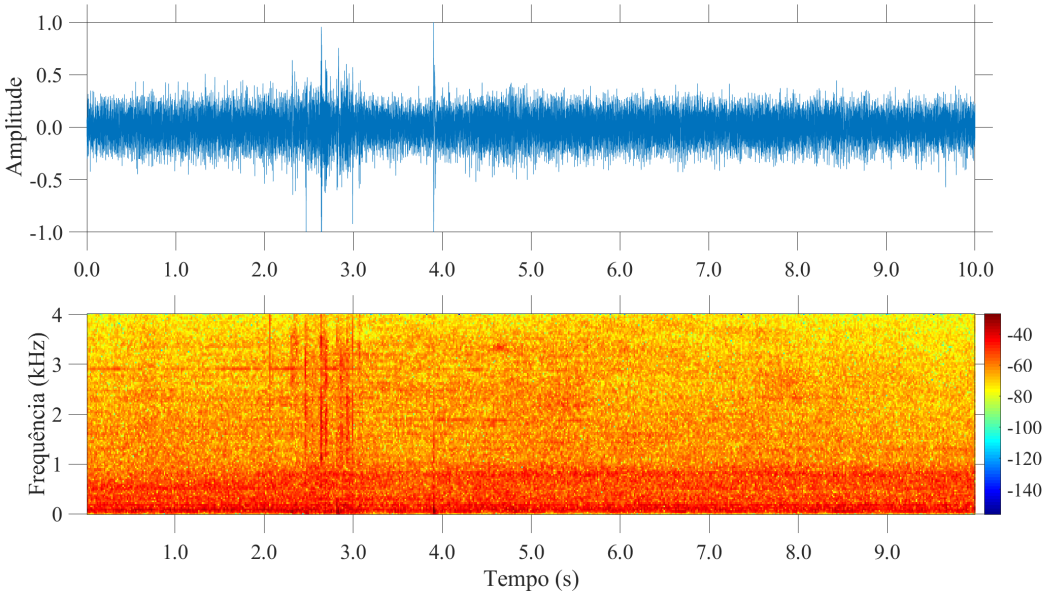
SNR/ tipo de ruído	0 dB	5 dB	7 dB	10 dB	15 dB	SNR aleatória (0 a 15 dB)	duração total
balbucios	1:26:26	1:26:26	1:26:26	1:26:26	1:26:26	1:26:26	8:38:36
carro	1:26:26	1:26:26	1:26:26	1:26:26	1:26:26	1:26:26	8:38:36
trem	1:26:26	1:26:26	1:26:26	1:26:26	1:26:26	1:26:26	8:38:36
aeroporto	1:26:26	1:26:26	1:26:26	1:26:26	1:26:26	1:26:26	8:38:36
duração total	5:45:44	5:45:44	5:45:44	5:45:44	5:45:44	5:45:44	34:34:24

Fonte: elaborado pelo autor.

PEARCE, 2000). Os sinais dos ruídos carro, balbucios, trem e aeroporto, no domínio do tempo (gráfico superior) e seu respectivo espectrograma (gráfico inferior) são apresentados nas Figuras 18, 19, 20 e 21, respectivamente. Os ruídos trem e balbucios possuem comportamentos estacionários. Já os ruídos carro e aeroporto contém segmentos não estacionários (entre $2,5$ s e 4 s para o ruído carro e entre 2 e 5 s para o ruído aeroporto). cada um dos ruídos possui duração de 10 s. Assim, uma vez que um determinado sinal de voz possui duração menor do que o tempo de duração do ruído, seleciona-se um início aleatório entre 0 e (10 - Δt) s, em que Δt é a duração do sinal de voz, com o objetivo de evitar o sobre-ajuste da RNP a um determinado período do ruído. Dito isso, somente uma quantidade aleatória de sinais ruidosos com o ruído carro e ruído aeroporto possui pulsos de curta duração.

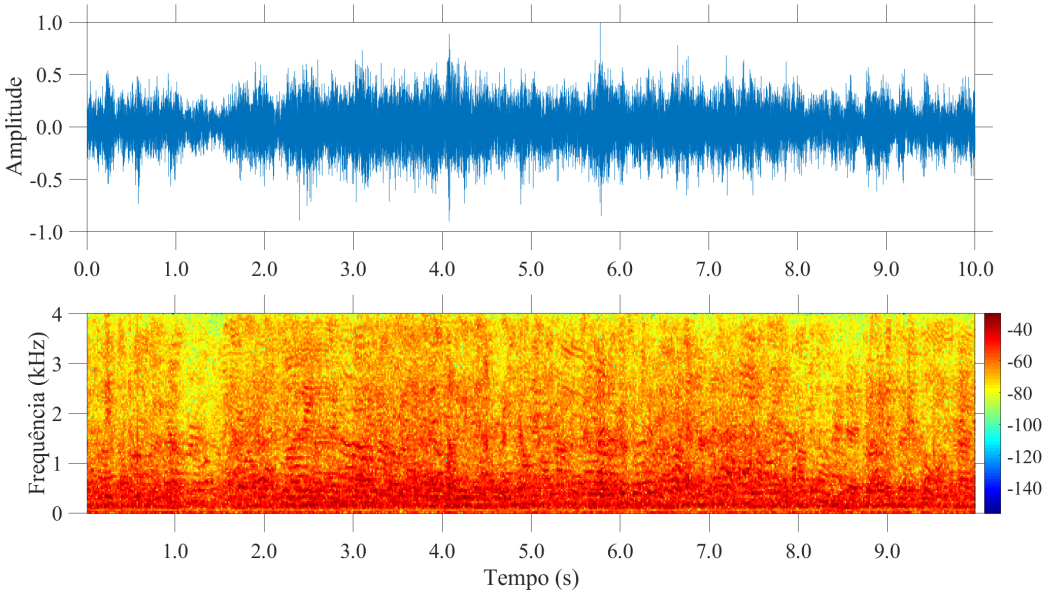
O conjunto de treinamento com múltiplas condições resultante é composto por quase 60 horas de áudio, enquanto o conjunto de validação e o conjunto de teste é composto por cerca de 17 horas e 17 minutos cada um.

Figura 18 – Forma de onda do ruído carro no domínio do tempo (gráfico superior) e seu respectivo espectrograma (gráfico inferior).



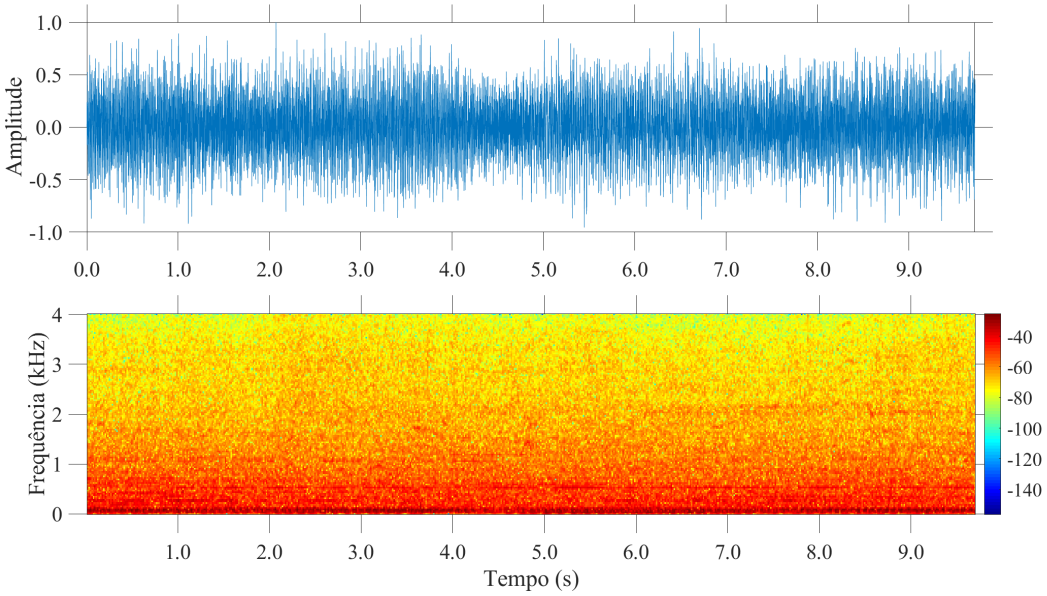
Fonte: elaborado pelo autor.

Figura 19 – Forma de onda do ruído balbucios no domínio do tempo (gráfico superior) e seu respectivo espectrograma (gráfico inferior).



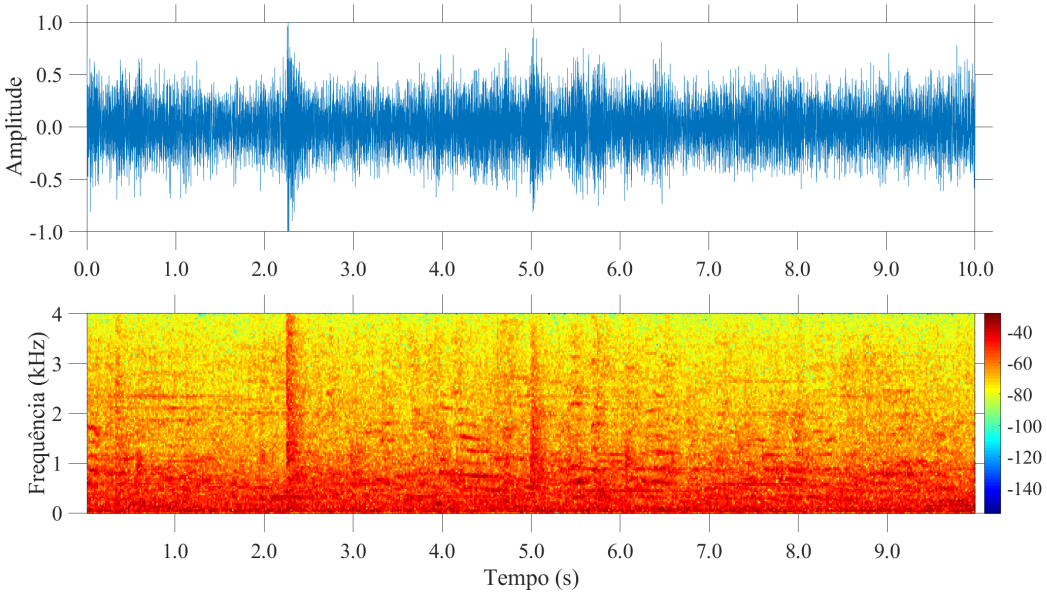
Fonte: elaborado pelo autor.

Figura 20 – Forma de onda do ruído trem no domínio do tempo (gráfico superior) e seu respectivo espectrograma (gráfico inferior).



Fonte: elaborado pelo autor.

Figura 21 – Forma de onda do ruído aeroporto no domínio do tempo (gráfico superior) e seu respectivo espectrograma (gráfico inferior).



Fonte: elaborado pelo autor.

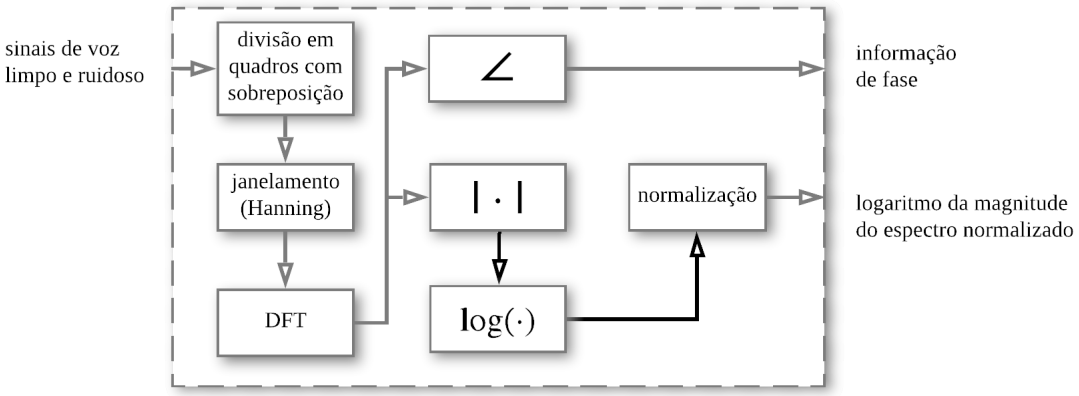
4.3 Extração de características

Cada arquivo de áudio do banco de dados de treinamento, validação e teste é reamostrado para 8 kHz para reduzir a quantidade de dados e, consequentemente, o tempo de treinamento da RNP e processamento dos dados. Como a maior parte da energia do sinal de voz se situa no intervalo de frequência entre 300 Hz e 4 kHz (limiar do critério de Nyquist-Shannon), as componentes de frequência descartadas acima de 4 kHz não comprometem de forma significativa a inteligibilidade das elocuções. Posteriormente, o sinal é segmentado em quadros de 32 ms (256 amostras por quadro) com sobreposição de 50%. Como dito no Capítulo 2, em quadros de 20 ms a 40 ms, as propriedades da voz apresentam poucas mudanças e o sinal pode ser considerado quase estacionário.

As operações de sobreposição e janelamento são utilizadas para reduzir o efeito das descontinuidades causadas pelo processo de segmentação, utilizando-se a janela de Hanning. Sobre cada quadro, aplica-se a TDF empregando a função FFT$(\cdot)$ do programa MATLAB (MATHWORKS, 2019). Tal função se baseia na implementação do algoritmo da transformada rápida de Fourier proposta por Frigo e Johnson (2015). O comprimento da TDF possui o mesmo tamanho do quadro ($N_s = N_{FFT} = 256$ amostras). É importante frisar que a TDF é simétrica em relação ao eixo y para representar as frequências negativas (a transformada de Fourier é definida no intervalo de $-\infty$ a $+\infty$). Por conta disso, são processados somente os $\frac{N_{FFT}}{2} + 1 = 129$ pontos de frequência. Para obter a magnitude e a fase, utiliza-se as funções ABS$(\cdot)$ e ANGLE$(\cdot)$, respectivamente. Em seguida, aplica-se o logaritmo sobre a magnitude do espectro para evitar a saturação das funções de ativação nas camadas ocultas. O vetor de características de voz resultante $\mathbf{O}$ possui dimensão 129×1.

Cada vetor de características é normalizado utilizando a média μ_X e a variância σ_x, calculadas a partir do conjunto de dados de treinamento, da seguinte forma $\mathbf{x}_{normalizado} = \frac{\mathbf{x}-\mu_\mathbf{X}}{\sigma_x}$. A normalização reduz a variabilidade do conjunto de dados, aproxima a média do conjunto de dados de treinamento para zero e evita a saturação das funções de ativação nas camadas ocultas. A expansão de quadros é utilizada para capturar o contexto espectral entre sequências de quadros adjacentes. Para uma expansão de quadros de tamanho três, o vetor de características resultante $\mathbf{O}_{expansao}$ possui dimensão $(3 \times 129, 1)$=$(387, 1)$, isto é, é adicionado um quadro à esquerda e um quadro à direita em relação ao quadro que está sendo atualmente processado. A sequência de passos para a extração do logaritmo da magnitude do espectro de frequência do sinal de voz normalizado é ilustrado em diagrama de blocos na Figura 22.

Figura 22 – Etapa de extração de características.

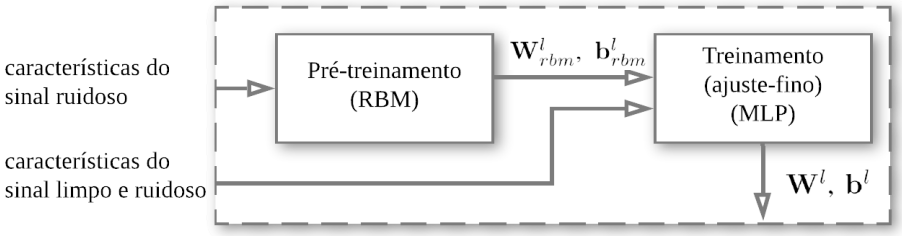


Fonte: elaborado pelo autor.

4.4 Rede neural profunda

A proposta de utilização de RNP para melhoria da qualidade do sinal de voz consiste em processar o logaritmo da magnitude do espectro de frequência do sinal ruidoso (entrada), de tal forma que se possa estimar o logaritmo da magnitude do espectro do sinal limpo (saída). A arquitetura utilizada para a RNP é uma estrutura de múltiplas camadas de perceptrons convencional com as unidades totalmente conectadas, como apresentado na Figura 15, cujos estágios envolvidos no treinamento da RNP são ilustrados em diagrama de blocos na Figura 23.

Figura 23 – Etapa de treinamento da RNP.



Fonte: elaborado pelo autor.

4.4.1 Pré-treinamento

O algoritmo (1) é usado para realizar o treinamento inicial dos parâmetros $\{\mathbf{W}, \mathbf{b}\}$ baseado em MRBs. Para uma arquitetura com três camadas escondidas, empilha-se uma MRB Gaussiana-Bernoulli e duas MBRs Bernoulli-Bernoulli. Para o processo de aprendizagem, utiliza-se uma taxa de aprendizagem de $\lambda = 0.0008$ para todas as MRBs e um tamanho de

mini-grupo M_b de 32 amostras. O momento inicial é de 0,5 para as primeiras 10 épocas e 0,9 para as 10 épocas restantes, totalizando $N_{epocas} = 20$ épocas, conforme sugerido por (XU *et al.*, 2015). Realiza-se o pré-treinamento em grupos de 16 arquivos por conta da limitação dos recursos computacionais disponíveis. A matriz de pesos é inicializada com um desvio padrão de $0,01$, enquanto os *bias* das camadas ocultas são inicializados com 0s. Este valor para o desvio-padrão permite que cada neurônio opere na escala linear da função sigmoide no início do treinamento (YU; DENG, 2015), já que valores altos para o desvio padrão saturam (com um ou zero) vários neurônios.

O algoritmo (1) apresentado a seguir é adaptado e traduzido de (YU; DENG, 2015). $\mathbb{S}$ é o espaço amostral, $\mathbf{O}^m$ é uma observação com m amostras (características do sinal ruidoso), $\mathbf{V^0}$ é a camada visível inicial, $\mathbf{H^0}$ é a camada escondida inicial, $P(\mathbf{H}|\mathbf{V^0})$ é a probabilidade a posteriori de $\mathbf{H}$ dado $\mathbf{V^0}$ e P($\mathbf{H}|\mathbf{V^{n+1}}$) é a probabilidade a posteriori de $\mathbf{H}$ dado $\mathbf{V^{n+1}}$. $\nabla_{\mathbf{W}}J$, $\nabla_{\mathbf{a}}J$ e $\nabla_{\mathbf{b}}J$ são os gradientes do erro quadrático médio com relação à $\mathbf{W}$, $\mathbf{a}$ e $\mathbf{b}$, respectivamente. I é a matriz identidade. $\Delta\mathbf{W}_t$, $\Delta\mathbf{a}_t$ e $\Delta\mathbf{b}_t$ são os gradientes médios da matriz de pesos, do vetor da camada visível e do vetor da camada escondida, respectivamente.

4.4.2 Treinamento (ajuste fino)

O algoritmo de retropropagação[2] é utilizado para segundo estágio de aprendizagem dos parâmetros $\{\mathbf{W}, \mathbf{b}\}$ da RNP. A sigmoide é utilizada como função de ativação dos neurônios nas camadas ocultas, enquanto a função linear é utilizada como função de ativação na camada de saída. São utilizadas 20 épocas para o treinamento. Diversas configurações da RNP contendo uma, duas e três camadas escondidas com diferentes quantidades de neurônios são avaliadas. O melhor resultado é alcançado com três camadas escondidas contendo 2048 neurônios em cada uma das camadas. Além disso, obtém-se um desempenho maior, isto é, uma convergência mais rápida com menor erro quadrático médio, quando se utiliza taxas de aprendizagem λ^l diferentes para cada uma das l camadas. A melhor configuração encontrada consiste nas seguintes taxas de aprendizagem: $\lambda^1 = 0.0512$, $\lambda^2 = 0.0256$, $\lambda^3 = 0.0128$, $\lambda^4 = 0.00016$. Um momento ρ_o inicial de 0.5 é utilizado para as 10 primeiras épocas e um momento final de 0.9 para as 10 últimas épocas. Por conta das limitações dos recursos computacionais, restringe-se o treinamento a grupos (lotes) de 16 arquivos de áudio. Os algoritmos (2) e (3) apresentam os códigos-base utilizados para o desenvolvimento dos estágios de propagação e retropropagação, respectivamente.

[2] Expressão traduzida do inglês *backpropagation*

Algorithm 1 Divergência contrastiva para treinar MRB.

função TRAINMRBWITHCD($\mathbb{S}$=$\mathbf{O}^m \leq m \leq M$, N)

▷ $\mathbb{S}$ é o conjunto de treinamento com M amostras

▷ N é o número de passos do algoritmo divergência contrastiva

Inicializar as matrizes de peso $\mathbf{W}^l$ de forma aleatória com distribuição Gaussiana com desvio padrão α =0,01.

Inicializar os vetores da camada visível **a** e da camada escondida **b** com zero.

Inicializar os gradientes $\nabla_{\mathbf{W}}J$, $\nabla_{\mathbf{a}}J$ e $\nabla_{\mathbf{b}}J$ com zero.

Laço de repetição ($epoca = 1; epoca < N_{epocas}; epoca \leftarrow epoca + 1$) **faça**

Selecionar aleatoriamente **O** mini grupos com $\mathbf{M}_b$ amostras.

