UNIVERSIDADE FEDERAL DO MARANHÃO

Programa de Pós-Graduação em Ciência da Computação

Maurício César Pinto Pessoa

Remoção de ruídos aditivos e segmentação de palavras-chave em áudios

São Luís
2018

UNIVERSIDADE FEDERAL DO MARANHÃO - UFMA
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

MAURÍCIO CÉSAR PINTO PESSOA

REMOÇÃO DE RUÍDOS ADITIVOS E SEGMENTAÇÃO DE PALAVRAS-CHAVE EM ÁUDIOS

SÃO LUÍS
2018

MAURÍCIO CÉSAR PINTO PESSOA

REMOÇÃO DE RUÍDOS ADITIVOS E SEGMENTAÇÃO DE PALAVRAS-CHAVE EM ÁUDIOS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da UFMA como requisito para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Tiago Bonini Bochartt

São Luís

2018

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Núcleo Integrado de Bibliotecas/UFMA

César Pinto Pessoa, Maurício.
Remoção de ruídos aditivos e segmentação de palavras-chave em áudios / Maurício César Pinto Pessoa. - 2018.
82 p.

Orientador(a): Tiago Bonini Borchartt.
Dissertação (Mestrado) - Programa de Pós-graduação em Ciência da Computação/ccet, Universidade Federal do Maranhão, Universidade Federal do Maranhão, 2018.

1. Processamento de áudio. 2. Redes geradoras adversárias. 3. Remoção de ruídos. 4. Segmentação de áudio. 5. Wavelets. I. Bonini Borchartt, Tiago. II. Título.

Dissertação de autoria de Maurício César Pinto Pessoa, sob o título **"Remoção de ruídos aditivos e segmentação de palavras-chave em áudios"**, apresentado ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal do Maranhão, como requisito para obtenção do título de Mestre em Ciência da Computação, aprovado em 29 de agosto de 2018 pela comissão examinadora constituída pelos doutores:

Prof. Dr. Tiago Bonini Borchartt
Orientador
Universidade Federal do Maranhão

Prof. Dr. Geraldo Braz Junior
Examinador Interno
Universidade Federal do Maranhão

Prof. Dr. André Luiz Brandão
Examinador Externo
Universidade Federal do ABC

A todos que acreditaram em mim e não me deixaram desistir.

Agradecimentos

Agradeço a Deus, por me dar saúde e a força necessária para superar os desafios da construção desse trabalho, além de colocar na minha vida todos as pessoas abaixo mencionadas.

Aos meus pais, Conceição e Gerisval, por me criarem e sempre insistirem na importância dos estudos, e por todo o amor, carinho e conhecimento recebido no passar dos anos. Esse mérito também é de vocês.

Ao meu irmão Alexandre, pela inúmeras discussões e troca de ideias sobre esse trabalho, sua ajuda fez a qualidade desse trabalho melhorar muito.

Agradeço à Amanda, minha querida namorada e companheira, pela paciência, amor e carinho, e principalmente por não me deixar desistir desse mestrado e ter me ajudado a colocar a cabeça no lugar nas horas difíceis.

Ao meu amigo e orientador, Tiago Bonini, que sempre buscou me ajudar a finalizar esse trabalho da melhor maneira possível.

Ao amigo Geraldo Braz Junior, que em diversos momentos também me deu apoio no desenvolvimento deste trabalho.

A todos os demais que compõem a PPGCC-UFMA: professores, alunos e funcionários.

Aos amigos de laboratório, o VIP Lab foi para nós como uma segunda casa durante esses dois anos. Em especial a: Italo Francyles, sem o qual eu não teria concluído a experimentação do trabalho. Jullyana Fialho, que também muito me ajudou em diversos pontos. E a todos os demais do laboratório cujo nomes não estão aqui, mas que me ajudaram e me apoiaram.

Agradeço aos demais amigos da UFMA, que em várias conversas de corredor foram me dando novas ideias para incrementar esse trabalho. E em especial ao Matheus Menezes, e Matheus Lisboa que também ajudaram na experimentação.

Aos meus amigos estrangeiros, que também sempre me ajudaram no que puderam, principalmente nos momentos de dificuldade. Em especial agradeço a: Adam McNeilly, Mehdi Bendriss, Eric Cugota, Dave Matthew, Raghav Sood, Matthieu Honel, Tim Castelijns, Layne Lund, Mark O' Sullivan, e os demais da *"Room 15"*, vocês são demais.

A CAPES, por viabilizar essa pesquisa através da bolsa de estudos.

Como de praxe, também agradeço ao Google e ao StackOverflow, pela quase infinita fonte de conhecimento que essas ferramentas provêm.

Aos pesquisadores Nikolay Shmyrev e Harmandeep Singh, cuja ajuda foi fundamental em algumas etapas do trabalho.

Por fim, agradeço a todos os demais que de alguma forma contribuíram para que esse trabalho se tornasse realidade.

"A ciência nunca resolve um problema sem criar pelo menos outros dez."

(George Bernard Shaw)

Resumo

A presença de ruídos aditivos é um dos principais problemas em sistemas de reconhecimento de áudio digital, pois dificultam a etapa de segmentação dos trechos relevantes de áudio, além de reduzir o desempenho dos classificadores. O principal objetivo desse trabalho é desenvolver um método de remoção de ruído e segmentação em arquivos de áudio digital, com foco nos arquivos gerados pelo método de observação direta, onde um observador grava em áudio todas as ações executadas pelo espécime observado de forma codificada em *Bite Categories*. Esse método pré-processa os arquivos de áudio a fim de normalizá-los e de reduzir sua dimensionalidade, posteriormente sendo utilizada a rede geradora adversária SEGAN para a remoção dos ruídos. A etapa de segmentação do áudio começa com um pré-processamento que atenua os vales do sinal e enfatiza os picos, de forma similar à normalização do sinal, seguido da aplicação de uma função de silenciamento de vales, com base no desvio padrão e escore padronizado. A segmentação é realizada a partir de uma função de mapeamento que encontra os tempos de início e fim de cada segmento com base na detecção de silêncios usando janelas deslizantes com sobreposição. Os testes de remoção de ruídos foram realizados através de um estudo duplo-cego, utilizando questionários com escala de *Likert* unipolar de 5 pontos e uma base de áudios compilada pelo autor, de forma a medir subjetivamente a qualidade do método, onde se obteve uma média 3,56 de 5 na remoção de ruídos e média 4,14 de 5 na qualidade geral do áudio. Os testes de segmentação foram realizados a partir de uma segunda base de áudios compilada pelo autor, onde se obteve um coeficiente de similaridade de Dice de 85,10% para os áudios sem ruído, 77,95% para os áudios ruidosos e 76,12% para os áudios com o ruído removido através da SEGAN. Após a apresentação dos resultados, compara-se o desempenho dos métodos propostos com alguns trabalhos relacionados presentes na literatura.

Palavras-chaves: processamento de áudio, remoção de ruídos, segmentação de áudio, redes geradoras adversárias, wavelets.

Abstract

The presence of additive noise is one of the main problems in digital audio recognition systems as they make it difficult to segment the audio relevant portions and may also reduce classifier performance. The main objective of this work is to develop a method of noise removal and segmentation in digital audio files generated by the direct observation method. This method is where an observer records, in audio, all the actions taken by a given specimen, coded in bite categories. This method preprocesses the audio files in order to normalize them and reduce their dimensionality, after which the SEGAN neural network is used to remove the noise. The audio segmentation step begins with a pre-processing that attenuates the signal valleys and emphasizes the peaks, similar to signal normalization. The pre-processing is followed by the application of the valley silencing function, based on the standard deviation and standardized score. Segmentation is performed by using a mapping function that finds the start and end times of each segment, using silence detection and overlapping sliding windows. The noise removal tests were performed through a double-blind study, using questionnaires with an unipolar 5-point Likert scale and an audio dataset compiled by the author, in order to subjectively measure the method's quality. Quality scores reached an average of 3.56 out of 5 on noise removal and an average of 4.14 out of 5 on overall audio quality. The segmentation tests were performed from a second audio dataset compiled by the author, and obtained Dice scores of 85.10% on the noiseless audios, 77.95% on the noisy audios, and 76.12% on the audios that had their noise removed through the SEGAN network. After the results are presented, a comparison is made between the obtained results and some related works currently present in the literature.

Keywords: Audio processing, noise removal, audio segmentation, generative adversarial networks, wavelets.

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

ADC	*Analog-Digital Converter*
ASR	*Automatic Speech Recognition*
BC	*Bite Categories*
BCG	*Bite Coding Grid*
CNN	*Convolutional Neural Networks*
CSD	Coeficiente de Similaridade de *Dice*
D	Discriminante
DCT	*Discrete Cosine Transform*
DEMAND	*Diverse Environments Multichannel Acoustic Noise Database*
DFT	*Discrete Fourier Transform*
FFT	*Fast Fourier Transform*
G	Gerador
GAN	*Generative Adversarial Networks*
Hz	*Hertz*
MF	*Mel-Frequency*
MFCC	*Mel-Frequency Cepstrum Coeficients*
MLP	*Multilayer Perceptron*
PRE	Precisão
RNA	Redes Neurais Artificiais
SCI	Segmento Corretamente Identificado
SEGAN	*Speech Enhancement Generative Adversarial Network*
SEN	Sensibilidade