$\mathbf{V}^0 \leftarrow \mathbf{O}$ ▷ Fase positiva

$\mathbf{H}^0 \leftarrow P(\mathbf{H} = 1|\mathbf{V}^0)$ ▷ Eq. (2.28)

$\nabla_{\mathbf{W}}J \leftarrow \mathbf{H}^0(\mathbf{V}^0)^T$

$\nabla_{\mathbf{a}}J \leftarrow$ SOMAR LINHAS($\mathbf{V}^0$)

$\nabla_{\mathbf{b}}J \leftarrow$ SOMAR LINHAS($\mathbf{H}^0$)

Laço de repetição n ← 0; n < N; n ← n+1 **faça** ▷ Fase negativa

$\mathbf{H}^n \leftarrow \mathbb{I}(\mathbf{H}^n > aleatorio(0,1))$ ▷ Amostragem, $\mathbb{I}(\cdot)$ é a função característica

$\mathbf{V}^{n+1} \leftarrow P(\mathbf{V}|\mathbf{H}^n)$ ▷ Eq. (2.29) se $\mathbf{v} \in \mathbb{R}^{N_v \times 1}$; Eq. (2.30) se $\mathbf{v} \in \{0,1\}^{N_v \times 1}$

$\mathbf{H}^{n+1} \leftarrow$ P($\mathbf{H} = 1|\mathbf{V}^{n+1}$) ▷ eq. (2.28)

Fim Laço de repetição

$\nabla_{\mathbf{W}}J_t \leftarrow \nabla_{\mathbf{W}}J_{t-1} - \mathbf{H}^N(\mathbf{V}^N)^T$ ▷ Eq. (2.37)

$\nabla_{\mathbf{a}}J_t \leftarrow \nabla_{\mathbf{a}}J_{t-1} -$ SOMAR LINHAS($\mathbf{V}^N$) ▷ Eq. (2.38)

$\nabla_{\mathbf{b}}J_t \leftarrow \nabla_{\mathbf{b}}J_{t-1} -$ SOMAR LINHAS($\mathbf{H}^N$) ▷ Eq. (2.39)

$$\Delta W_t \leftarrow \rho \Delta W_{t-1} + (1-\rho)\frac{1}{M_b}\sum_{m=1}^{M_b} \nabla_{\mathbf{W_t}} J_{NLM}(\mathbf{W}, \mathbf{a}, \mathbf{b}; \mathbf{v}^m) \quad \triangleright \text{ Eq. (2.34)}$$

$$\Delta a_t \leftarrow \rho \Delta a_{t-1} + (1-\rho)\frac{1}{M_b}\sum_{m=1}^{M_b} \nabla_{\mathbf{a_t}} J_{NLM}(\mathbf{W}, \mathbf{a}, \mathbf{b}; \mathbf{v}^m) \quad \triangleright \text{ Eq. (2.35)}$$

$$\Delta b_t \leftarrow \rho \Delta b_{t-1} + (1-\rho)\frac{1}{M_b}\sum_{m=1}^{M_b} \nabla_{\mathbf{b_t}} J_{NLM}(\mathbf{W}, \mathbf{a}, \mathbf{b}; \mathbf{v}^m) \quad \triangleright \text{ Eq. (2.36)}$$

$W_{t+1} \leftarrow W_t - \lambda \Delta W_t$ ▷ Atualizar **W**. Eq. (2.34)

$a_{t+1} \leftarrow a_t - \lambda \Delta a_t$ ▷ Atualizar **a**. Eq. (2.35)

$b_{t+1} \leftarrow b_t - \lambda \Delta b_t$ ▷ Atualizar **b**. Eq. (2.36)

Fim Laço de repetição

Retorne *rbm*= $\{\mathbf{W}, \mathbf{a}, \mathbf{b}\}$

Fim função

Fonte: (YU; DENG, 2015, p.87, adaptação e tradução nossa).

Algorithm 2 Propagação

função FORWARDCOMPUTATION($\mathbf{O}$)
$\mathbf{V}^0 \leftarrow \mathbf{O}$ ▷ Cada coluna de $\mathbf{O}$ é um vetor de observação
Laço de repetição ($l \leftarrow$; $l < L$; $l \leftarrow l+1$) **faça** ▷ L é o número total de camadas
$\mathbf{Z}^l \leftarrow \mathbf{W}^l\mathbf{V}^{l-1} + \mathbf{b}^l$
$\mathbf{V}^l \leftarrow f(\mathbf{Z}^l)$ ▷ f(·) é a função sigmoide
Fim Laço de repetição
$\mathbf{Z}^L \leftarrow \mathbf{W}^L\mathbf{V}^{L-1} + \mathbf{b}^l$
$\mathbf{V}^L \leftarrow \mathbf{Z}^L$
Retorne $\mathbf{V}^L$
Fim função

Fonte: (YU; DENG, 2015, p.59, adaptação e tradução nossa).

Algorithm 3 Retropropagação

função BACKPROPAGATION($\mathbb{S} = \{(\mathbf{o}^m, \mathbf{d}^m) | 0 \leq m \leq M\}$)
▷ $\mathbb{S}$ é o conjunto de treinamento com M amostras
▷ Parâmetros $\{\mathbf{W}_0^l, \mathbf{b}_0^l\}$ provenientes do pré-treinamento, em que $0 < l \leq L$
▷ L é o número total de camadas

Enquanto o total de épocas não for atingido **faça**
Selecione aleatoriamente um mini grupo $\{\mathbf{O}, \mathbf{Y}\}$ com M_b amostras
Chamar função FORWARDCOMPUTATION($\mathbf{O}$)
$\mathbf{E}_t^L \leftarrow \mathbf{V}_t^L - \mathbf{D}$ ▷ Eq. (2.47)
$\mathbf{G}_t^L \leftarrow \mathbf{E}_\mathbf{t}^\mathbf{L}$
Laço de repetição ($l \leftarrow L$; $l > 0$; $l \leftarrow l-1$) **faça**
$\nabla_{\mathbf{W}_t^l} J_{EQM}(\mathbf{W}, \mathbf{b}; \mathbf{x}, \mathbf{d}) \leftarrow \mathbf{G}_t^l \left(\mathbf{v}_t^{l-1}\right)^T$ ▷ Eq. (2.46)
$\nabla_{\mathbf{b}_t^l} J_{EQM}(\mathbf{W}, \mathbf{b}; \mathbf{x}, \mathbf{d}) \leftarrow$ SOMAR LINHAS$(\mathbf{G}_t^l)$ ▷ Eq. (2.48)
$\mathbf{W}_{t+1}^l \leftarrow \rho\mathbf{W}_t^l - (1-\rho)\lambda^l \frac{1}{M_b}\left(\nabla_{\mathbf{W}_t^l} J_{EQM}(\mathbf{W}, \mathbf{b}; \mathbf{x}, \mathbf{d}) + \gamma\nabla_{\mathbf{W_t}}^l \mathbf{R}\right)$ ▷ Eq. (2.42)
$\mathbf{b}_{t+1}^l \leftarrow \rho\mathbf{b}_t^l - (1-\rho)\lambda^l \frac{1}{M_b}\nabla_{\mathbf{b}_t^l} J_{EQM}(\mathbf{W}, \mathbf{b}; \mathbf{x}, \mathbf{d})$ ▷ Eq. (2.43)
$\mathbf{E}_t^{l-1} \leftarrow (\mathbf{W}_t^l)^T \mathbf{G}_t^l$ ▷ Eq. (2.52)
Se $l > 1$ **então**
$\mathbf{G}_t^{l-1} \leftarrow f'(\mathbf{Z}_t^{l-1}) \bullet \mathbf{E}_t^{l-1}$ ▷ Eq. (2.53)
Fim Se
Fim Laço de repetição
Fim Enquanto
Retorne *dnn*=$\{\mathbf{W}^l, \mathbf{b}^l\}$ ▷ em que $0 \leq l \leq L$
Fim função

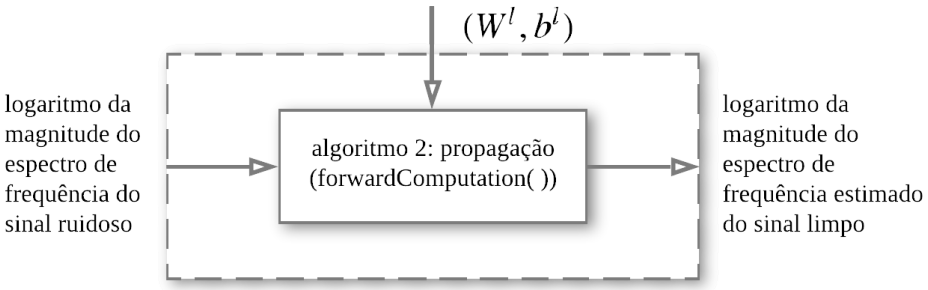
Fonte: (YU; DENG, 2015, p.66, adaptação e tradução nossa).

O Apêndice B apresenta os parâmetros e hiperparâmetros utilizados na configuração da RNP.

4.4.3 Decodificação

O processo de decodificação da RNP é ilustrado na Figura 24. Uma vez que os parâmetros **W** e **b** são obtidos na etapa de treinamento, obtém-se a estimativa do logaritmo da magnitude do espectro do sinal limpo a partir do logaritmo da magnitude do espectro de frequência do sinal ruidoso usando o algoritmo (2). Dada uma observação **O**, a saída da rede neural é obtida calculando os vetores de ativação utilizando a equação (2.22) para as camadas 1 a $L-1$ e a equação (2.24) para a camada de saída L.

Figura 24 – Etapa de decodificação da RNP.



Fonte: elaborado pelo autor.

4.5 Reconstrução do sinal

O bloco de reconstrução do sinal de voz é apresentado na Figura 25. A desnormalização do logaritmo da magnitude do espectro de frequência (i.e., características) do sinal de voz é realizada com a média μ_x e a variância σ_x do conjunto de dados de treinamento. Em seguida, aplica-se a operação de exponenciação (processo inverso à operação de logaritmo aplicada na etapa de extração de características).

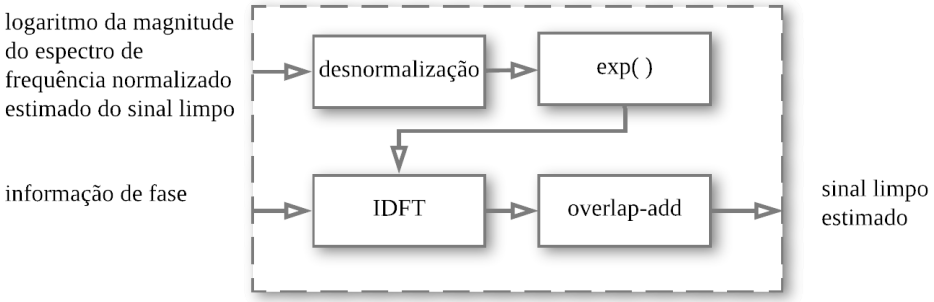
Neste estágio, deve-se espelhar os $\frac{N_{FFT}}{2} = 128$ pontos de frequência, a fim de representar as frequências negativas e transformar corretamente o quadro para o domínio do tempo ($N_s = 256$). A informação de fase do sinal ruidoso, obtida a partir da TDF na etapa de extração de características, é usada para calcular a Transformada Discreta de Fourier Inversa (TDFI). Loizou (2013, p.106) afirma que o uso da informação de fase é considerada uma prática aceita

na reconstrução da estimativa do sinal limpo, porém, alguns autores discutem a importância da estimação de fase do sinal limpo, especialmente em sinais com baixas SNRs (GERKMANN *et al.*, 2012) (GERKMANN *et al.*, 2015).

De posse da estimativa do logaritmo da magnitude do espectro de frequência do sinal limpo e a informação de fase proveniente do sinal ruidoso, aplica-se a TDFI sobre cada vetor de saída para produzir uma estimativa dos quadros do sinal limpo no domínio do tempo.

Finalmente, o método soma e sobreposição, descrito na Seção 2.3.4 e proposto em (ALLEN, 1977), é empregado para sobrepor os quadros e obter uma estimativa do sinal limpo completo no domínio do tempo.

Figura 25 – Etapa de reconstrução do sinal.



Fonte: elaborado pelo autor.

4.6 Métricas de desempenho

Algoritmos de melhoria do sinal de voz objetivam aumentar a qualidade e a inteligibilidade do sinal. Enquanto a inteligibilidade se refere à capacidade de um grupo de ouvintes de escutar e entender as palavras proferidas, a qualidade se refere à avaliação dos sinais de voz por um grupo de ouvintes em uma escala pré-determinada ou à avaliação comparativa com sinais de referência. Loizou (2013) ressalta o compromisso existente entre a qualidade e a inteligibilidade do sinal de voz da seguinte maneira:

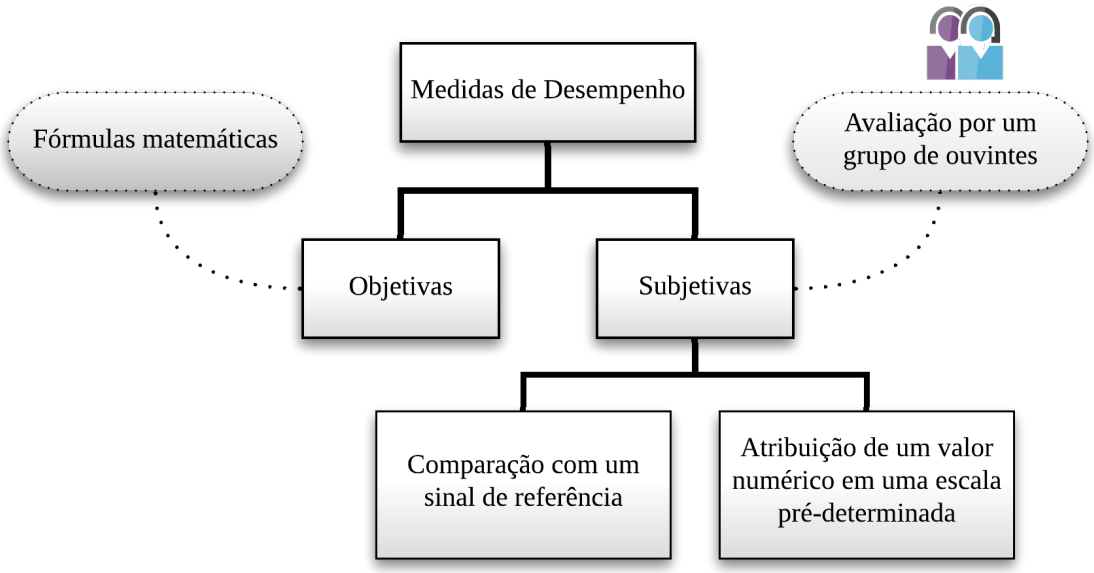
> É possível reduzir o ruído de fundo ao custo de introduzir distorção na voz, o que por sua vez, pode prejudicar a inteligibilidade. Assim, o principal desafio no desenvolvimento de algoritmos de aprimoramento do sinal de fala consiste reduzir o ruído sem introduzir qualquer distorção perceptível no sinal. (LOIZOU, 2013, p.1,tradução nossa).

Para verificar se o objetivo de atenuar o ruído aditivo de fundo e estimar o sinal de voz desejado limpo foi alcançado, após a aplicação dos algoritmos, é necessário estabelecer critérios de desempenho. Segundo Benesty *et al.* (2007), existem duas categorias de medidas de desempenho: subjetivas e objetivas. A Figura 26 apresenta um sumário de categorização das medidas de desempenho.

As métricas subjetivas se baseiam nos julgamentos de um grupo de pessoas para avaliar a qualidade e a inteligibilidade dos sinais de voz, baseado na habilidade de percepção, discriminação e experiência. Métricas subjetivas podem ser classificadas em duas categorias: aquelas baseadas numa preferência relativa em comparação a um dado sinal de referência e aquelas baseados na atribuição de um valor numérico indicando a percepção de qualidade do sinal de voz. Já as métricas objetivas envolvem fórmulas matemáticas, desenvolvidas a partir de técnicas de processamento de sinal, para estimar a qualidade, a inteligibilidade e o grau de correlação com as métricas subjetivas dos sinais de voz e/ou avaliar diferenças entre o sinal de voz ruidoso, o de referência e o processado.

De acordo com Benesty *et al.* (2007), no que diz respeito à qualidade do sinal de voz, o método subjetivo é mais adequado, uma vez que o julgamento do ouvinte é o considerado em última análise. Contudo, a avaliação subjetiva demanda esforço e tempo e os resultados são

Figura 26 – Classificação de medidas de desempenho: medidas objetivas utilizam fórmulas matemáticas, enquanto as subjetivas são avaliadas por um grupo de ouvintes.



Fonte: elaborado pelo autor.

considerados onerosos de se obter, uma vez que é necessária a disponibilidade de um grupo de avaliadores e o estabelecimento de condições controladas para avaliação. Por conta disso, a utilização de métricas objetivas é mais econômica, prática e mais utilizada. Hu e Loizou (2007) realizam um estudo comparativo de diversas métricas objetivas e avaliam a correlação com a inteligibilidade e métodos subjetivos.

Para esta dissertação, utilizam-se quatro métricas objetivas comumente utilizadas na literatura (LOIZOU, 2013) (XU *et al.*, 2014) (XU *et al.*, 2015): a distorção log-espectral (LSD)[3], a STOI, a PESQ e a TEP. As métricas PESQ e STOI possuem uma alta correlação com a inteligibilidade e testes subjetivos. A métrica LSD, por sua vez, aborda o aspecto da distorção espectral, complementar às outras duas métricas. Já a TEP avalia o aumento ou decaimento na taxa de reconhecimento de palavras em um SRAV, relativo à melhoria ou degradação de algum aspecto perceptivo da qualidade e/ou inteligibilidade de sinais de voz.

4.6.1 LSD

A LSD (GRAY *et al.*, 1980) mede o grau de distorção espectral entre o sinal de voz de referência e o sinal de voz degradado. Quanto mais próximo a zero, menor a distorção e, consequentemente, maior a qualidade. Ambos os sinais são segmentados em quadros de 25 ms com 60% de sobreposição (15 ms). Para a operação de janelamento, utiliza-se a janela de Hamming. Os momentos de silêncio são removidos, a partir de um detector de atividade de voz (VAD)[4]. A LSD calcula a média da distância euclidiana entre a magnitude do espectro de frequência dos quadros do sinal de voz de referência e dos quadros do sinal de voz degradado. Em situações extremas encontradas na literatura, foi encontrado o valor LSD de 11,14 para um sinal de voz com uma SNR de -5 dB. Embora seja possível encontrar valores maiores, limita-se o valor da LSD nesta dissertação ao intervalo entre 0 (melhor caso) e 20 (extremamente distorcido em frequência).

4.6.2 STOI

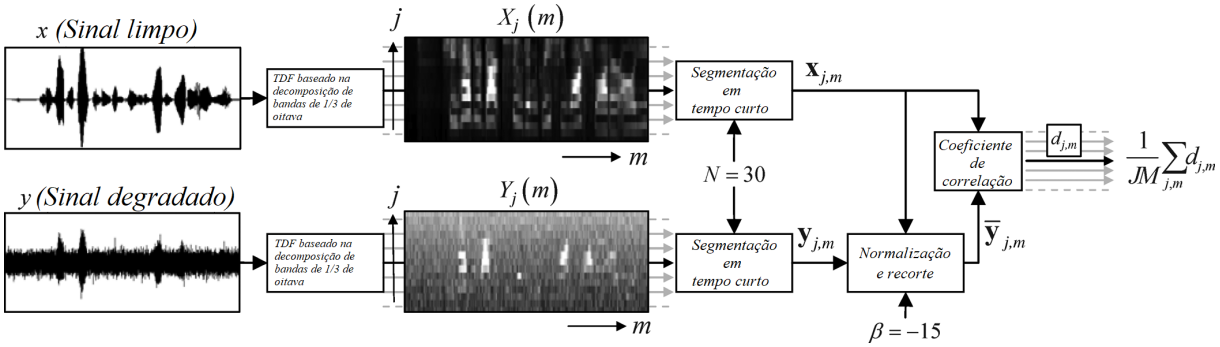
Segundo (TAAL *et al.*, 2010b), a STOI é uma métrica objetiva que possui correlação de 95% com a inteligibilidade, conceito relativo ao percentual médio de palavras entendidas corretamente por um grupo de pessoas. Esta métrica foi desenvolvida motivada pela necessidade

[3] Abreviatura do inglês *log-spectral distortion*.
[4] Abreviatura do inglês *voice activity detector*.

de substituir testes subjetivos, pois estes são custosos tanto do ponto de vista financeiro quanto do ponto de vista temporal. Varia numa escala de zero a um, em que zero significa um sinal de voz ininteligível e 1 um sinal de voz completamente inteligível.

A estrutura básica da métrica STOI é apresentada na Figura 27. A medida é função de um sinal de voz limpo de referência e de um sinal de voz a ser avaliado (sinal de voz degradado). Ambos os sinais são divididos em bandas de um terço de oitava, baseado na TDF, normalizados e recortados. Em seguida, segmentos curtos no tempo (384 ms) de ambos os sinais são comparados por meio de um coeficiente de correlação. Em seguida, é calculada a média das medidas de inteligibilidade intermediárias de curto tempo, resultando em um valor escalar, que possui uma relação monotônica crescente com a inteligibilidade da voz (TAAL *et al.*, 2011).

Figura 27 – Estrutura básica da métrica STOI.



Fonte: (TAAL *et al.*, 2011, tradução nossa).

A medida de inteligibilidade intermediária é definida como uma estimativa do coeficiente de correlação linear entre as unidades tempo-frequência do sinal limpo $\mathbf{x_j(m)}$ (definida na notação vetorial) e do sinal de voz degradado ($\mathbf{y_j(m)}$) (TAAL *et al.*, 2011)

$$d_{j,m} = \frac{(\mathbf{x}_{j,m} - \mu_{\mathbf{x}_{j,m}})^T (\overline{\mathbf{y}}_{j,m} - \mu_{\overline{\mathbf{y}}_{j,m}})}{||\mathbf{x}_{j,m} - \mu_{\mathbf{x}_{j,m}}|| \cdot ||\overline{\mathbf{y}}_{j,m} - \mu_{\overline{\mathbf{y}}_{j,m}}||}, \tag{4.1}$$

em que m representa o m-ésimo quadro, $\mathbf{x}_{j,m}$ representa o envelope de tempo curto do sinal de voz limpo, $\mathbf{y}_{j,m}$ representa o envelope de tempo curto do sinal de voz a ser avaliado, $\overline{\mathbf{y}}_{j,m}$ é uma versão normalizada e recortada de y, $\mu_{(\cdot)}$ representa a média das amostras do vetor correspondente e j representa a j-ésima banda de um terço de oitava, $||\cdot||$ representa a norma l_2. Finalmente, a média da inteligibilidade intermediária é calculada em todas as bandas e quadros

$$d = \frac{1}{J \cdot M} \sum_j \sum_m \mathbf{d_{j,m}} \tag{4.2}$$

em que M representa a quantidade total de quadros e J o número de bandas de 1/3 oitava. O código na plataforma MATLAB é disponibilizado pelos autores em (TAAL *et al.*, 2010a).