SNR	Signal-to-noise ratio
SD	Segmentos Detectados
SII	Segmento Incorretamente Identificado

Sumário

1 Introdução

Com os avanços tecnológicos e o constante aumento de poder computacional, além do advento de novos algoritmos, o reconhecimento automático de voz, do inglês *Automatic Speech Recognition* (ASR), tornou-se uma alternativa viável para diversas aplicações computacionais que antes não eram possíveis (CODEN; BROWN; SRINIVASAN, 2002). Porém, antes de se conseguir realizar um reconhecimento automático de fala, é necessário encontrar e extrair do áudio os segmentos relevantes ao contexto da aplicação (HYOUNG-GOOK; MOREAU; SIKORA, 2005).

Uma característica dificultadora para esse tipo de aplicação consiste de ruídos aditivos presentes na maioria dos ambientes acústicos, o que limita o desempenho da segmentação, pois a presença desses ruídos é um dos fatores que mais reduzem a eficácia dos algoritmos de segmentação de áudio (SHRAWANKAR; THAKARE, 2010). Portanto, é fundamental a existência de boas técnicas que ajudem a atenuar ou remover os ruídos do áudio, efetivamente reduzindo a relação sinal-ruído, do inglês *Signal-to-noise ratio (SNR)* e consequentemente aumentando a eficiência das segmentações (SHRAWANKAR; THAKARE, 2010).

No contexto brasileiro, podemos citar como estudo de caso para a remoção de ruídos aditivos e a segmentação automática de áudios o método de observação direta de animais conhecidos como pequenos ruminantes.

A análise do comportamento ingestivo de pequenos ruminantes é de grande importância para a implantação de estratégias de suplementação alimentar destes animais, além da correta conservação dos seus ambientes pastorais (AGREIL; MEURET, 2004; TORRES-ACOSTA; CASTRO, 2014).

Os ruminantes são mamíferos herbívoros, cujo conjunto gástrico é constituído de vários estômagos. O nome ruminante tem origem na ação dos animais ruminarem os alimentos, ou seja, após inicialmente ingerir o alimento, ele é regurgitado na forma de um pequeno bolo, que é então novamente mastigado para ser posteriormente deglutido (SOEST, 1994; DIJKSTRA; FORBES; FRANCE, 2005; ZEN; SANTOS; MONTEIRO, 2014).

Segundo a Pesquisa da Pecuária Municipal (PPM) realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE), no Brasil, os pequenos ruminantes com maior presença

no mercado pecuário são os ovinos e caprinos, como ovelhas e cabras, que juntos superam 26 milhões de cabeças no país (IBGE, 2015).

A observação direta é um dos métodos mais utilizados para quantificar e qualificar o comportamento ingestivo desses animais, pois essa metodologia é predominantemente não invasiva e é adequada para aquisição de dados em ambientes pastorais com a presença de vegetação heterogênea (AGREIL; MEURET, 2004).

Esse método consiste em observar continuamente um determinado animal durante um período de tempo e anotar em um caderno, ou registrar em um gravador de áudio, o seu comportamento alimentício em tempo real para posterior formatação e análise. As características como o tamanho da mordida, o tipo e a parte da vegetação que o animal ingeriu, entre outras, são mapeadas para 33 categorias de mordidas (*Bite Categories*, ou BC) com base em uma grade de mordidas (*Bite Coding Grid* (BCG)), ilustrada pela Figura 1 (TORRES-ACOSTA; CASTRO, 2014).

Cada categoria na BCG representa um tipo diferente de mordida realizada pelo animal, levando em consideração o tamanho da bocada, o tipo de vegetação consumida, a altura em que se encontrava o alimento, entre outros fatores (TORRES-ACOSTA; CASTRO, 2014).