4.6.3 PESQ

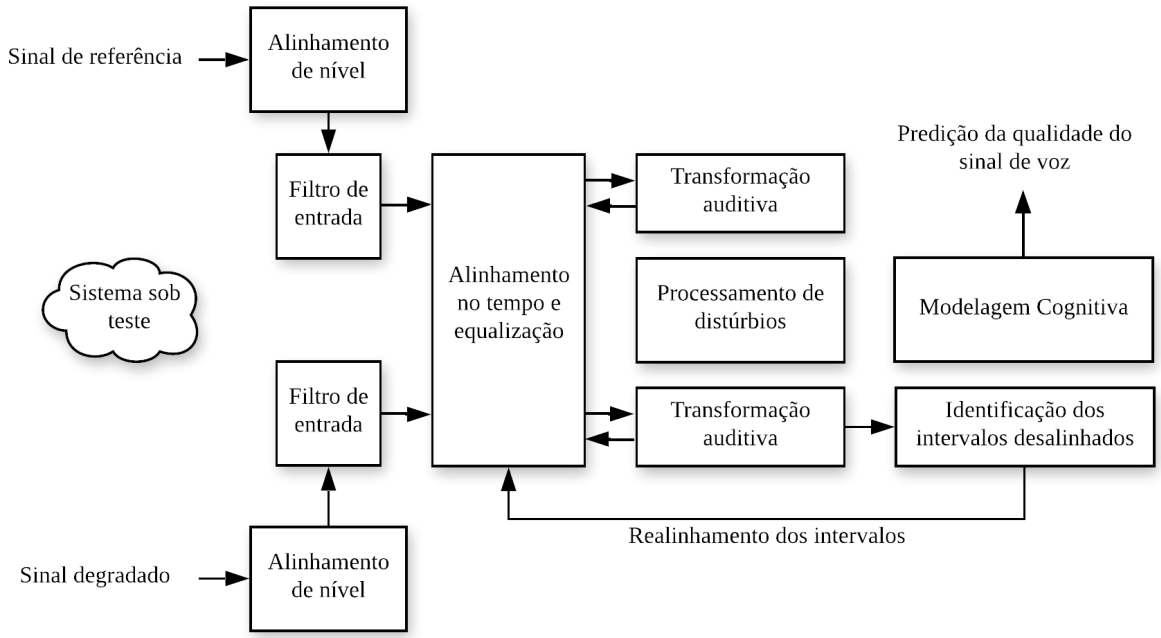
No ano 2000, foi realizada uma competição organizada pelo ITU-T para selecionar uma métrica objetiva capaz de apresentar um desempenho confiável em uma vasta quantidade de *codecs* (codificadores/decodificadores) e condições de rede (LOIZOU, 2013). A métrica PESQ (avaliação perceptiva da qualidade da voz) foi selecionada e padronizada pelo ITU-T na recomendação P.862 em substituição à recomendação P.861, que era baseada na métrica *Perceputal Speech Quality Measure* (PSQM) (percepção de qualidade de voz) (BEERENDS; STEMERDINK, 1994). A PESQ é uma métrica objetiva que possui alta correlação com métricas subjetivas (RIX *et al.*, 2001), é baseada nas características psicoacústicas do ouvido humano e varia numa escala de um a cinco, em que um é considerado sinal de qualidade insuficiente e cinco uma qualidade excelente.

A estrutura desta métrica é apresentada na Figura 28. De acordo com Rix *et al.* (2001), o modelo inicia pelo alinhamento de ambos os sinais a um nível de audição padrão. Estes são filtrados (usando uma FFT) com um filtro de entrada para modelar um aparelho de telefone. Os sinais são alinhados no tempo e depois processados através de uma transformação auditiva semelhante à da métrica PSQM. A transformação também envolve a equalização para a filtragem linear no sistema e para a variação do ganho. Dois parâmetros de distorção são extraídos dos distúrbios (a diferença entre as transformações dos sinais) e são agrupados no domínio do tempo e frequência e mapeados para uma previsão opinião média subjetiva (MOS)[5].

Em uma avaliação de desempenho de sete métricas objetivas [segSNR, *Weighted Spectral Slope* (WSS), PESQ, *Log-Likelihood Ratio* (LLR), distância Distância *Itakura-Saito* (IS), distância Cepstrum e fwSNRseg] em termos de predição de qualidade do aprimoramento do sinal de voz ruidoso, Hu e Loizou (2007) demonstraram que a PESQ possui a maior correlação ($\rho = 0,89$) com a qualidade geral dentre as métricas avaliadas.

[5] Acrônimo do inglês *Subjective Opinion Score*.

Figura 28 – Estrutura básica da métrica PESQ.



Fonte: (RIX *et al.*, 2001, tradução nossa).

4.6.4 TEP

A Taxa de Erro de Palavra (TEP[6]) é uma métrica que indica o percentual de palavras incorretas reconhecidas por um SRAV. É definida por Ali e Renals (2018)

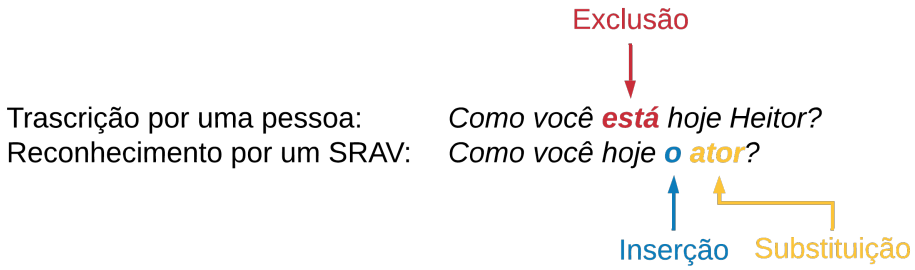
$$TEP = \frac{N_{sub} + N_{exc} + N_{ins}}{N_{total}} \cdot 100\%, \tag{4.3}$$

em que N_{sub}, N_{exc}, N_{ins} e N_{total} são, respectivamente, o número de palavras substituídas, o número de palavras excluídas, o número de palavras inseridas e o número total de palavras na hipótese de transcrição com relação ao sinal de voz de referência. Um valor próximo a 100% indica que o sinal está bastante corrompido e o/ou o SRAV não foi capaz de transcrever corretamente as palavras. Um valor próximo a 0% indica que houve poucos erros, sejam inserções, exclusões e/ou substituições de palavras, no resultado do reconhecimento.

Para avaliação da TEP dos sinais de voz ruidosos e dos sinais de voz limpos estimados, utiliza-se a função SPEECH2TEXT, disponível em MathWorks, Audio Toolbox Team (2020). Tal função permite a conexão com interfaces de programação de aplicações (do inglês, *application programming interfaces*) baseadas em nuvem para reconhecimento de voz usando a plataforma *MATrix LABoratory* (MATLAB) da empresa MathWorks (MATHWORKS, 2019). As APIs

[6] Também é conhecida como WER, acrônimo do inglês *Word Error Rate*.

Figura 29 – Exemplos de erros no processo de reconhecimento de voz considerados no cálculo da TEP.



Fonte: elaborado pelo autor.

de reconhecimento de voz suportadas incluem: Google, IBM e Microsoft. Através da função, são enviados os áudios dos sinais de voz ruidosos e dos sinais limpos estimados para o sistema de reconhecimento automático em nuvem da empresa Google. A partir do recebimento da transcrição desse SRAV, compara-se a sentença reconhecida com a frase fornecida pelo banco de dados TIMIT, a fim de obter o resultado da TEP. Para o cálculo da TEP, utiliza-se a função WER, desenvolvida e disponibilizada por Polityko (2020).

Um quadro-resumo das métricas de desempenho utilizadas e seus respectivos intervalos de variação são apresentados na Tabela 5.

Tabela 5 – Sumário das métricas de avaliação de desempenho.

Métrica	Descrição	Intervalo
LSD	métrica objetiva que mede a distorção espectral	0 a 20*
STOI	métrica objetiva que possui correlação com o grau de inteligibilidade.	0 a 1,0
PESQ	métrica objetiva que possui correlação com métricas subjetivas	1,0 a 5,0
TEP	métrica objetiva que indica o percentual de palavras incorretas reconhecidas por um SRAV	0 a 100%

Fonte: elaborado pelo autor.

4.7 Ambiente de desenvolvimento

Os algoritmos são implementados em *scripts* na plataforma MATLAB da empresa MathWorks (MATHWORKS, 2019). Tivemos à nossa disposição um computador portátil com processador Intel Core i7 6700HQ, memória RAM de 16 GB (2x8GB, arquitetura de canal duplo), placa de vídeo NVIDIA GeForce GTX 960M, disco rígido com capacidade 1 TB e taxa média de leitura/escrita de 96 MB/s, e uma unidade de estado sólido[7] M.2 com capacidade de 256 GB e taxa média de leitura/escrita de aproximadamente 390 MB/s.

Utilizando o banco de dados completo, o tempo de treinamento da rede neural com três camadas ocultas é de aproximadamente 72 horas. O tempo necessário para avaliar o banco de dados de teste completo é de aproximadamente 67 horas. A demora no tempo de teste se deve principalmente à conexão com a API *Google Cloud Speech-to-Text* que é utilizada para avaliar a TEP. Cada arquivo de áudio leva cerca de 4 a 20 segundos (depende do tamanho do arquivo e da conexão de rede) para obter uma resposta de reconhecimento da API. Como o banco de dados completo é composto por 20.160 arquivos, resulta em um tempo médio de aproximadamente 67 horas. Ainda que de forma minoritária, o tempo computacional dos cálculos das métricas LSD, STOI e PESQ também contribuem para esse tempo elevado.

[7] *Solid-State Drive* (SSD).

5 RESULTADOS E DISCUSSÃO

Este Capítulo apresenta os principais resultados obtidos para os experimentos dos modelos subtração espectral, filtro de Wiener e RNP. A Seção 5.1 mostra o comparativo de desempenho dos resultados obtidos pelos modelos a partir das métricas objetivas LSD, STOI, PESQ e TEP. As Seções 5.2.1, 5.2.2 e 5.2.3 apresentam algumas estimativas de sinais limpo obtidos pelos algoritmos subtração espectral, filtro de Wiener e RNP, respectivamente. O Capítulo se encerra com uma discussão geral acerca dos resultados na Seção 5.3.

5.1 Avaliação de desempenho

Os sinais de voz ruidosos foram produzidos a partir dos áudios da base de dados TIMIT (GAROFOLO *et al.*, 1988) e os áudios dos ruídos da base de dados AURORA-2 (HIRSCH; PEARCE, 2000). Dos 6.300 áudios da base de dados TIMIT, 4.620 (73,3% do total) foram utilizados para o treinamento da RNP, 840 (13,33%) para o conjunto de validação e 840 para teste. Para áudio do conjunto de treinamento, acrescentam-se três tipo de ruídos (balbucios, trem e carro) e se estabelecem quatro níveis de SNR. Além disso, é acrescentada uma categoria com sinais de voz com SNRs aleatórias geradas com distribuição uniforme no intervalo entre 0 e 15 dB. No total, são estabelecidas 15 combinações (ruído e SNR) diferentes para o conjunto de treinamento. Para os conjuntos de validação e teste, são utilizados os mesmos cenários do conjunto de treinamento, acrescidos do ruído aeroporto e da SNR 7 dB, totalizando 24 cenários para o conjunto de validação e teste.

Os resultados das métricas LSD, STOI, PESQ e TEP estão presentes nas Tabelas 6, 7, 8 e 9, respectivamente. A RNP produz os menores valores para a métrica LSD, quando comparada aos algoritmos clássicos, incluindo os cenários SNR 7 dB e a categoria de SNRs aleatórias. Se por um lado o ruído aditivo é atenuado nos algoritmos clássicos, por outro, a distorção espectral aumenta. Devido a isso, a redução na LSD nos algoritmos clássicos não é tão significativa quanto na RNP. Em relação às categorias de SNR avaliadas, a redução média da RNP é de 35% e da subtração espectral 14%. Já no filtro de Wiener, há um pequeno aumento de 1%.

Em relação à métrica STOI, a RNP apresenta um aumento de 7% em relação ao sinal ruidoso, a subtração espectral uma redução de 1% e o filtro de Wiener uma redução de 4%. O maior ganho de inteligibilidade é obtido pela RNP na SNR 0 dB, com 17%. Nesse mesmo

Tabela 6 – Resultados da métrica LSD, obtidos pelos algoritmos para as diferentes categorias de SNR.

SNR	Sinal Ruidoso	Subtração Espectral	Filtro de Wiener	RNP
0 dB	1,75	1,43	1,62	**1,09**
5 dB	1,43	1,25	1,52	**0,95**
7 dB*	1,42	1,19	1,43	**0,90**
10 dB	1,26	1,10	1,34	**0,84**
15 dB	1,01	0,94	1,17	**0,75**
Aleatória* (dB)	1,38	1,19	1,41	**0,90**
Média	1,42	1,21	1,42	0,91

Fonte: elaborado pelo autor.

cenário, o filtro de Wiener apresenta o pior resultado, com redução de 6%. Na categoria de SNRs aleatórias, o ganho obtido é de 6% para a RNP. Já na SNR 15 dB, não houve aumento no valor da STOI.

Os resultados da Tabela 7 indicam que a RNP é capaz de aumentar o nível de inteligibilidade nas SNRs 0, 5 e 10 dB. Além disso, é eficaz mesmo em níveis de SNRs para os quais não é treinada, como em 7 dB e na categoria de SNRs aleatórias, o que comprova a capacidade de generalização da RNP. Contudo, em 15 dB, os resultados apontam que a RNP não produz aumento na inteligibilidade, o que indica uma possível limitação para a estrutura e a configuração da RNP utilizada nesta dissertação.

Tabela 7 – Resultados da métrica STOI, obtidos pelos algoritmos para as diferentes categorias de SNR.

SNR	Sinal Ruidoso	Subtração Espectral	Filtro de Wiener	RNP
0 dB	0,70	0,70	0,66	**0,82**
5 dB	0,80	0,80	0,77	**0,88**
7* dB	0,84	0,83	0,80	**0,90**
10 dB	0,88	0,87	0,85	**0,92**
15 dB	0,94	0,93	0,91	**0,94**
Aleatoria* (dB)	0,84	0,83	0,80	**0,89**
Média	0,83	0,83	0,80	0,89

Fonte: elaborado pelo autor.

Em relação à métrica PESQ, além dos algoritmos clássicos e da RNP utilizada nesta dissertação, são mostrados, na Tabela 8, os resultados do trabalho realizado por Xu *et al.* (2014). Salienta-se que a comparação não é direta, uma vez que os treinamentos das RNPs são realizados com tipos de ruídos e níveis de SNR diferentes. Xu *et al.* (2014) consideram os cenários de

treinamento com SNRs -5 dB, 0 dB, 5 dB, 10 dB, 15 dB e 20 dB e os ruídos balbucios, restaurante e rua. Já para os cenários de teste, acrescentam a SNR 7 dB e os ruídos carro e sala de exposições. Contudo, a comparação é útil para servir como referência para avaliação de desempenho dos resultados apresentados. Para essa métrica, a RNP alcança um aumento médio de 29%, em relação ao sinal ruidoso. Já os aumentos na subtração espectral e no filtro de Wiener são de 14% e 9%, respectivamente. O maior ganho, de 40%, é obtido a partir das estimativas dos sinais de voz limpos, utilizando a RNP, para os sinais degradados com SNR 0 dB. A maior PESQ é obtida com SNR 15 dB também para a RNP, com valor de 3,46. Os resultados da RNP obtidos nesta dissertação são superiores aos resultados dos níveis de SNRs equivalentes ao trabalho de Xu *et al.* (2014).

Tabela 8 – Resultados da métrica PESQ, obtidos pelos algoritmos para as diferentes categorias de SNR.

SNR	Sinal Ruidoso	Subtração Espectral	Filtro de Wiener	RNP	RNP 2
0 dB	1,92	2,28	2,09	**2,68**	2,41
5 dB	2,25	2,61	2,48	**2,99**	2,78
7 dB*	2,39	2,73	2,63	**3,10**	2,95
10 dB	2,59	2,92	2,84	**3,25**	3,10
15 dB	2,92	3,22	3,17	**3,46**	3,36
Aleatoria* (dB)	2,46	2,74	2,65	**3,10**	-
Média	2,42	2,75	2,64	3,10	2,92

Fonte: elaborado pelo autor.

Finalmente, os resultados da métrica TEP são expostos na Tabela 9. Na média geral, constata-se uma redução de 3% da TEP para a RNP, um aumento de 9% para a subtração espectral e um aumento de 34% para o filtro de Wiener.

Para as SNRs 0, 5 e 7 dB, a RNP obtém uma redução no erro de aproximadamente 14%, 9% e 5%, respectivamente. Contudo, em 15 dB e na categoria de SNRs aleatórias, observa-se um aumento na taxa de erro de palavra de 6% e 2%, respectivamente. Embora haja redução na LSD e um aumento da PESQ e da STOI, ocorre um incremento na taxa de erro de palavra para SNRs acima de 10 dB. Nesses cenários, o compromisso existente entre a redução de ruído aditivo e a atenuação e o espalhamento de energia das principais harmônicas que constituem os fonemas das palavras fica evidente. Consequentemente, os efeitos provocados pela decodificação imprecisa da RNP impactam na taxa de erro de palavras em um SRAV.

Tabela 9 – Resultados da métrica TEP, obtidos pelos algoritmos para as diferentes categorias de SNR.

SNR	Sinal Ruidoso	Subtração Espectral	Filtro de Wiener	RNP
0 dB	0,66	0,73	0,87	**0,57**
5 dB	0,45	0,50	0,66	**0,41**
7 dB*	0,41	0,44	0,58	**0,39**
10 dB	**0,35**	0,38	0,46	**0,35**
15 dB	**0,31**	0,32	0,36	0,33
Aleatoria* (dB)	**0,41**	0,46	0,56	0,42
Média	0,43	0,47	0,58	0,41

Fonte: elaborado pelo autor.

Os resultados da métrica TEP, obtidos pelos algoritmos para os diferentes tipos de ruídos, são apresentados na Tabela 10. Para os ruídos balbucios e aeroporto, a RNP é capaz de reduzir a taxa de erro, porém, para o ruído trem o resultado é praticamente o mesmo do sinal ruidoso. Já para o ruído aeroporto, é constatado um pequeno aumento.

Os desempenhos dos algoritmos clássicos são especialmente ruins no cenário com o ruído balbucios, enquanto que, para o ruído trem, apresentam o melhor resultado. Tais algoritmos aumentam a taxa de erro de palavra nos quatro tipos de ruídos por causa da distorção espectral, do espalhamento de energia em frequência, e do ruído musical.

Tabela 10 – Resultados da métrica TEP, obtidos pelos algoritmos para os diferentes tipos de ruídos.

SNR	Sinal Ruidoso	Subtração Espectral	Filtro de Wiener	RNP
Aeroporto	**0,41**	0,48	0,58	0,44
Balbucios	0,49	0,58	0,68	**0,42**
Carro	0,49	0,52	0,64	**0,42**
Trem	**0,35**	0,39	0,47	0,36
Média	0,43	0,49	0,59	0,41

Fonte: elaborado pelo autor.

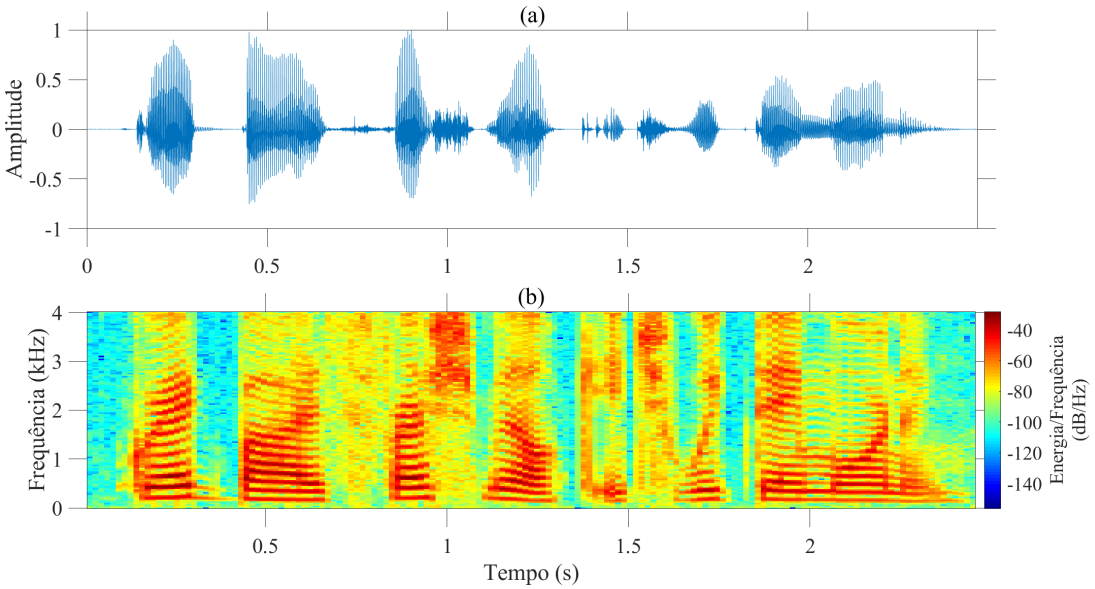
5.2 Análise gráfica

Para analisar a melhoria ou a degradação da qualidade e da inteligibilidade, são ilustrados os sinais de voz limpos estimados a partir de cada um dos modelos, nos cenários de 0 dB (pior caso) e 10 dB (segundo melhor caso), degradados com os ruídos balbucios (segundo melhor caso) e aeroporto (pior caso). Toma-se como referência um sinal de voz limpo do banco de dados de teste, contendo uma elocução feminina da seguinte frase "*Greg buys fresh milk each weekday morning*". A representação tanto no (a) domínio do tempo quanto no (b) domínio da frequência são apresentados na Figura 30.

A escolha por uma elocução feminina se deve ao fato de que as mulheres possuem uma maior quantidade de energia em componentes de alta frequência na voz. Em diversas técnicas de redução de ruído (filtro de Wiener e subtração espectral, por exemplo), tais componentes são atenuadas, o que torna este cenário mais desafiador. Os sinais degradados com ruído balbucios e SNR 0 dB, ruído aeroporto e SNR 0 dB, ruído balbucios e SNR 10 dB, e ruído aeroporto e SNR 10 dB, são apresentados nas Figuras 31, 32, 33 e 34, respectivamente.