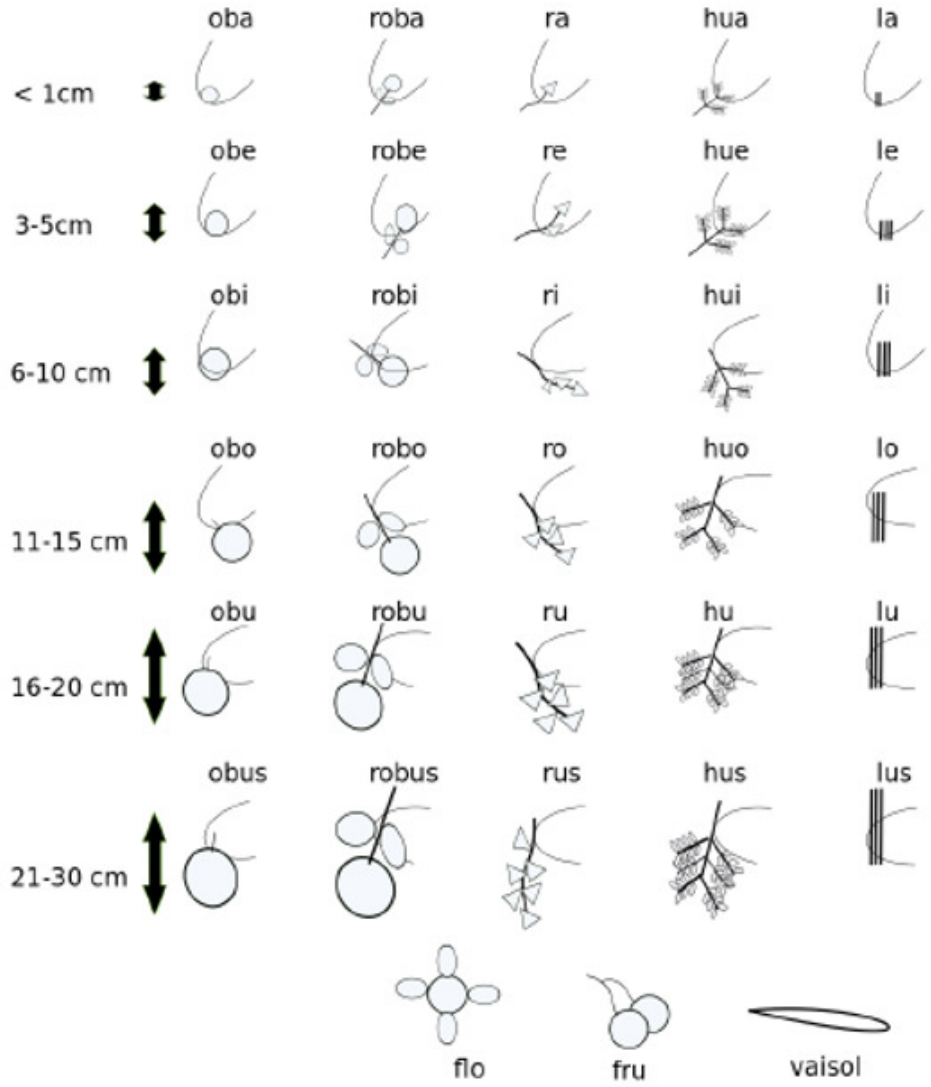
Os arquivos de áudio digital gerados no método de observação direta através de um gravador contêm trechos de fala com informações semânticas. Essas informações consistem das palavras faladas pelo locutor, ou nesse caso, das BC observadas. Uma das formas de analisar posteriormente esse conteúdo resume-se em colocar uma pessoa para ouvir o áudio e transcrever as informações relevantes no formato desejado, seja uma transcrição total, ou somente das palavras-chave.

Entretanto, devido ao grande volume de informações, essa prática manual torna-se inviável para a maioria das aplicações, fazendo-se necessária a presença de algum sistema automático de segmentação e reconhecimento de fala (HYOUNG-GOOK; MOREAU; SIKORA, 2005).

1.1 Objetivo

Destaca-se nesta seção os objetivos (geral e específicos) deste trabalho a serem desenvolvidos no decorrer da metodologia.

Figura 1 – *Bite Coding Grid.*



Fonte: Agreil e Meuret (2004).

1.1.1 Objetivo geral

O objetivo desse trabalho é desenvolver um método de remoção de ruídos e segmentação automática dos BC presentes nos arquivos de áudio gerados na observação direta de pequenos ruminantes.

1.1.2 Objetivos específicos

Para alcançar o objetivo geral deste trabalho, necessita-se passar por objetivos mais específicos, como destaca-se a seguir:

- Coleta de amostras de áudios com os BC que serão usados para treinamento e teste das técnicas de remoção de ruídos.
- Aplicar técnicas de filtragem e remoção de ruídos nos áudios utilizando redes neurais.
- Aplicar técnicas de segmentação de áudio, de forma a isolar os BC.
- Validar o desempenho da metodologia de acordo com as métricas de desempenho apresentadas na literatura.