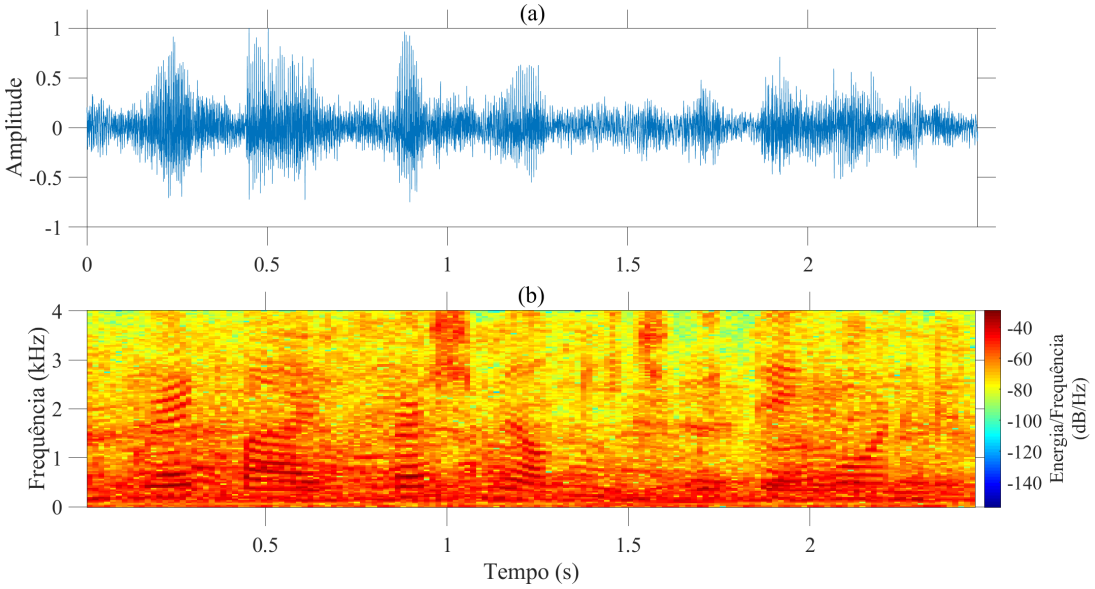
Ao comparar os sinais ruidosos com o sinal limpo, observa-se nos espectrogramas que uma grande quantidade de energia (cor vermelha e amarela) foi espalhada e distribuída (de acordo com o tipo de ruído) nos sinais ruidosos por diversas componentes de frequência ao longo do tempo. Já no domínio do tempo, a variância da amplitude torna-se maior (de acordo com o nível de SNR), tanto nos períodos de voz quanto nos períodos de silêncio.

Figura 30 – Forma de onda de uma elocução feminina, no (a) domínio do tempo, e (b) seu respectivo espectrograma.



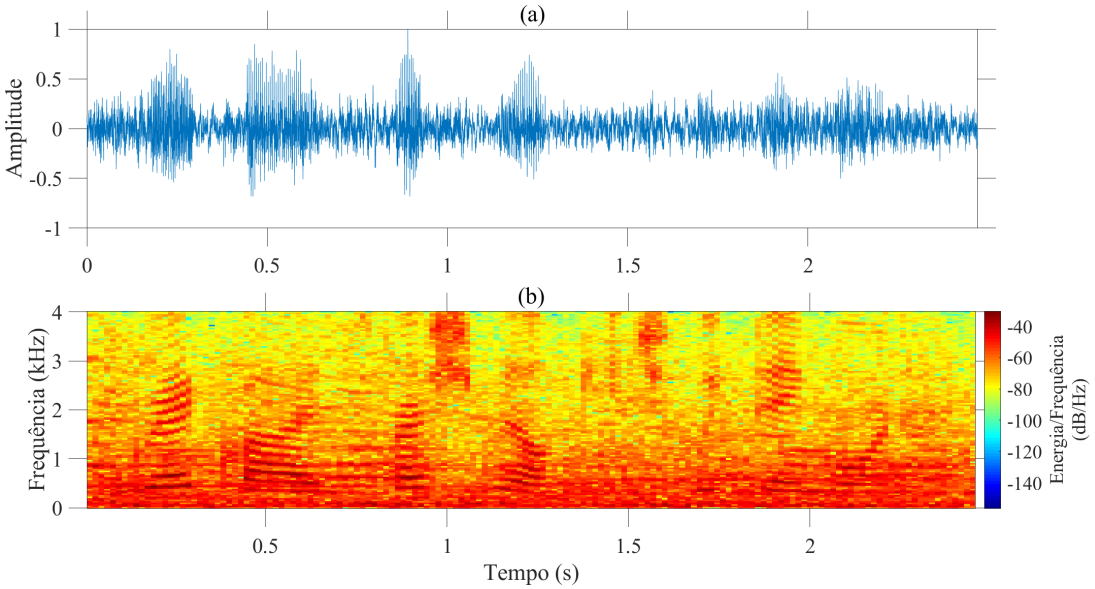
Fonte: elaborado pelo autor.

Figura 31 – (a) Sinal da Figura 30 degradado com ruído balbucios e SNR 0 dB no domínio do tempo, e (b) o respectivo espectrograma do sinal degradado.



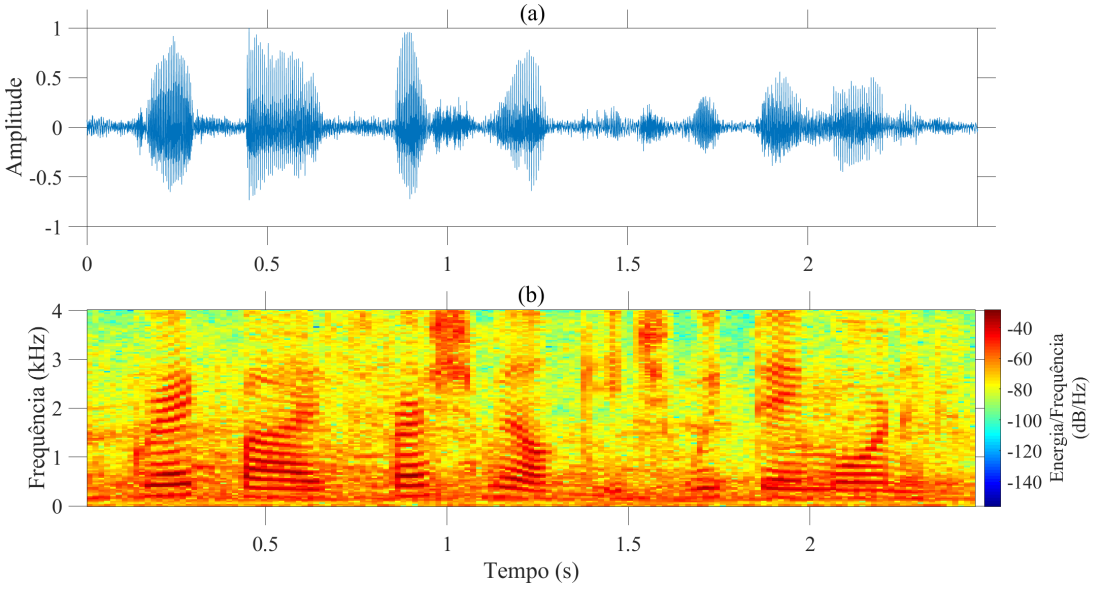
Fonte: elaborado pelo autor.

Figura 32 – (a) Sinal da Figura 30 degradado com ruído aeroporto e SNR 0 dB no domínio do tempo, e (b) o respectivo espectrograma do sinal degradado.



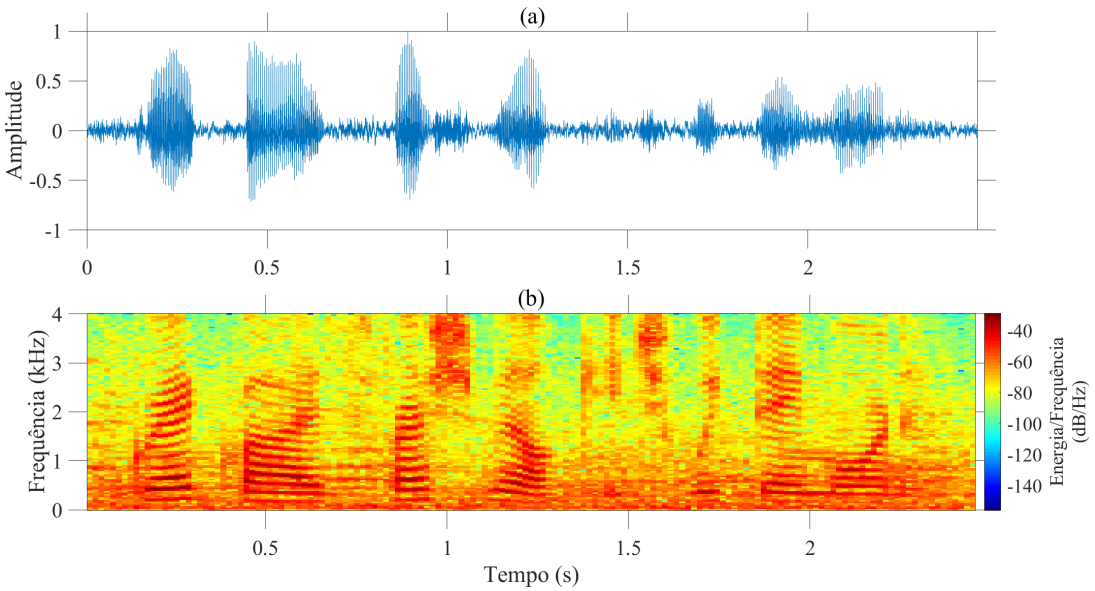
Fonte: elaborado pelo autor.

Figura 33 – (a) Sinal da Figura 30 degradado com ruído balbucios e SNR 10 dB no domínio do tempo, e (b) o respectivo espectrograma do sinal degradado.



Fonte: elaborado pelo autor.

Figura 34 – (a) Sinal da Figura 30 degradado com ruído aeroporto e SNR 10 dB no domínio do tempo, e (b) o respectivo espectrograma do sinal degradado.

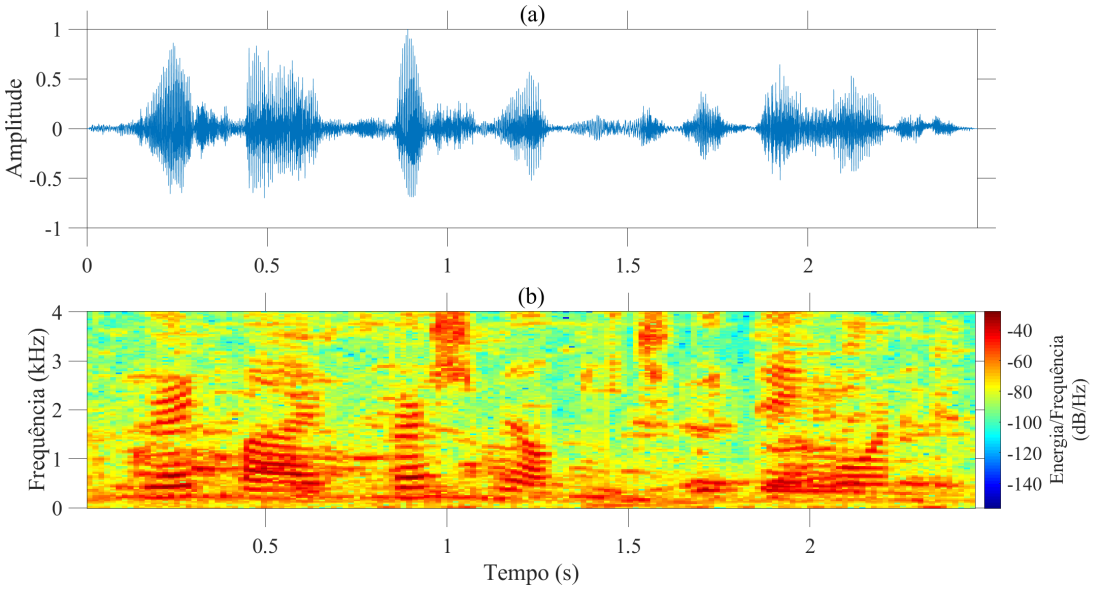


Fonte: elaborado pelo autor.

5.2.1 Resultados obtidos pela subtração espectral

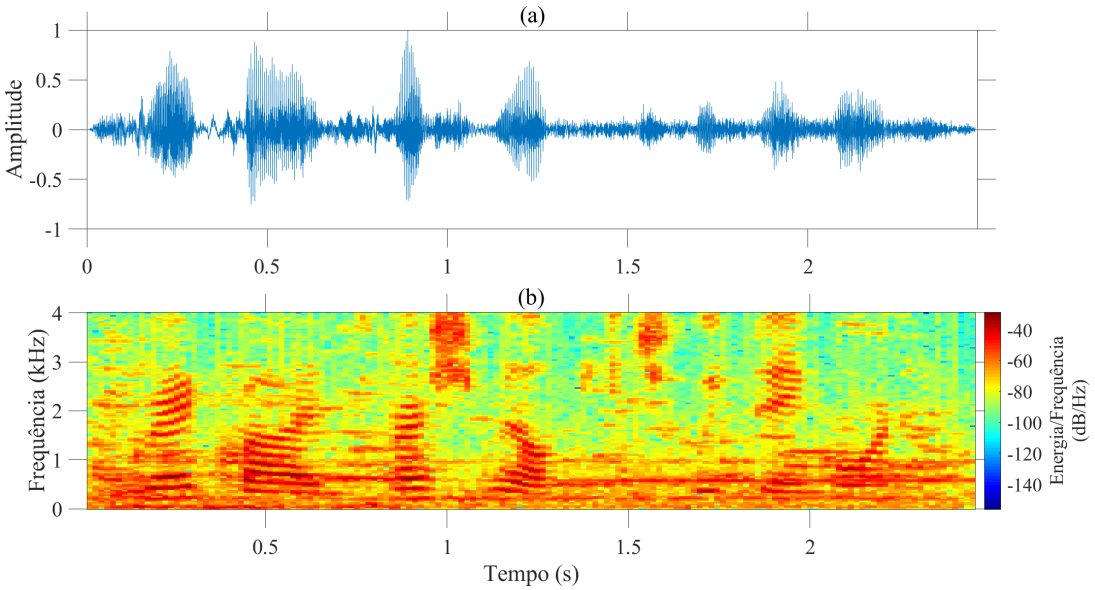
Esta Seção apresenta as estimativas dos sinais limpos utilizando o algoritmo da subtração espectral. Comparando as Figuras 35 e 36 com a Figuras 31 e 32, referente ao sinais ruidosos, pode-se observar a redução do ruído aditivo, principalmente nos períodos de silêncio. Contudo, esta redução tem um custo do aumento da distorção espectral nos quadros que possuem sinais de voz. Ao comparar as Figuras 35, 36 e 30, percebe-se que a energia dos sinais estimados é espalhada por diversas componentes de frequência. Tais componentes são importantes para formação dos fonemas e, uma vez que a energia dessas componentes se espalha, os fonemas se tornam distorcidos e menos inteligíveis. Com os sinais das Figuras 37 e 38, esse efeito é menor.

Figura 35 – Subtração espectral: (a) estimação do sinal limpo da Figura 30 no domínio do tempo, e (b) o respectivo espectrograma do sinal estimado, a partir do sinal da Figura 31, degradado com ruído balbucios e SNR de 0 dB.



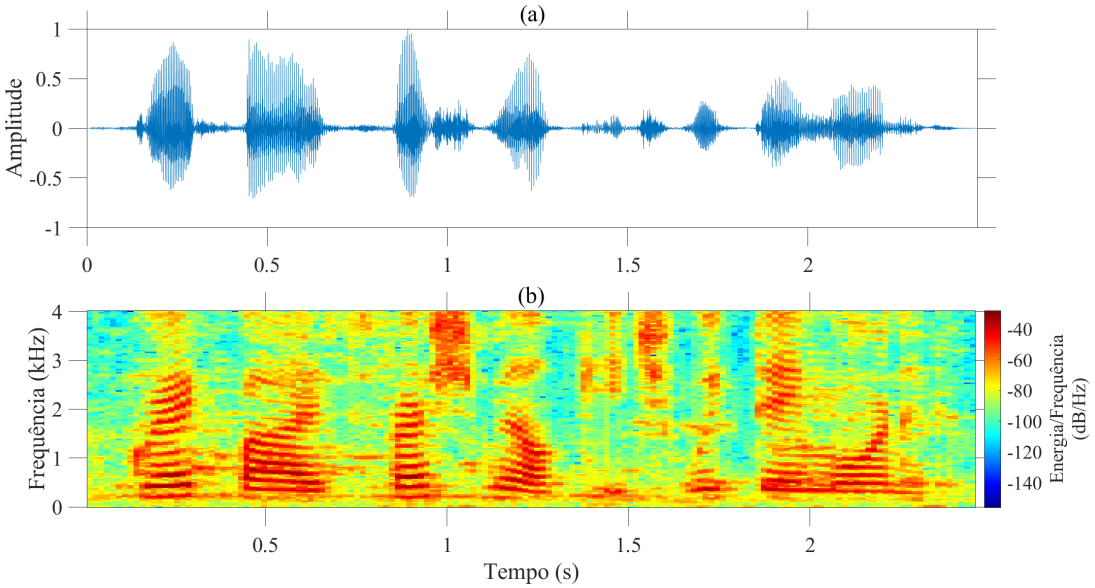
Fonte: elaborado pelo autor.

Figura 36 – Subtração espectral: (a) estimação do sinal limpo da Figura 30 no domínio do tempo, e (b) o respectivo espectrograma do sinal estimado, a partir do sinal da Figura 32, degradado com ruído aeroporto e SNR de 0 dB.



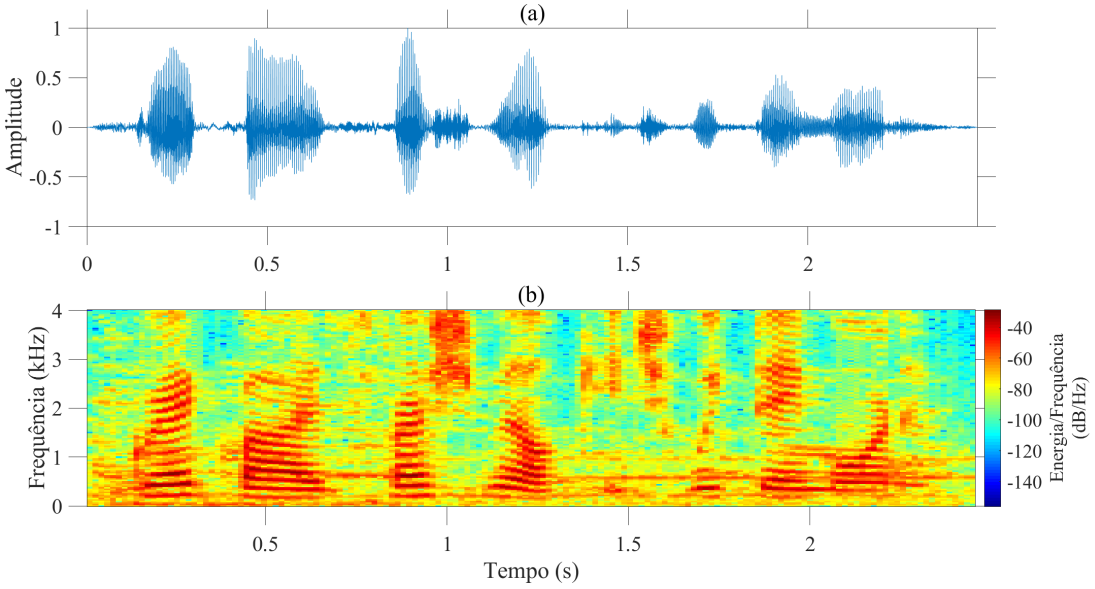
Fonte: elaborado pelo autor.

Figura 37 – Subtração espectral: (a) estimação do sinal limpo da Figura 30 no domínio do tempo, e (b) o respectivo espectrograma do sinal estimado, a partir do sinal da Figura 33, degradado com ruído balbucios e SNR de 10 dB.



Fonte: elaborado pelo autor.

Figura 38 – Subtração espectral: (a) estimação do sinal limpo da Figura 30 no domínio do tempo, e (b) o respectivo espectrograma do sinal estimado, a partir do sinal da Figura 34, degradado com ruído aeroporto e SNR de 10 dB.



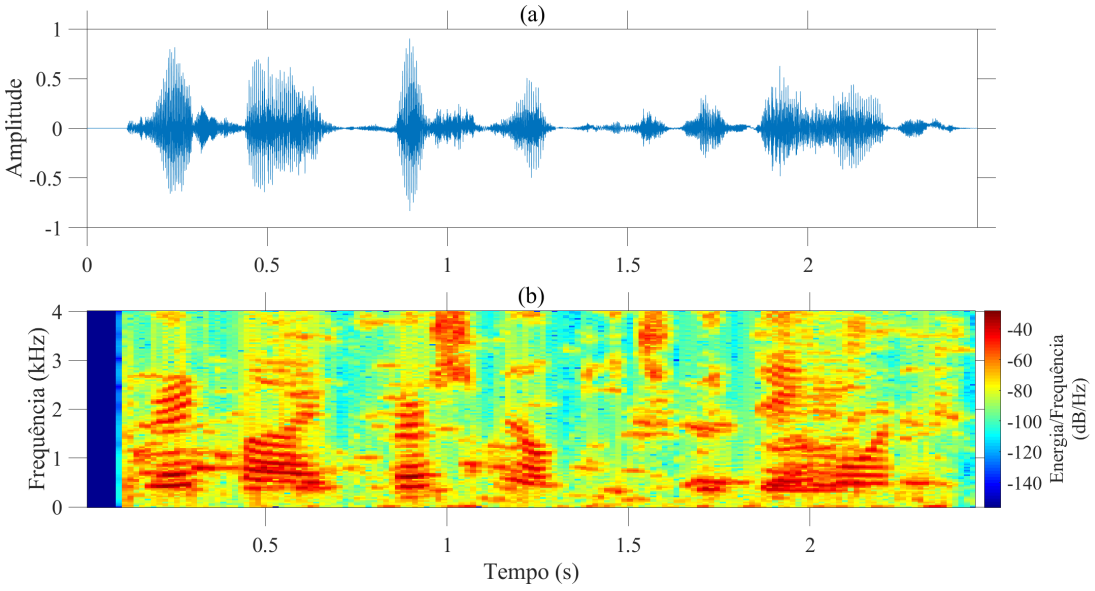
Fonte: elaborado pelo autor.

5.2.2 Resultados obtidos pelo filtro de Wiener

As estimativas dos sinais limpos utilizando o algoritmo do filtro de Wiener são apresentadas nesta Seção. Os efeitos da distorção em frequência, do espalhamento e da atenuação de energia das principais componentes harmônicas dos sinais de voz se tornam ainda mais perceptíveis quando analisamos os sinais de voz limpo estimados utilizando o filtro de Wiener, ilustrados nas Figuras 39, 40, 41 e 42 (mesmos cenários apresentados da seção anterior).

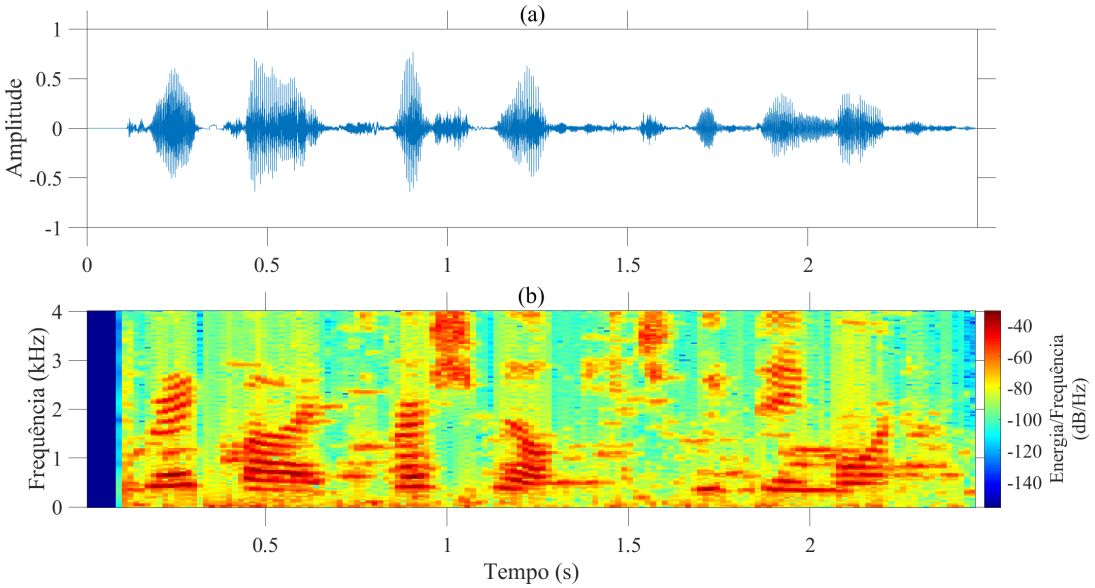
A versão do filtro de Wiener abordado nesta dissertação utiliza a estimativa da SNR a partir dos 10 primeiros quadros do sinal de tal forma que o período relativo a esses quadros é considerado silêncio. A tarja azul no início dos sinais, percebida nos espectrogramas das Figuras citadas, é relativo a esse período de silêncio em que o nível de energia é reduzido a zero. Ao analisar essas Figuras, percebe-se uma redução no nível de ruído nos períodos de silêncio. Contudo, ao analisar as componentes harmônicas formadoras dos fonemas, observa-se a atenuação e o espalhamento de energia para outras frequências. Por conta disso, ocorre a redução no nível de inteligibilidade do sinal ao reduzir o ruído. Isso confirma o compromisso existente entre a redução do ruído aditivo e o aumento da distorção espectral em algoritmos clássicos, como descrito por Loizou (2013), e se reflete nos resultados da LSD, STOI, PESQ e TEP.

Figura 39 – Filtro de Wiener: (a) estimação do sinal limpo da Figura 30 no domínio do tempo, e (b) o respectivo espectrograma do sinal estimado, a partir do sinal da Figura 31, degradado com ruído balbucios e SNR de 0 dB.



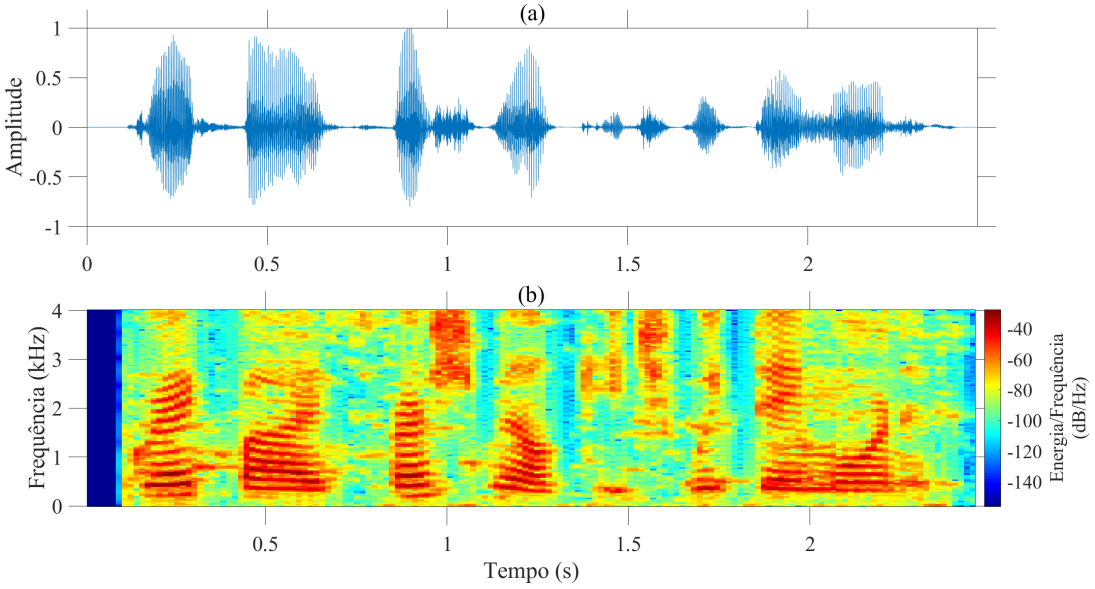
Fonte: elaborado pelo autor.

Figura 40 – Filtro de Wiener: (a) estimação do sinal limpo da Figura 30 no domínio do tempo, e (b) o respectivo espectrograma do sinal estimado, a partir do sinal da Figura 32, degradado com ruído aeroporto e SNR de 0 dB.



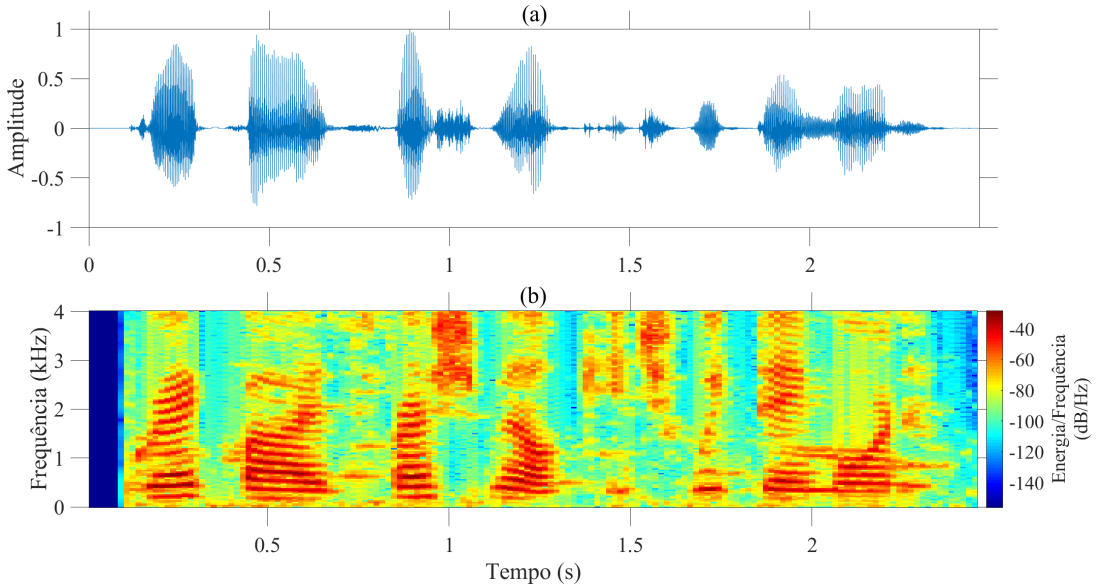
Fonte: elaborado pelo autor.

Figura 41 – Filtro de Wiener: (a) estimação do sinal limpo da Figura 30 no domínio do tempo, e (b) o respectivo espectrograma do sinal estimado, a partir do sinal da Figura 33, degradado com ruído balbucios e SNR de 10 dB.



Fonte: elaborado pelo autor.

Figura 42 – Filtro de Wiener: (a) estimação do sinal limpo da Figura 30 no domínio do tempo, e (b) o respectivo espectrograma do sinal estimado, a partir do sinal da Figura 34, degradado com ruído aeroporto e SNR de 10 dB.



Fonte: elaborado pelo autor.

5.2.3 Resultados obtidos pela rede neural profunda

Esta Seção apresenta os sinais de voz limpo estimados obtidos pela RNP. Comparando com os algoritmos anteriores, há uma semelhança maior entre os sinais de voz limpo e os estimados.

Na estimativa do sinal limpo do sinal degradado com SNR 0 dB e ruído balbucios, ilustrado na Figura 43, boa parte das principais componentes de frequência formadoras dos fonemas são restauradas. Isso pode ser observado no espectrograma nos períodos próximos a 0,25 s a 0,5 s. Contudo, no intervalo entre 1,8 s a 2,5 s, a reconstrução não é totalmente precisa, pois ocorre um pequeno espalhamento de energia (ex. em torno de 100 Hz a 1,2 kHz) e atenuação em componentes de baixa energia (ex. em torno de 1,5 Hz a 4 kHz).

Já para a estimativa do sinal limpo a partir do sinal ruidoso degradado com ruído aeroporto e SNR 0 dB, ilustrado na Figura 44, observa-se diversos períodos ruidosos, mesmo nos quadros de silêncio. Tal ruído não foi utilizado no treinamento da RNP e, por conta disso, o aumento de desempenho não é tão significativo quanto nos demais tipos de ruído. Neste caso, embora haja redução de ruído, o espalhamento de energia é maior, se comparado ao cenário anterior.

Na simulação de pior caso (ruído aeroporto e SNR 0 dB), embora haja uma redução de 23% na LSD (1,71 para 1,32), um aumento de 9% da STOI (0,70 para 0,76), um aumento de 17% na PESQ (1,94 para 2,27), ocorre um aumento de 5% na TEP (0,63 para 0,66). Este aumento é consequência da presença de ruído musical e do espalhamento de energia em algumas componentes de frequência formadoras dos fonemas nos sinais de voz estimados. O ruído musical fica evidente nos momentos de silêncio, os quais não foi possível reduzir de forma significativa os níveis de ruído aditivo de fundo para este cenário. Nesse caso, picos de energia espectrais isolados (não conectados), especialmente em altas frequências, soam como tons musicais ao ouvido do ser humano. Já nos momentos em que há voz, o ruído musical produz uma versão distorcida do sinal.

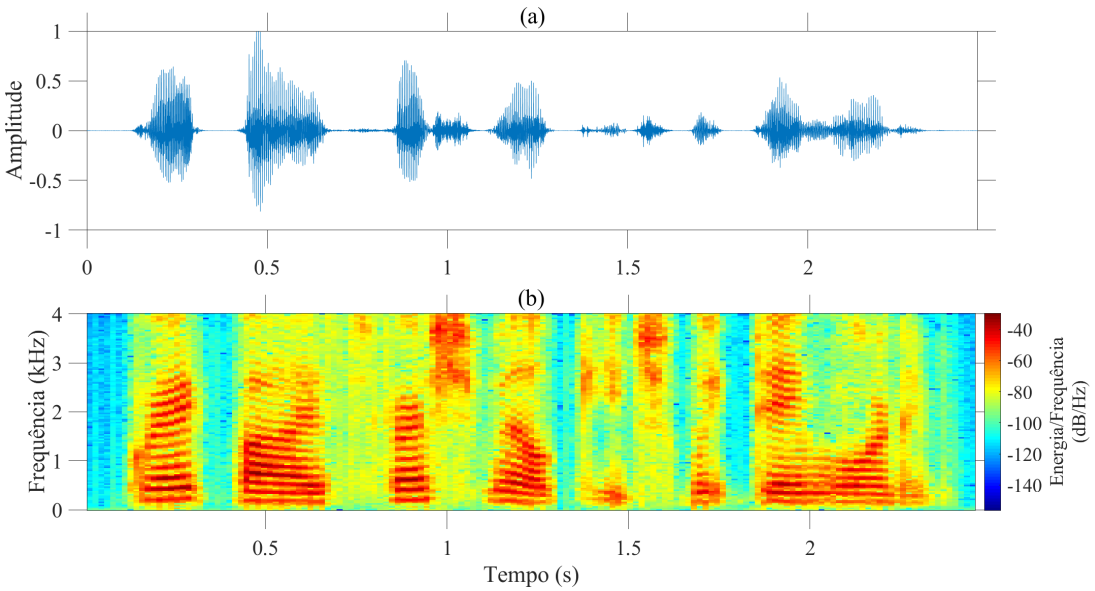
Os sinais estimados para o cenário de 10 dB, são apresentados nas Figuras 45 e 46. É notória a redução de ruído, principalmente nos períodos de silêncio, na Figura 45. Em geral, as principais componentes harmônicas formadoras dos fonemas também são reconstruídas de forma adequada. Já na Figura 46, também é possível notar ruído musical do processo de filtragem da RNP quando comparado ao sinal limpo, porém, este resultado é um indicativo da capacidade de generalização da RNP. Devido à quantidade de cenários diferentes, opta-se por apresentar no Apêndice A alguns dos demais resultados obtidos pela RNP, especificamente para os cenários com ruído trem (0 e 15 dB), carro (0 e 15 dB), balbucios e aeroporto (7 dB).

A análise comparativa do desempenho da RNP com e sem pré-treinamento é realizada a partir da observação dos erros médio quadrático do conjunto de dados de treinamento (EQM_t) e do conjunto de dados de validação (EQM_v) ao longo de 20 épocas nas Figuras 47 e 48. Para a RNP sem pré-treinamento, o menor valor para o EQM_v é obtido na época 15, com cerca de 0,4069. Já para a RNP com pré-treinamento, o menor valor obtido para o EQM_v é obtido na época 11, com cerca de 0,3998. Comparando ambos os gráficos, é possível observar que o EQM_t e EQM_v da RNP com pré-treinamento na primeira época é menor do que na RNP sem o pré-treinamento. Contudo, o pré-treinamento não impacta de forma significa nos valores mínimos do EQM_t e do EQM_v para a configuração de parâmetros e hiperparâmetros utilizada.

Para evitar o super-ajuste aos dados de treinamento e obter a maior capacidade de generalização possível para a RNP, salvam-se os parâmetros ($\boldsymbol{W}, \boldsymbol{b}$) da época com menor EQM_v, os quais são usados para a etapa de decodificação. Tal técnica é conhecida como parada preventiva[1] e evita o super-ajuste[2] da rede neural profunda aos dados de treinamento.

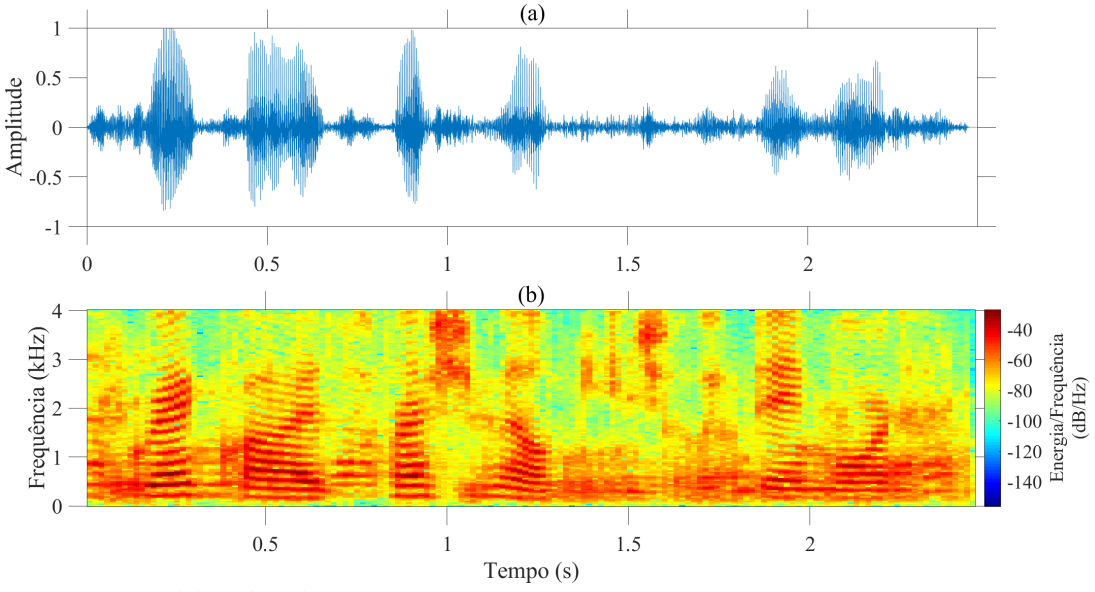
[1] Expressão traduzida do inglês *early stopping*.
[2] Expressão traduzida do inglês *overfitting*.

Figura 43 – RNP: (a) estimação do sinal limpo da Figura 30 no domínio do tempo, e (b) o respectivo espectrograma do sinal estimado, a partir do sinal da Figura 31, degradado com ruído balbucios e SNR de 0 dB.



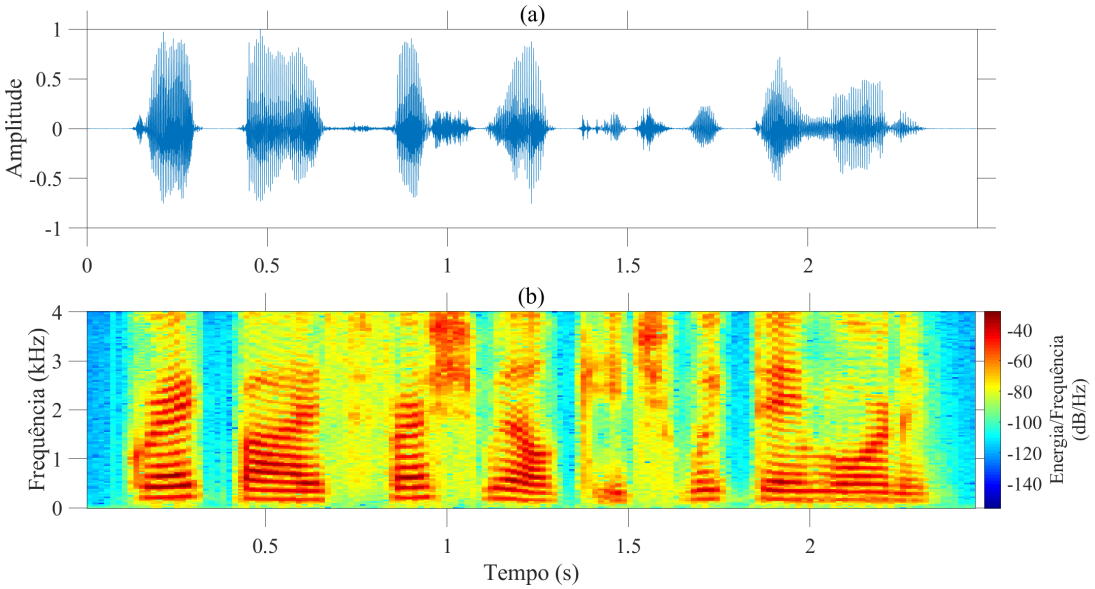
Fonte: elaborado pelo autor.

Figura 44 – RNP: (a) estimação do sinal limpo da Figura 30 no domínio do tempo, e (b) o respectivo espectrograma do sinal estimado, a partir do sinal da Figura 32, degradado com ruído aeroporto e SNR de 0 dB.



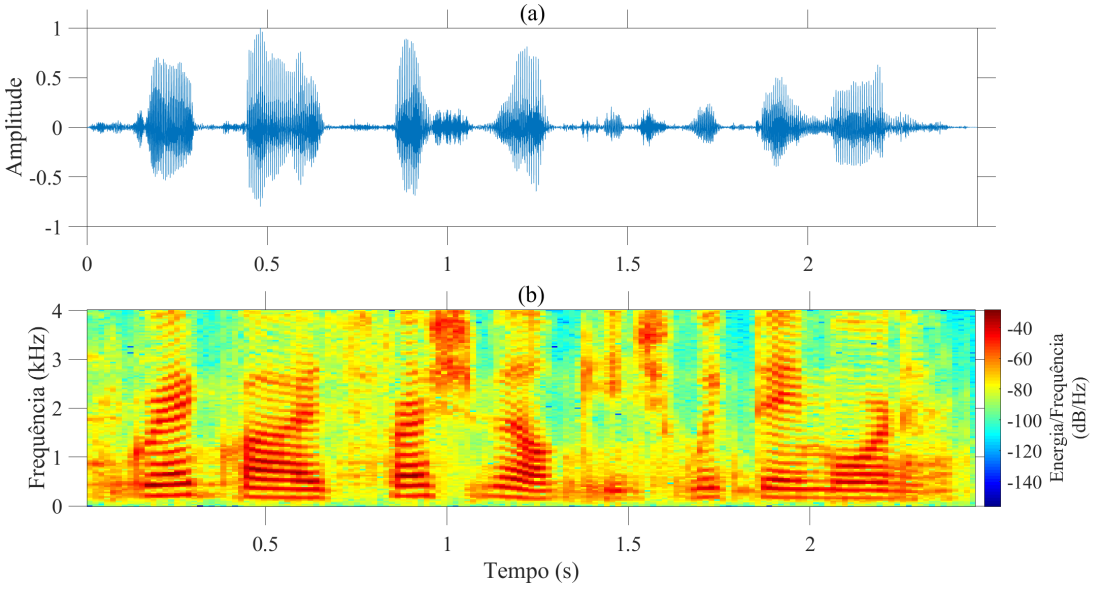
Fonte: elaborado pelo autor.

Figura 45 – RNP: (a) estimação do sinal limpo da Figura 30 no domínio do tempo, e (b) o respectivo espectrograma do sinal estimado, a partir do sinal da Figura 33, degradado com ruído balbucios e SNR de 10 dB.



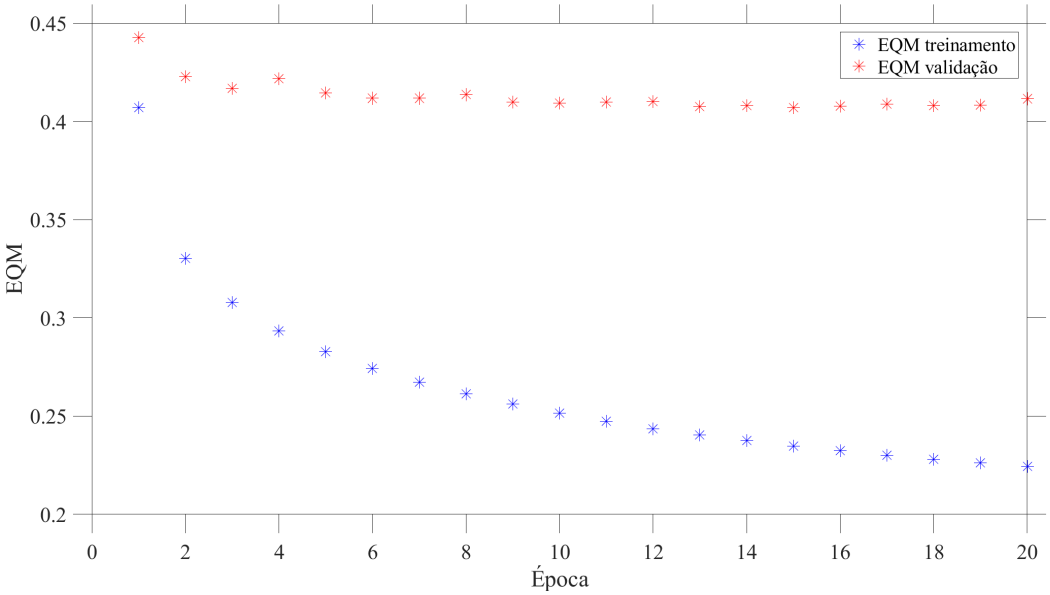
Fonte: elaborado pelo autor.

Figura 46 – RNP: (a) estimação do sinal limpo da Figura 30 no domínio do tempo, e (b) o respectivo espectrograma do sinal estimado, a partir do sinal da Figura 34, degradado com ruído aeroporto e SNR de 10 dB.



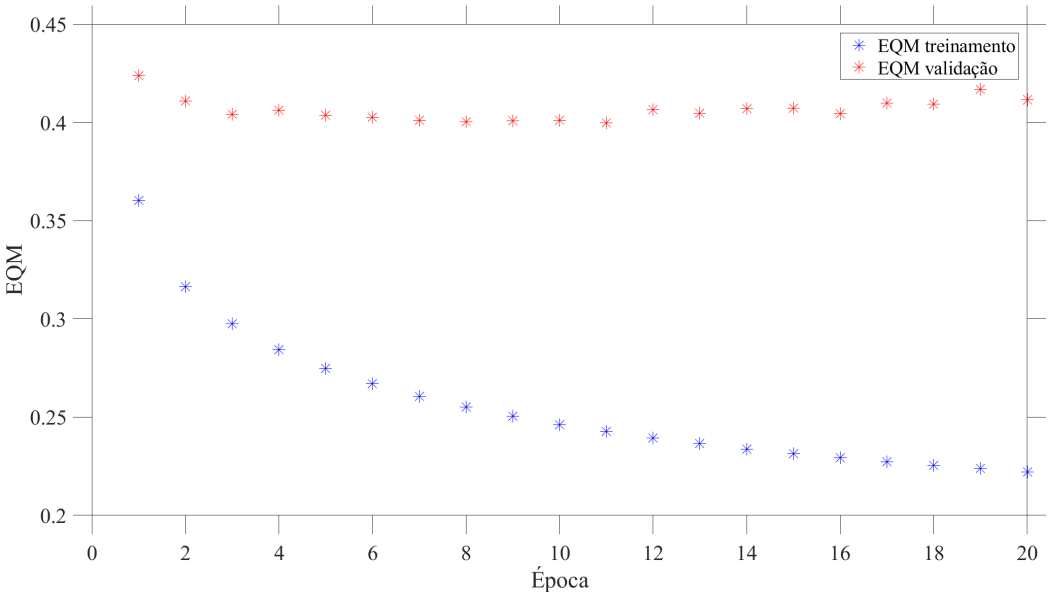
Fonte: elaborado pelo autor.

Figura 47 – RNP sem o pré-treinamento: evolução do erro médio quadrático relativo ao conjunto de dados de treinamento (em azul) (EQM_t) e ao conjunto de dados de validação (em vermelho) (EQM_v) do estágio de treinamento da RNP.



Fonte: elaborado pelo autor.

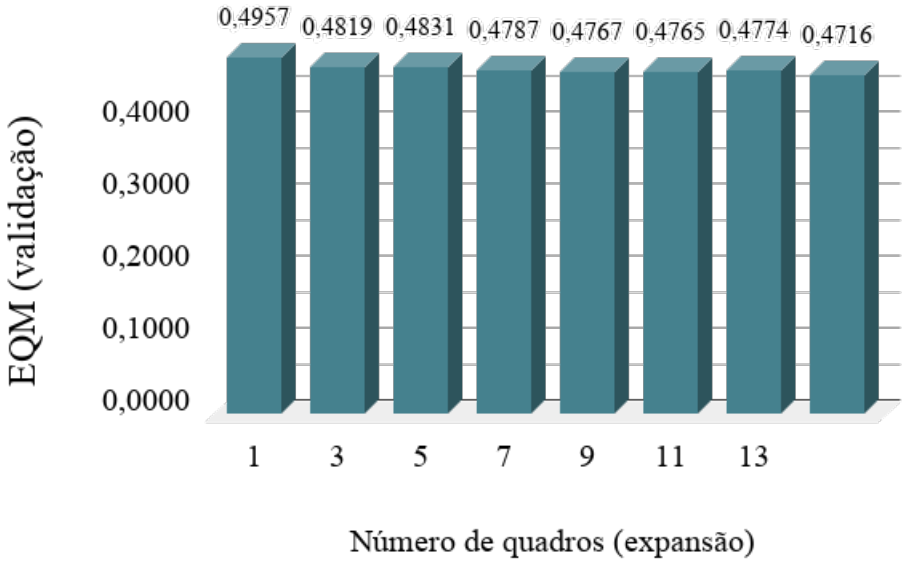
Figura 48 – RNP com o pré-treinamento: evolução do erro médio quadrático relativo ao conjunto de dados de treinamento (em azul) (EQM_t) e ao conjunto de dados de validação (em vermelho) (EQM_v) do estágio de treinamento da RNP.



Fonte: elaborado pelo autor.

O resultado de 7 testes utilizando valores diferentes para expansão de quadros é ilustrado na Figura 49. Quando são usados 11 quadros para a expansão, obtém-se o menor valor para o EQM_v. Para a obtenção desta análise comparativa, utiliza-se somente 10% do banco de dados de treinamento devido o custo computacional elevado para processar o conjunto completo de dados. Por conta disso, os valores do EQM_v são elevados se comparados aos valores obtidos para o EQM_v obtido com o banco de dados completo.

Figura 49 – Resultados dos testes com expansão de quadros, avaliado com 10% dos arquivos do banco de dados de treinamento.



Fonte: elaborado pelo autor.

5.3 Discussão

Os resultados demonstram a capacidade da RNP de realizar o mapeamento espectral não linear para reduzir ruído aditivo e, consequentemente, aumentar a qualidade e a inteligibilidade dos sinais de voz degradados. A metodologia baseada na RNP supera os modelos clássicos de subtração espectral e filtro de Wiener em todos os cenários de ruídos e níveis de SNR propostos. Em relação à métrica LSD, o resultado médio da RNP é 36% menor comparado ao filtro de Wiener e 25% menor com relação à subtração espectral. Na métrica STOI, o resultado médio da RNP é 12% superior ao obtido pelo filtro de Wiener e 8% superior ao resultado obtido pela subtração espectral. Na métrica PESQ, a RNP é superior em 17% comparado ao filtro de Wiener e 13% em relação à subtração espectral. Já na métrica TEP, a RNP apresenta um resultado 29% menor que o filtro de Wiener e 13% menor que a subtração espectral.

O algoritmo da subtração espectral é um algoritmo computacionalmente simples e realiza a redução de ruído aditivo, porém, nota-se a presença de ruído musical nos sinais de voz limpo estimados pelo algoritmo. A presença de ruído musical é mais evidente nos sinais estimados a partir dos degradados com níveis baixos de SNR ($\leq$ 10 dB).

Embora o filtro de Wiener seja computacionalmente mais custoso do que a subtração espectral, a versão utilizada nesta dissertação não produziu bons resultados para os cenários propostos quando comparado aos demais algoritmos. Dos resultados, fica evidente o compromisso existente entre a redução de ruído aditivo e o aumento da distorção espectral e do ruído musical, presentes nas estimativas dos sinais limpos.

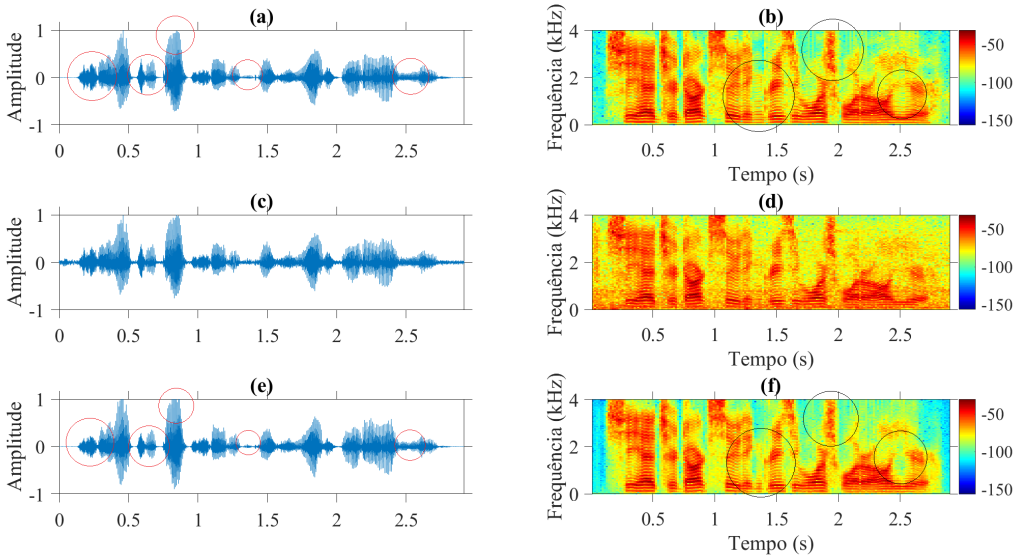
Embora a RNP tenha apresentado significativo aumento de desempenho nas métricas objetivas LSD, STOI e PESQ, constata-se que o aumento geral na TEP não foi significativo. Como demonstrado na Tabela 9, a RNP aponta para um aumento da TEP para a SNR 15 dB. Através de alguns testes, pode-se notar que a TEP do sinal com SNR 15 dB se aproxima do resultado da TEP obtido para o sinal limpo.

Na prática, a RNP está atenuando e espalhando energia em algumas componentes de frequência que impactam, ainda que de forma mínima, no aumento da taxa de erro de palavra. As Figuras 50a., 50c. 50e. apresentam o comparativo entre o sinal de voz limpo de um homem, o sinal degradado com ruído carro e SNR 15 dB e o sinal de voz limpo estimado pela RNP, respectivamente. As Figuras 50b., 50d. 50e. são os respectivos espectrogramas desses sinais. Os círculos em vermelho no domínio do tempo e em preto nos espectrogramas são pequenas diferenças percebidas entre as Figuras do sinal limpo e do estimado. Tais diferenças estão

relacionadas aos efeitos provocados pela RNP comentados anteriormente.

Uma possível solução para esse problema consiste em treinar a RNP com sinais com SNRs > 15 dB e/ou com os próprios sinais de voz limpos, para que ela possa aprender as especificidades do problema nestes cenários com pouco ruído aditivo.

Figura 50 – (a) Sinal de voz limpo de um homem no domínio do tempo, (c) sinal degradado com SNR 15 dB e ruído carro, (e) sinal de voz limpo estimado pela RNP. (b), (d) e (f) são os espectrogramas desses sinais.



Fonte: elaborado pelo autor.

A disponibilidade de um banco de dados heterogêneo, com uma grande quantidade de arquivos de áudio de voz (acima de 50 horas) é fundamental para o bom desempenho da rede neural. Há a necessidade de aumentar a capacidade de generalização da RNP para diversos cenários de ruído e SNR. Conhecer e modelar as condições do mundo físico carece de uma grande quantidade de dados e que englobe os mais diversos ambientes de utilização de SRAVs. A presença de algum tipo de ruído desconhecido, que a RNP não é capaz de generalizar, pode conduzir a baixas taxas de acerto de reconhecimento.

A partir dos resultados apresentados, para os algoritmos da subtração espectral, do filtro de Wiener e da rede neural profunda, algumas conclusões podem ser obtidas.

6 CONCLUSÕES E TRABALHOS FUTUROS

Esta dissertação descreve uma proposta para redução de ruído aditivo em sinais de voz em cenários simulados do mundo físico, utilizando uma estrutura de rede neural profunda (RNP). Para o conjunto de dados de treinamento, são utilizados os ruídos balbucios, carro e trem, e os níveis de SNR 0, 5, 10 e 15 dB. Já para o conjunto de de validação e teste, além dos cenários estabelecidos no conjunto de treinamento, são acrescentados o ruído aeroporto e o nível de SNR 7 dB. Ademais, são utilizados níveis de SNRs geradas aleatoriamente com distribuição de probabilidade uniforme, no intervalo de 0 a 15 dB, tanto para o conjunto de treinamento quanto para o conjunto de teste. Cada combinação de tipo de ruído e nível de SNR constitui um cenário, totalizando 15 para o conjunto de treinamento e 24 para o conjunto de teste.

A RNP proposta nesta dissertação se baseia numa arquitetura de múltiplas camadas de perceptrons convencional, com as unidades de camadas adjacentes totalmente conectadas. O treinamento completo é realizado utilizando os estágios de pré-treinamento e treinamento (ajuste fino), baseados nos algoritmos máquina restrita de Boltzmann e re-propagação, respectivamente. Os algoritmos clássicos subtração espectral não linear e o filtro de Wiener, com estimação de SNR a priori, são utilizados como referências para análise comparativa. A avaliação de desempenho é realizada pelas métricas objetivas LSD, STOI, PESQ e a taxa de erro de palavra (TEP).

Acerca dos resultados, a RNP supera os algoritmos clássicos em todos os cenários de SNR, inclusive dos sinais degradados com ruído aeroporto e SNR de 7 dB que não são utilizados no treinamento. Contudo, os resultados para a TEP não são satisfatórios para níveis de SNR acima de 10 dB, embora seja superior aos algoritmos clássicos. Apesar das melhorias, as métricas LSD, STOI e PESQ não são capazes de avaliar de forma precisa as atenuações e o espalhamento de energia nas principais componentes harmônicas que formam os fonemas. Tais efeitos impactam no resultado da TEP. Para a solução desse problema, propõe-se a utilização de sinais com SNRs mais elevadas (ex. 20 dB) e/ou a utilização de sinais de voz limpos no treinamento da RNP. Devido ao elevado custo computacional para o treinamento e validação da RNP, tais sinais não foram utilizados nesta dissertação. Para processar 60 horas de áudio do conjunto de treinamento e 17 horas de áudio para o conjunto de validação, são necessárias cerca de 72 horas utilizando os recursos computacionais disponíveis para a simulação. Esse tempo já é obtido com a otimização do código para a leitura dos arquivos de áudio de forma rápida e com a utilização da unidade de processamento gráfica em operações que demandam cálculos

matemáticos expressivos, com grandes matrizes.

Nesta dissertação, emprega-se uma metodologia para extração de características de sinais de voz para o treinamento da RNP. Além disso, é disponibilizado como informação anexa um conjunto de valores de parâmetros e hiperparâmetros para o treinamento da RNP aplicada à redução de ruído aditivo em sinais de voz. Uma outra contribuição desta dissertação é a utilização de um sistema de reconhecimento automático de voz (SRAV) para avaliar a TEP. Emprega-se uma interface de programação de aplicações (API) que realiza a conexão com o SRAV em nuvem da empresa Google. Tal interface permite o acesso às transcrições do reconhecimento desse sistema. A partir daí, as sentenças fornecidas por esse SRAV e as sentenças do banco de dados são comparadas para avaliar a TEP.

Trabalhos futuros

Dos resultados obtidos, faz-se necessário investigar a correlação existente entre as métricas LSD, STOI e PESQ com a TEP. É possível, por exemplo, aplicar métodos de regressão (ex. método dos mínimos quadrados) para estimar a influência de cada uma dessas métricas na TEP. Tal investigação não foi realizada nesta dissertação e fica como sugestão de trabalho futuro.

Sugere-se também a elaboração de um grande banco de dados (acima de 2.000 elocuções) em português, com sinais de voz limpos, contendo elocuções de homens e mulheres, de diferentes regiões dialéticas do Brasil, seguindo os padrões internacionais. Tal banco de dados pode colaborar com o desenvolvimento técnico-científico nacional na área de processamento de sinais de voz.

Para a continuidade desta dissertação, sugere-se ainda a investigação dos seguintes pontos:

(i) utilização de SNRs flutuantes nos arquivos de áudio, simulando cenários do mundo físico;
(ii) utilização de uma quantidade maior de diferentes tipos de ruído (estacionários, não estacionários, de banda larga, de banda estreita), para aumentar a capacidade de generalização da RNP;
(iii) treinamento da RNP com áudios com SNRs geradas aleatoriamente com distribuição uniforme no intervalo de -10 a 20 dB;
(iv) cenários mais complexos contendo combinações de dois ou mais tipos de ruídos;
(v) outras estruturas para a rede neural profunda, como a rede neural convolucional, a rede neural recorrente, ou combinações delas;

(vi) utilização de técnicas de regularização e otimização;

(vii) o uso de taxas de aprendizagem e tamanho do mini-grupo dinâmicas para cada uma das camadas;

(viii) a utilização de outros tipos de características dos sinais de voz;

(ix) a eficiência de outros algoritmos de pré-treinamento;

(x) correlação entre as métricas de desempenho LSD, STOI e PESQ e a taxa de erro de palavra (TEP), estabelecendo pesos que indiquem quantitativamente as respectivas influências;

(xi) treinamento e validação cruzada com bancos de dados distintos;

(xii) o treinamento da RNP utilizando quadros de curta (20 a 40 ms) e longa duração (80 a 300 ms);

(xiii) impacto de um detector de atividade de voz;

(xiv) redução de ruído juntamente com a redução da reverberação;

(xv) utilização de dois ou mais canais de áudio, possivelmente considerando também um arranjo de sensores.

(xvi) efeito da aplicação de um algoritmo de estimação de fase do sinal limpo.

REFERÊNCIAS

ABHANG, P. A.; GAWALI, B. W.; MEHROTRA, S. C. **Introduction to EEG-and speech-based emotion recognition**. 3. ed. London, UK: Academic Press, 2016. ISBN 978-0-12-804490-2.

ABREU, C. C. E. de. **Melhoramento de sinais de voz baseado na identificação de padrões ruidosos**. Tese (Doutorado em Engenharia Elétrica) – Faculdade de Engenharia, Programa de Pós-Graduação em Engenharia Elétrica: Automação, Universidade Estadual Paulista, Ilha Solteira, 2017.

ALI, A.; RENALS, S. Word error rate estimation for speech recognition: e-WER. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (VOLUME 2: SHORT PAPERS), 56. **Proceedings [...]**. Melbourne, Victoria, Australia: Association for Computational Linguistics, 2018. p. 20–24.

ALLEN, J. Short term spectral analysis, synthesis, and modification by discrete fourier transform. **IEEE Transactions on Acoustics, Speech, and Signal Processing**, IEEE, v. 25, n. 3, p. 235–238, 1977.

BEERENDS, J. G.; STEMERDINK, J. A. A perceptual speech-quality measure based on a psychoacoustic sound representation. **Journal of the Audio Engineering Society**, Audio Engineering Society, v. 42, n. 3, p. 115–123, 1994.

BENESTY, J. **Fundamentals of Speech Enhancement**. Cham, Switzerland: Springer, 2018. ISBN 978-3-319-74524-4.

BENESTY, J.; SONDHI, M. M.; HUANG, Y. **Springer handbook of speech processing**. Montreal, Canada: Springer, 2007. ISBN 978-3-540-49125-5.

BOONE, D. R.; MCFARLANE, S. C.; BERG, S. L. V. **The voice and voice therapy**. 9. ed. Upper Saddle River, NJ, USA: Pearson Education, 2014. ISBN 978-0-13-300702-2.

CAPES. **Catálogo de teses e dissertações**. 2016. https://catalogodeteses.capes.gov.br/catalogo-teses/. Acesso em: 12 de fevereiro de 2020.

CENTER, P. R. **Nearly half of Americans use digital voice assistants, mostly on their smartphones**. 2017. Disponível em: https://www.pewresearch.org/fact-tank/2017/12/12/nearly-half-of-americans-use-digital-voice-assistants-mostly-on-their-smartphones/. Acesso em: 7 de janeiro de 2020.

CHEN, J.; BENESTY, J.; HUANG, Y.; DOCLO, S. New insights into the noise reduction wiener filter. **IEEE Transactions on audio, speech, and language processing**, IEEE, v. 14, n. 4, p. 1218–1234, 2006.

D'AVILA, H. **Som Fricativo Sonoro //: Modificações Vocais**. Monografia (Especialização em Fonoaudiologia) – Departamento de Otorrino-Fonoaudiologia, Universidade Federal de Santa Maria, Santa Maria, 2005.

DAVIS, K. H.; BIDDULPH, R.; BALASHEK, S. Automatic recognition of spoken digits. **The Journal of the Acoustical Society of America**, ASA, v. 24, n. 6, p. 637–642, 1952.

DU, J.; HUO, Q. A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions. In: ANNUAL CONFERENCE OF THE INTERNATIONAL SPEECH COMMUNICATIONS ASSOCIATION (INTERSPEECH 2008), 9. **Proceedings [...]**. Brisbane, Australia, 2008. p. 569–572.

ELSEVIER. **Science Direct**. 1997. Disponível em: https://www.sciencedirect.com/. Acesso em: 15 de fevereiro de 2020.

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.

FANT, G. **Acoustic theory of speech production**: With calculations based oon x-ray studies of russian articulations. Netherlands: Walter de Gruyter, 1970.

FRIGO, M.; JOHNSON, S. **FFTW software**. 2015. Disponível em: http://www.fftw.org. Acesso em: 29 de janeiro de 2020.

FURUI, S. 50 years of progress in speech and speaker recognition research. **ECTI Transactions on Computer and Information Technology (ECTI-CIT)**, v. 1, n. 2, p. 64–74, 2005.

GAROFOLO, J. S.; LAMEL, L. F.; FISHER, W. M.; FISCUS, J. G.; PALLETT, D. S. Getting started with the DARPA TIMIT CD-ROM: an acoustic phonetic continuous speech database. **National Institute of Standards and Technology (NIST), Gaithersburgh, MD**, v. 107, p. 16, 1988.

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, 27. **Proceedings [...]**. Eilat, Israel: IEEE, 2012. p. 1–5.

GLOBAL INDUSTRY ANALYSTSINC. **Expanding Applications & Technology Developments Drive Growth in the Global Voice and Speech Recognition Technology Market**. 2018. Disponível em: https://strategyr.blogspot.com/2018/12/expanding-applications-technology.html?m=0. Acesso em: 6 de janeiro de 2020.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. Massachusetts, USA: MIT press, 2016. Disponível em: http://www.deeplearningbook.org. Acesso em: 22 de janeiro de 2020.

GOOGLE. **Google Acadêmico**. 2004. Disponível em: https://scholar.google.com.br/. Acesso em: 12 de fevereiro de 2020.

GRAY, R.; BUZO, A.; GRAY, A.; MATSUYAMA, Y. Distortion measures for speech processing. **IEEE Transactions on Acoustics, Speech, and Signal Processing**, IEEE, v. 28, n. 4, p. 367–376, 1980.

GUSMÃO, C. de. S.; CAMPOS, P. H.; MAIA, M. E. O. O formante do cantor e os ajustes laríngeos utilizados para realizá-lo: uma revisão descritiva. **Per Musi**, SciELO Brasil, n. 21, p. 43–50, 2010.

HAN, K.; WANG, Y.; WANG, D.; WOODS, W. S.; MERKS, I.; ZHANG, T. Learning spectral mapping for speech dereverberation and denoising. **IEEE Transactions on Audio, Speech, and Language Processing**, IEEE Press, v. 23, n. 6, p. 982–992, 2015.

HAVELOCK, D.; KUWANO, S.; VORLÄNDER, M. **Handbook of signal processing in acoustics**. New York, NY, USA: Springer Science & Business Media, 2008. v. 1. ISBN 978-0-387-77698-9.

HAYKIN, S. **Neural networks and learning machines, 3/E**. 3. ed. Upper Saddle River, New Jersey, USA: Pearson Education India, 2009. ISBN 978-0-13-147139-9.

HILL, P. **Audio and Speech Processing with MATLAB**. Boca Raton, FL, USA: CRC Press, 2019. ISBN 978-1-4987-6274-8.

HINTON, G. E. A practical guide to training restricted Boltzmann machines. In: **Neural networks: Tricks of the trade**. 2. ed. Toronto, Canada: Springer, 2012. p. 599–619. ISBN 978-3-642-35288-1.

HIRSCH, H.-G.; PEARCE, D. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: ASR2000-AUTOMATIC SPEECH RECOGNITION: CHALLENGES FOR THE NEW MILLENIUM ISCA TUTORIAL AND RESEARCH WORKSHOP (ITRW). **Proceedings [...]**. Paris, France, 2000. ISCA Archive. Disponível em: http://www.isca-speech.org/archive_open/asr2000. Acesso em: 22 de janeiro de 2020.

HU, Y.; LOIZOU, P. C. Evaluation of objective quality measures for speech enhancement. **IEEE Transactions on audio, speech, and language processing**, IEEE, v. 16, n. 1, p. 229–238, 2007.

HUANG, X.; BAKER, J.; REDDY, R. A historical perspective of speech recognition. **Commun. ACM**, v. 57, n. 1, p. 94–103, 2014.

IBM. **Pioneering Speech Recognition**. 2015. Disponível em: https://www.ibm.com/ibm/history/ibm100/us/en/icons/speechreco/transform/. Acesso em: 8 de dezembro de 2020.

IEEE. **IEEE Explore**. 2004. Disponível em: https://ieeexplore.ieee.org/Xplore/home.jsp. Acesso em: 15 de fevereiro de 2020.

INGLE, J. G. P. V. K. **Digital Signal Processing Using Matlab: A Problem Solving Companion**. 4. ed. Boston, USA: Cengage Learning, 2016.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION (ISO). **ISO 12.001**: Noise emitted by machinery and equipment - rules for the drafting and presentation of a noise test code. [*S. l.*], 2010. Disponível em: https://www.iso.org/standard/21256.html.

JUANG, B.-H.; RABINER, L. R. Automatic speech recognition–a brief history of the technology development. **Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara**, v. 1, p. 67, 2005.

LATHI, B. P.; GREEN, R. A. **Linear systems and signals**. 2. ed. New York, USA: Oxford University Press, 2005.

LECUN, Y. Efficient learning and second-order methods. **A tutorial at NIPS**, Denver, USA, 1993.

LI, J.; DENG, L.; HAEB-UMBACH, R.; GONG, Y. **Robust automatic speech recognition: a bridge to practical applications**. Oxford, UK: Academic Press, 2016. ISBN 978-0-12-802398-3.

LIMA, I. de. A. **Redução de ruído sonoro aplicada reconhecimento automático de voz**. Dissertação (Mestrado em Engenharia Elétrica) – Centro de Engenharia Elétrica e Informática, Programa de Pós-Graduação em Engenharia Elétrica: Processamento da Informação/Comunicações, Universidade Federal de Campina Grande, Campina Grande, 2014.

LIU, D.; SMARAGDIS, P.; KIM, M. Experiments on deep learning for speech denoising. In: ANNUAL CONFERENCE OF THE INTERNATIONAL SPEECH COMMUNICATION ASSOCIATION, 15. **Proceedings [...]**. Singapore: International Speech Communication Association (ISCA), 2014. p. 2685–2689.

LOCKWOOD, P.; BOUDY, J. Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars. **Speech communication**, Elsevier, v. 11, n. 2-3, p. 215–228, 1992.

LOIZOU, P. NOIZEUS: A noisy speech corpus for evaluation of speech enhancement algorithms. **Speech Communication**, v. 49, p. 588–601, 2017.

LOIZOU, P. C. **Speech enhancement: theory and practice**. 2. ed. Boca Raton, FL, USA: CRC press, 2013. ISBN 978-1-4665-0422-6.

LOMBARD, E. Le signe de l'elevation de la voix. **Ann. Mal. de L'Oreille et du Larynx**, p. 101–119, 1911.

MADISCH, I.; HOFMAYER, S.; FICKENSCHER, H. **Research Gate**. 2008. Disponível em: https://www.researchgate.net/. Acesso em: 15 de fevereiro de 2020.

MARKIT, I. **Growing Virtual Personal Assistant Market Expands Significantly into New Vehicles**. 2019. Disponível em: https://news.ihsmarkit.com/prviewer/release_only/slug/automotive-growing-virtual-personal-assistant-market-expands-significantly-new-vehicle. Acesso em: 8 de janeiro de 2020.

MATHWORKS. **MATLAB Version 9.6.0 (R2019a)**. Natick, Massachusetts: The MathWorks Inc., 2019. Disponível em: https://www.mathworks.com/products/matlab.html.

MATHWORKS, Audio Toolbox Team. **Automatic speech-to-text conversion**. MATLAB Central File Exchange, 2020. Disponível em: https://www.mathworks.com/matlabcentral/fileexchange/65266-speech2text. Acesso em: 29 de janeiro de 2020.

MCLOUGHLIN, I. V. **Speech and audio processing: a MATLAB-based approach**. Cambridge, UK: Cambridge University Press, 2016. ISBN 978-1-107-08546-6.

MIHOV, S. G.; IVANOV, R. M.; POPOV, A. N. Denoising speech signals by wavelet transform. **Annual Journal Of Electronics**, n. 6, p. 2–5, 2009.

OPPENHEIM, A.; SCHAFER, R. **Processamento em tempo discreto de sinais**. 3. ed. São Paulo: Pearson Education do Brasil, 2013. ISBN 978-85-8143-102-4.

PARK, S. R.; LEE, J. A fully convolutional neural network for speech enhancement. **CoRR**, abs/1609.07132, 2016. Disponível em: http://arxiv.org/abs/1609.07132. Acesso em: 22 de janeiro de 2020.

PEARSONS, K. S.; BENNETT, R. L.; FIDELL, S. **Speech levels in various noise environments**. [*S. l.*]: Office of Health and Ecological Effects, Office of Research and Development, 1977.

PIGNATARI, S. S. N.; ANSELMO-LIMA, W. T. **Tratado de otorrinolaringologia**. [*S. l.*]: Rio de Janeiro: Elsevier, 2018.

POLITYKO, E. **Word Error Rate**. MATLAB Central File Exchange, 2020. Disponível em: https://www.mathworks.com/matlabcentral/fileexchange/55825-word-error-rate. Acesso em: 5 de março de 2020.

POULARIKAS, A. D. **Adaptive filtering**: Fundamentals of least mean squares with MATLAB®. Boca Raton, FL, USA: CRC Press, 2015. ISBN 978-1-4822-5336-8.

REDDY, D. R. Speech recognition by machine: A review. **Proceedings of the IEEE**, IEEE, v. 64, n. 4, p. 501–531, 1976.

RIX, A. W.; BEERENDS, J. G.; HOLLIER, M. P.; HEKSTRA, A. P. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, 26. **Proceedings [...]**. Salt Lake City, Utah, USA, 2001. v. 2, p. 749–752.

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. **Psychological review**, American Psychological Association, v. 65, n. 6, p. 386, 1958.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. **nature**, Nature Publishing Group, v. 323, n. 6088, p. 533–536, 1986.

SALAKHUTDINOV, R.; MNIH, A.; HINTON, G. Restricted Boltzmann machines for collaborative filtering. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 24. **Proceedings [...]**. Corvalis, Oregon, EUA: Association for Computing Machinery, New York, NY, United States, 2007. p. 791–798. ISBN 978-1-59593-793-3.

SANTOS, J. C. M. dos. **Redução de ruído em sinais de voz combinando filtro de Kalman e transformada wavelet**. Dissertação (Mestrado em Engenharia Elétrica) – Faculdade de Engenharia Elétrica, Programa de Pós-Graduação em Ciência, Universidade Federal de Uberlândia (UFU), Uberlândia, 2015.

SCALART, P. *et al.* Speech enhancement based on a priori signal to noise estimation. In: INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING CONFERENCE, 21. **Proceedings [...]**. Washington, USA: IEEE Computer Society, 1996. v. 2, p. 629–632.

SCHNUPP, J.; NELKEN, I.; KING, A. **Auditory neuroscience: Making sense of sound**. London, England: MIT press, 2011. ISBN 978-0-262-11318-2.

SILVA, L. A. da. **Filtros de Kalman no tempo e freqüência discretos combinados com subtração espectral**. Dissertação (Mestrado em Engenharia Elétrica) – Escola de Engenharia de São Carlos, Programa de Pós-Graduação em Engenharia Elétrica, Universidade de São Paulo, São Carlos, 2007.

TAAL, C. H.; HENDRIKS, R. C.; HEUSDENS, R.; JENSEN, J. **Matlab implementation of the Short-Time Objective Intelligibility (STOI)**. 2010. Disponível em: http://insy.ewi.tudelft.nl/content/short-time-objective-intelligibility-measure. Acesso em: 28 de janeiro de 2020.

TAAL, C. H.; HENDRIKS, R. C.; HEUSDENS, R.; JENSEN, J. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, 26. **Proceedings [...]**. Dallas, EUA, 2010. p. 4214–4217.

TAAL, C. H.; HENDRIKS, R. C.; HEUSDENS, R.; JENSEN, J. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. **IEEE Transactions on Audio, Speech, and Language Processing**, IEEE, v. 19, n. 7, p. 2125–2136, 2011.

TAHA, T. M.; HUSSAIN, A. A survey on techniques for enhancing speech. **International Journal of Computer Applications**, v. 179, n. 17, p. 1–14, 2018.

WEISSER, A.; BUCHHOLZ, J. M. Conversational speech levels and signal-to-noise ratios in realistic acoustic conditions. **The Journal of the Acoustical Society of America**, Acoustical Society of America, v. 145, n. 1, p. 349–360, 2019.

XU, Y.; DU, J.; DAI, L.-R.; LEE, C.-H. An experimental study on speech enhancement based on deep neural networks. **IEEE Signal processing letters**, IEEE, v. 21, n. 1, p. 65–68, 2014.

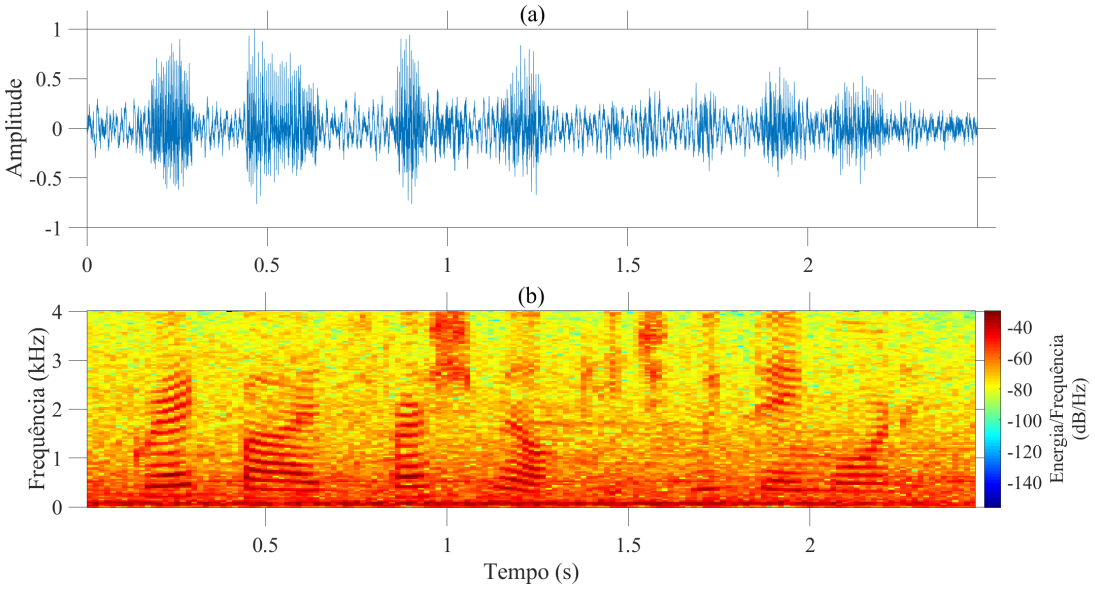
XU, Y.; DU, J.; DAI, L.-R.; LEE, C.-H. A regression approach to speech enhancement based on deep neural networks. **IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)**, IEEE Press, v. 23, n. 1, p. 7–19, 2015.

YOUNG, S. Frederick jelinek 1932-2010: The pioneer of speech recognition technology. **Speech and Language Processing Technical Committee Newsletter**, IEEE Signal Processing Society, 2010. Disponível em: http://archive.signalprocessingsociety.org/technical-committees/list/sl-tc/spl-nl/2010-11/jelinek/. Acesso em: 8 de dezembro de 2020.

YU, D.; DENG, L. **Automatic speech recognition: A deep learning approach**. [*S. l.*]: Springer, 2015. v. 2. ISBN 978-1-4471-5779-3.

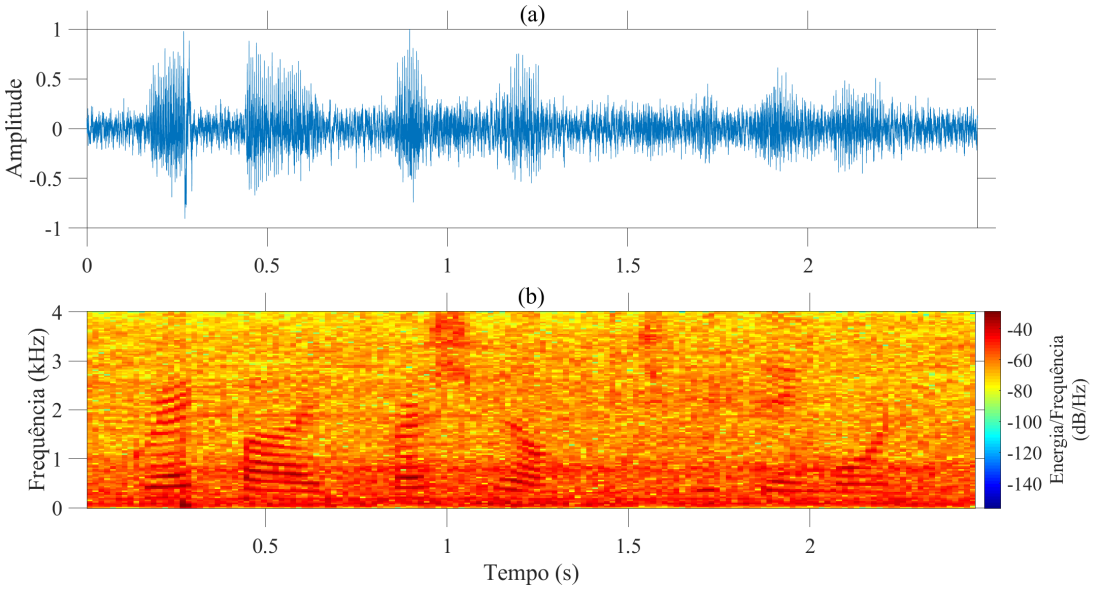
APÊNDICE A – RESULTADOS COMPLEMENTARES

Figura 51 – (a) Sinal da Figura 30 degradado com ruído trem e SNR 0 dB no domínio do tempo e (b) o respectivo espectrograma do sinal degradado.



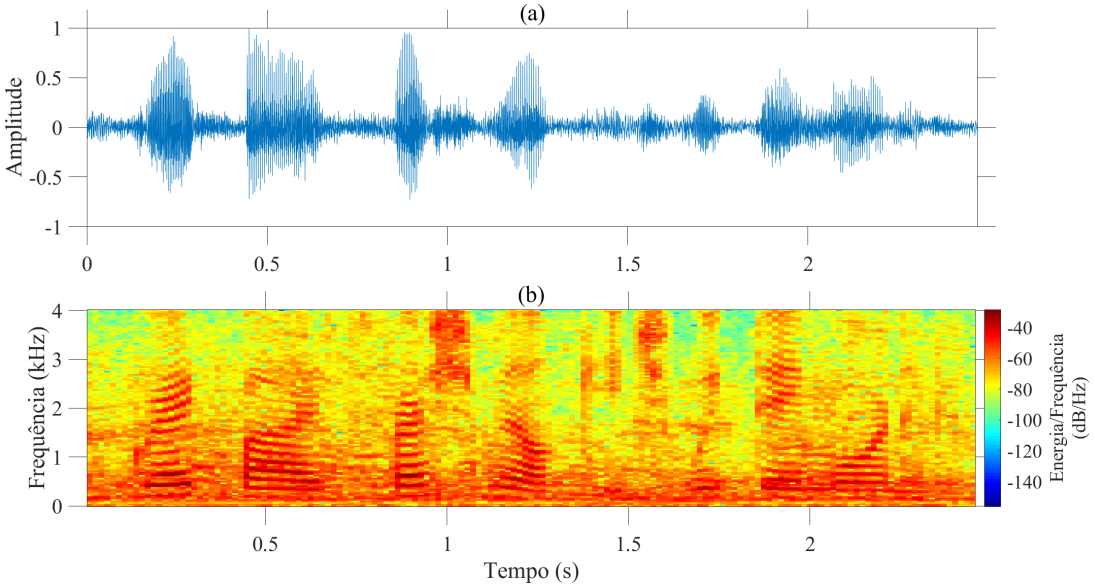
Fonte: elaborado pelo autor.

Figura 52 – (a) Sinal da Figura 30 degradado com ruído carro e SNR 0 dB no domínio do tempo e (b) o respectivo espectrograma do sinal degradado.



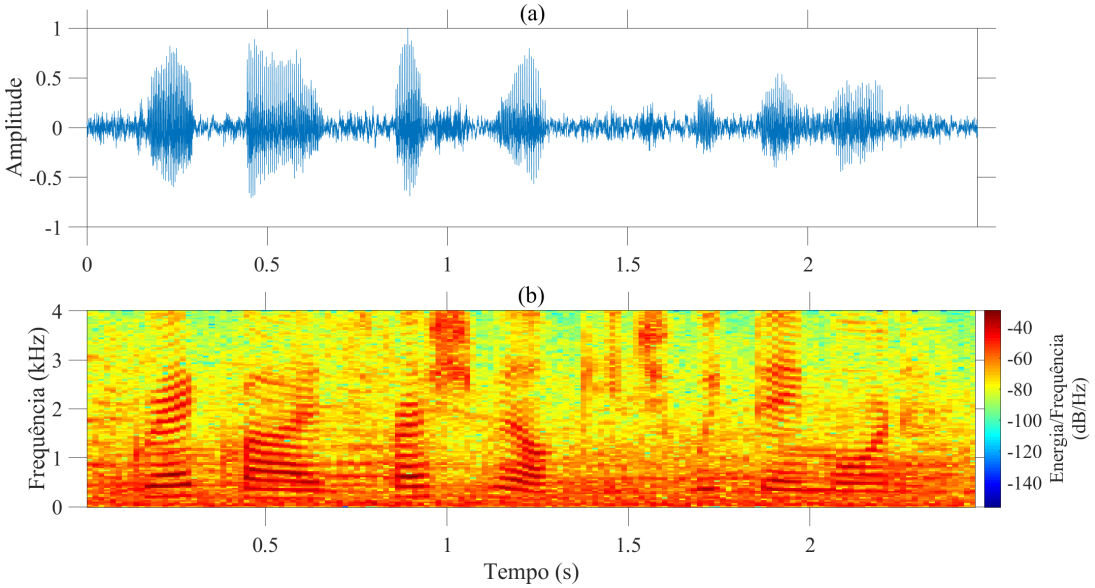
Fonte: elaborado pelo autor.

Figura 53 – (a) Sinal da Figura 30 degradado com ruído balbucios e SNR 7 dB no domínio do tempo e (b) o respectivo espectrograma do sinal degradado.



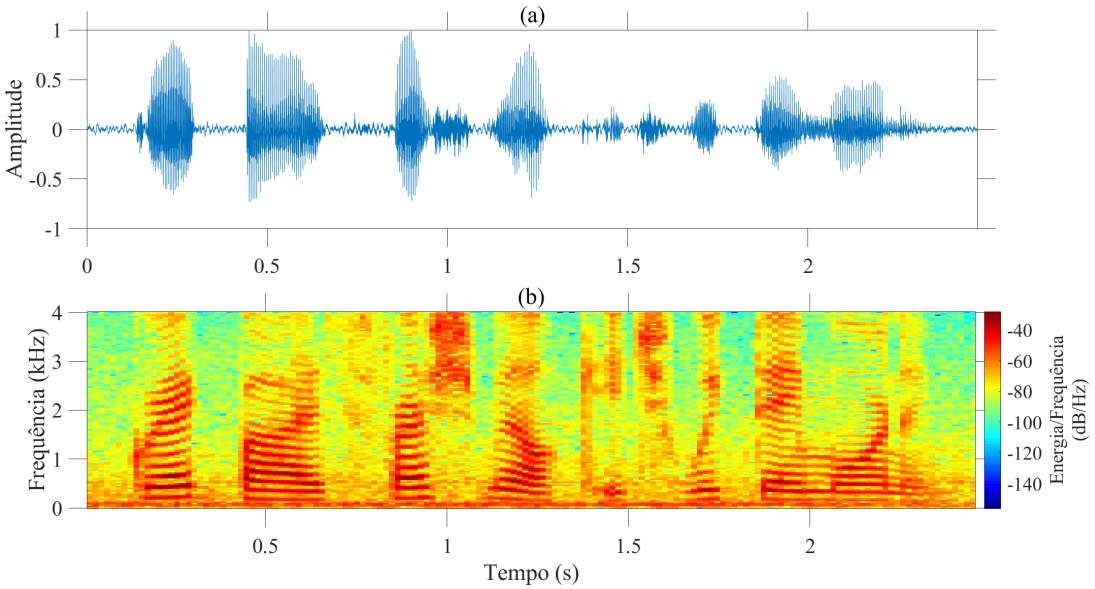
Fonte: elaborado pelo autor.

Figura 54 – (a) Sinal da Figura 30 degradado com ruído aeroporto e SNR 7 dB no domínio do tempo e (b) o respectivo espectrograma do sinal degradado.



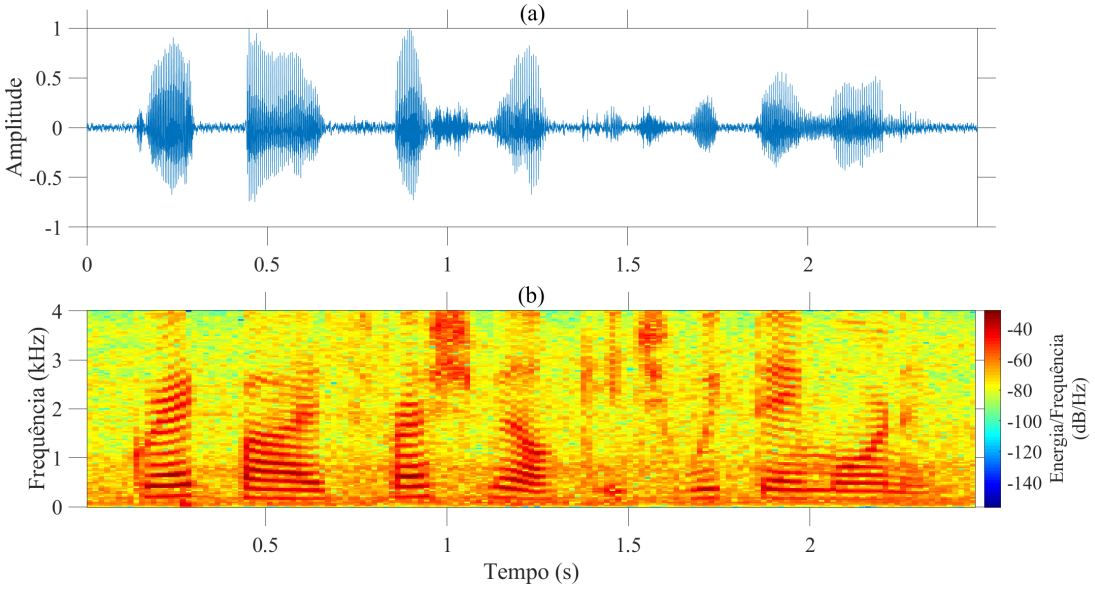
Fonte: elaborado pelo autor.

Figura 55 – (a) Sinal da Figura 30 degradado com ruído trem e SNR 15 dB no domínio do tempo e (b) o respectivo espectrograma do sinal degradado.



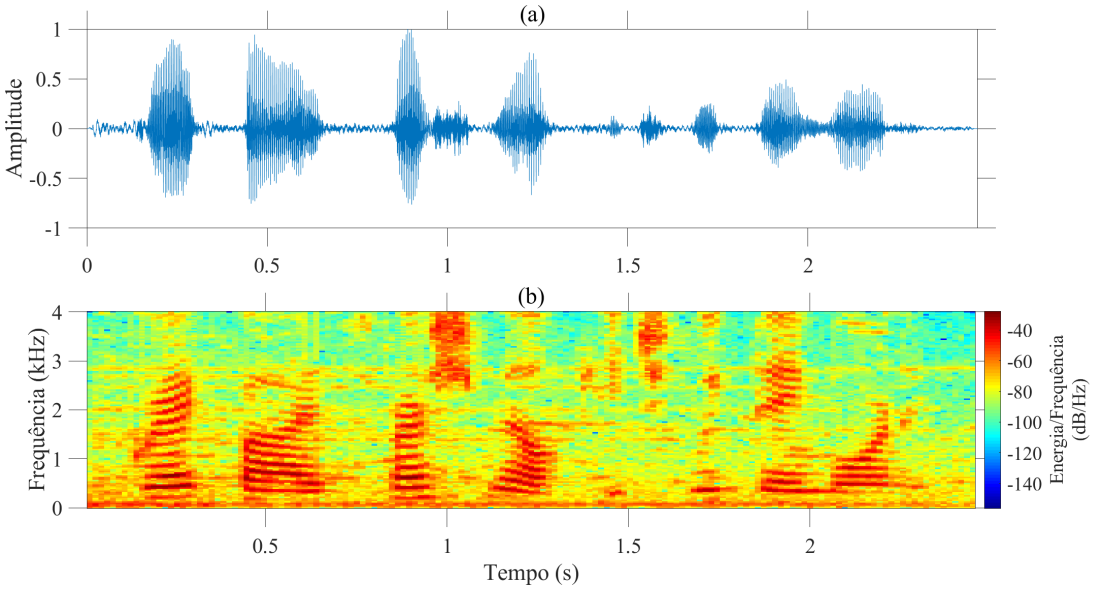
Fonte: elaborado pelo autor.

Figura 56 – (a) Sinal da Figura 30 degradado com ruído carro e SNR 15 dB no domínio do tempo e (b) o respectivo espectrograma do sinal degradado.



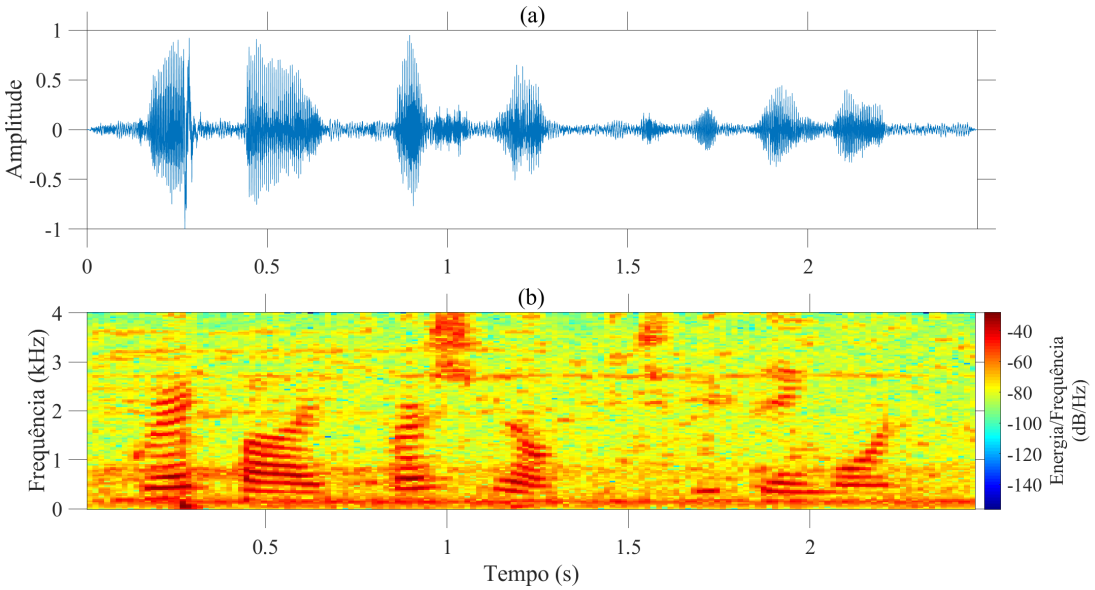
Fonte: elaborado pelo autor.

Figura 57 – Rede neural profunda (RNP): (a) estimação do sinal limpo da Figura 30 no domínio do tempo e (b) o respectivo espectrograma do sinal estimado, a partir do sinal da Figura 51, degradado com ruído trem e SNR de 0 dB.



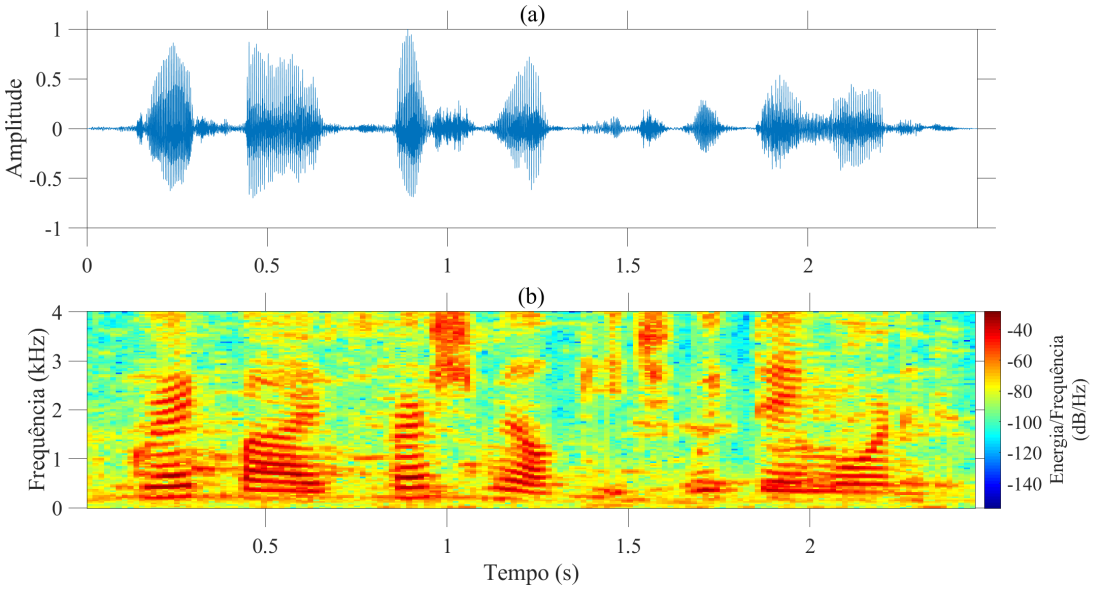
Fonte: elaborado pelo autor.

Figura 58 – Rede neural profunda (RNP): (a) estimação do sinal limpo da Figura 30 no domínio do tempo e (b) o respectivo espectrograma do sinal estimado, a partir do sinal da Figura 52, degradado com ruído carro e SNR de 0 dB.



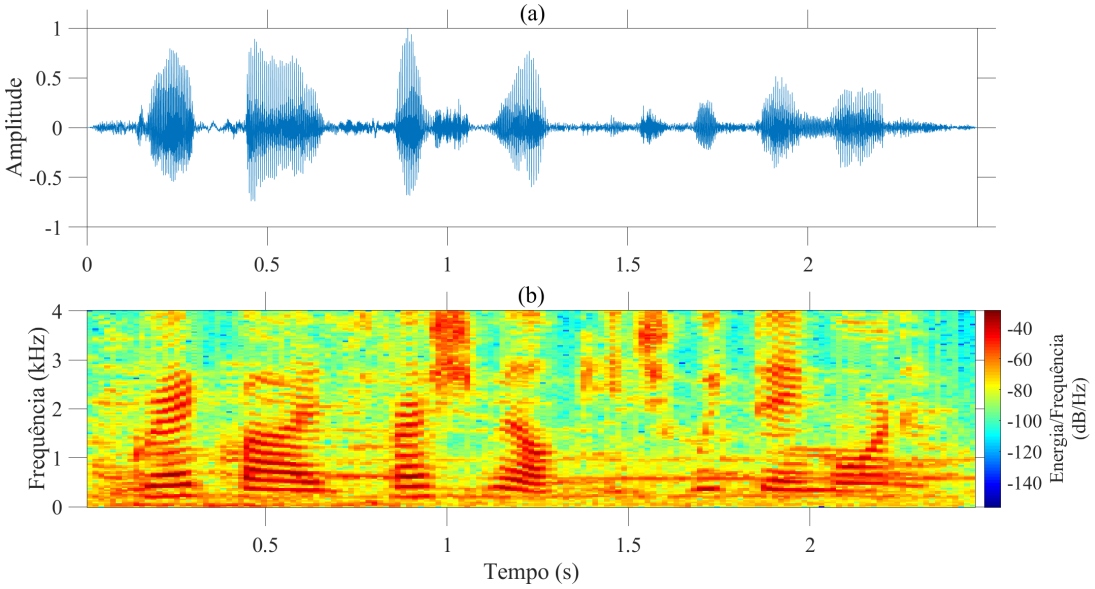
Fonte: elaborado pelo autor.

Figura 59 – Rede neural profunda (RNP): (a) estimação do sinal limpo da Figura 30 no domínio do tempo e (b) o respectivo espectrograma do sinal estimado, a partir do sinal da Figura 53, degradado com ruído balbucios e SNR de 0 dB.



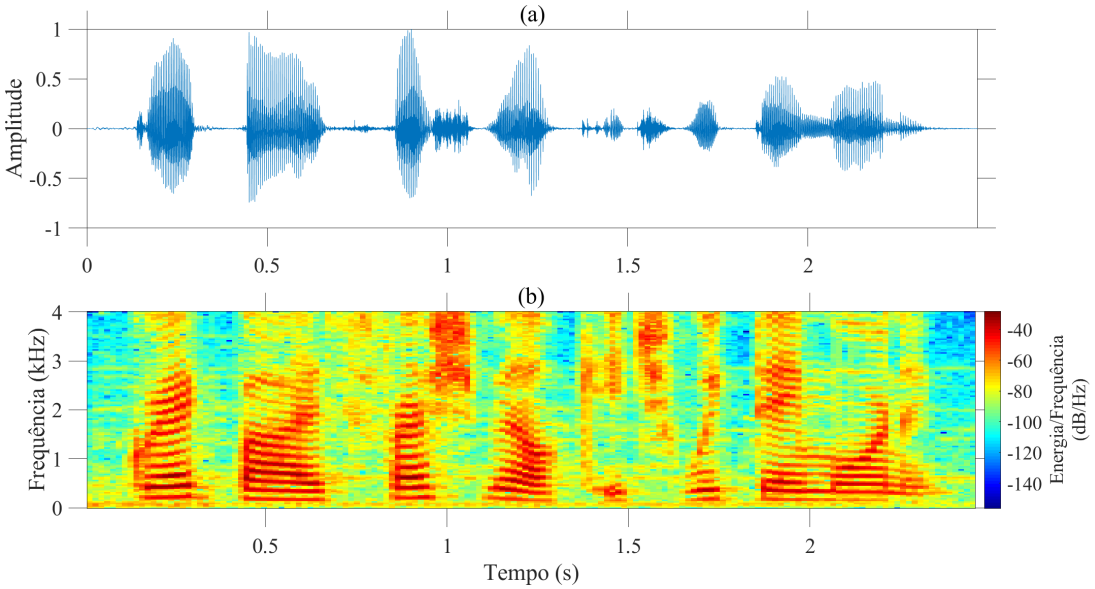
Fonte: elaborado pelo autor.

Figura 60 – Rede neural profunda (RNP): (a) estimação do sinal limpo da Figura 30 no domínio do tempo e (b) o respectivo espectrograma do sinal estimado, a partir do sinal da Figura 54, degradado com ruído aeroporto e SNR de 0 dB.



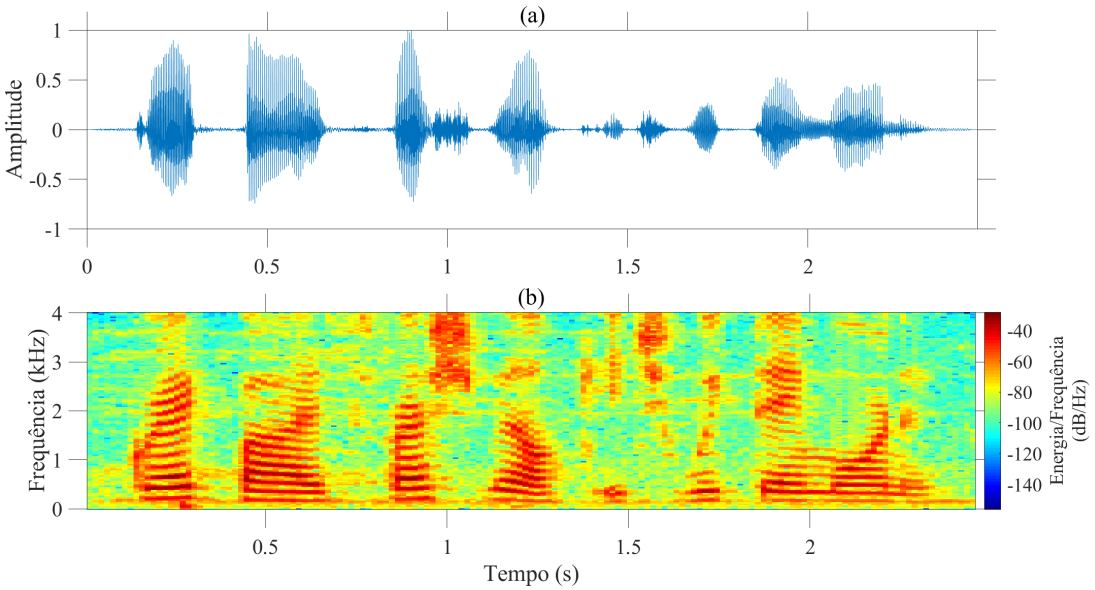
Fonte: elaborado pelo autor.

Figura 61 – Rede neural profunda (RNP): (a) estimação do sinal limpo da Figura 30 no domínio do tempo e (b) o respectivo espectrograma do sinal estimado, a partir do sinal da Figura 55, degradado com ruído trem e SNR de 15 dB.



Fonte: elaborado pelo autor.

Figura 62 – Rede neural profunda (RNP): (a) estimação do sinal limpo da Figura 30 no domínio do tempo e (b) o respectivo espectrograma do sinal estimado, a partir do sinal da Figura 56, degradado com ruído carro e SNR de 15 dB.



Fonte: elaborado pelo autor.

APÊNDICE B – PARÂMETROS DA RNP

Este apêndice lista o conjunto de valores e intervalos de valores de parâmetros e hiperparâmetros utilizados nos testes da RNP. Aspectos gerais, relativos à extração de características, ao processamento do algoritmo, ao número de camadas e neurônios da rede neural e à inicialização dos pesos sinápticos da matriz de pesos, são apresentados na Tabela 11. Parâmetros e hiperparâmetros relativos à etapa de pré-treinamento são descritos na Tabela 13. Já as configurações utilizadas na etapa de treinamento (ajuste-fino) são listadas na Tabela 12. O decaimento ponderado (regularização) não foi avaliado devido ao alto custo computacional necessário para se avaliar os resultados[1]. Estabelece-se também a condição de variabilidade do parâmetro, i.e., se permanece fixo ou variável ao longo dos testes. A escolha dos valores deve levar em consideração as recomendações de Yu e Deng (2015), Haykin (2009), Hinton (2012), de tal forma que a RNP não divirja.

Tabela 11 – Parâmetros gerais de configuração da RNP.

Parâmetro/Hiper-parâmetro		Valor/Intervalo	Configuração utilizada
frequencia de amostragem	fixo	8000	8000
tamanho do quadro (ms)	fixo	32	32
expansão de quadros	variável	1 a 21	11
sobreposição de quadros (%)	fixo	50 %	50 %
% de arquivos (treinamento)	variável	10 % a 100 %	100%
quota de treinamento	fixo	16, 32 ou 64	16
número de camadas	variável	1 a 3	3
número de neurônicos na 1a camada	variável	entre 256 e 8192	2048
número de neurônicos na 2a camada	variável	entre 256 e 8192	2048
número de neurônicos na 3a camada	variável	entre 256 e 8192	2048
número de neurônicos na 4a camada	variável	entre 256 e 8192	-
desvio padrão para inicialização das matrizes de peso	fixo	1e-2	1e-2

Fonte: elaborado pelo autor.

[1] O cálculo da norma das matrizes de peso consome bastante tempo a cada lote de arquivos processados (cerca de 3 segundos/lote).

Tabela 12 – Parâmetros do estágio de pré-treinamento da RNP.

Parâmetro/Hiper-parâmetro		Valor/Intervalo	Configuração utilizada
número de épocas para o pré-treinamento	fixo	20	20
taxa de aprendizagem da 1a camada (pré-treinamento)	variável	1e-6 a 1e-3	5,12e-4
taxa de aprendizagem da 2a camada (pré-treinamento)	variável	1e-6 a 1e-3	2,56e-4
taxa de aprendizagem da 3a camada (pré-treinamento)	variável	1e-6 a 1e-3	1,28e-4
taxa de aprendizagem da 4a camada (pré-treinamento)	variável	1e-7 a 1e-4	2e-7
momento inicial (pré-treinamento)	fixo	0,5	5e-1
momento final (pré-treinamento)	fixo	0,9	9e-1
número de passos (divergência contrastiva)	fixo	1	1

Fonte: elaborado pelo autor.

Tabela 13 – Parâmetros do estágio de treinamento (ajuste-fino) da RNP.

Parâmetro/Hiper-parâmetro		Valor/Intervalo	Configuração utilizada
número de épocas (treinamento)	fixo	20	20
taxa de aprendizagem da 1a camada (treinamento)	variável	8e-4 a 5e-2	5,12e-2
taxa de aprendizagem da 2a camada (treinamento)	variável	8e-4 a 5e-2	2,56e-2
taxa de aprendizagem da 3a camada (treinamento)	variável	8e-4 a 5e-2	1,28e-2
taxa de aprendizagem da 4a camada (treinamento)	variável	8e-4 a 5e-2	1,6e-4
fator gama do decaimento ponderado (treinamento)	variável	1e-3 a 1e-5	0
momento inicial (treinamento)	fixo	5e-1	5e-1
momento final (treinamento)	fixo	9e-1	9e-1
dropout (treinamento)	variável	0 a 10 %	4%
tamanho do mini-grupo	variável	32, 64 ou 128	32

Fonte: elaborado pelo autor.