1.2 Contribuições do trabalho

A metodologia proposta engloba as seguintes contribuições ao meio científico:

- Levantamento de amostras de áudios com os BC, podendo ser utilizada para treinar diversas arquiteturas de redes neurais.
- Levantamento de amostras de áudio paralelos, para serem utilizados no treinamento de redes neurais especializadas em remoção de ruídos.
- Metodologia para a remoção de ruídos aditivos e segmentação automática de pequenos códigos, como BC, em arquivos de áudio.

1.3 Organização do Trabalho

Além do capítulo introdutório, ainda há mais 5 capítulos completando esta dissertação. No Capítulo 2 são apresentados os trabalhos relacionados. O Capítulo 3 trata da fundamentação teórica, onde são mostrados os conceitos e técnicas necessárias para o entendimento da metodologia desenvolvida. No Capítulo 4 será apresentada a metodologia proposta, que está dividida em cinco etapas: aquisição da base de áudios, pré-processamento, remoção de ruídos, geração de base de testes e segmentação. O Capítulo 5 apresenta a experimentação realizada neste trabalho. O Capítulo 6 apresenta os resultados e discussões de cada etapa da metodologia desenvolvida. Por fim, no Capítulo 7, expõem-se as conclusões obtidas a partir da metodologia desenvolvida e mencionam-se os trabalhos para futuros aprimoramentos.

2 Trabalhos relacionados

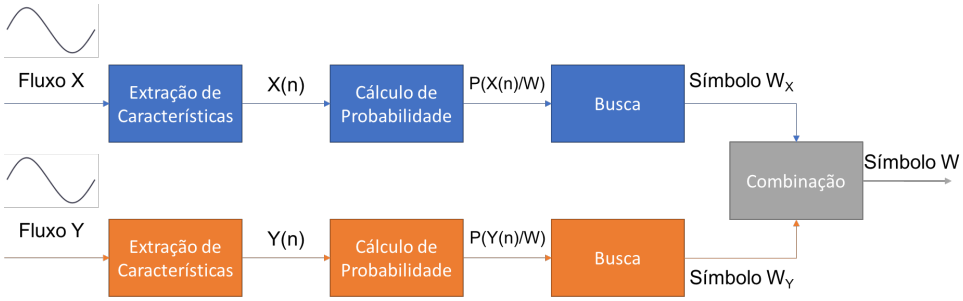
Os sistemas de reconhecimento de voz tem sua eficiência diretamente ligada à uma série de fatores, tais como o nível de ruído em relação ao sinal original a ser identificado, o tamanho do vocabulário, os modelos utilizados, as diferenças de sotaque ou idiomas dos locutores, as características utilizadas, entre outros (MORENO, 1996; LI, 2005; HYOUNG-GOOK; MOREAU; SIKORA, 2005). Nesta seção são apresentados alguns trabalhos que se propõem a resolver um ou mais desses fatores de forma a tornar os sistemas de reconhecimento mais eficientes.

A tese publicada por Moreno (1996) demonstra porque os sistemas de reconhecimento de voz possuem uma degradação de acurácia em ambientes ruidosos, além de propor métodos baseados em dados e em modelos matemáticos dos ambientes visando reduzir os erros gerados por esses ruídos. Para amenizar esse problema, o autor propôs duas famílias de algoritmos de compensação de ruído, denominadas *Multivariate-Gaussian-Based Cepstral Normalization* (RATZ), que busca compensar o ruído a partir das características de entrada, e o *Statistical Reestimation* (STAR), que busca compensar o ruído a partir da estrutura interna do modelo acústico. O RATZ obteve resultados equivalentes aos de sistemas sem ruído estando em ambientes com a relação sinal-ruído (SNR) (JOHNSON, 2006) de $15db$, enquanto o STAR obteve esse desempenho em ambientes com o SNR de $5db$.

Em Li (2005) é proposto um método de fusão de hipóteses paralelas. Esse método consiste em combinar os resultados individuais gerados pelo ASR com base em fontes de características distintas, de forma a maximizar o número de acertos do sistema. Ou seja, dado dois ou mais fluxos de características diferentes, o resultado final do sistema será a melhor combinação probabilística desses resultados gerados em paralelo, conforme ilustrado na Figura 2. Esse método obteve uma melhora de 15% quando comparado a sistemas com apenas um fluxo de características.

Ainda em Li (2005), vale ressaltar que o autor enfatiza que a performance dos ASR depende de dois fatores críticos. O primeiro fator consiste de quais características foram extraídas do sinal de entrada e sua respectiva relevância. O segundo fator são quais modelos estatísticos estão sendo usados para representar os símbolos da linguagem. Entre esses dois fatores, o primeiro pode ser considerado mais importante, devido ao fato dos

Figura 2 – Diagrama do método de fusão de hipóteses.

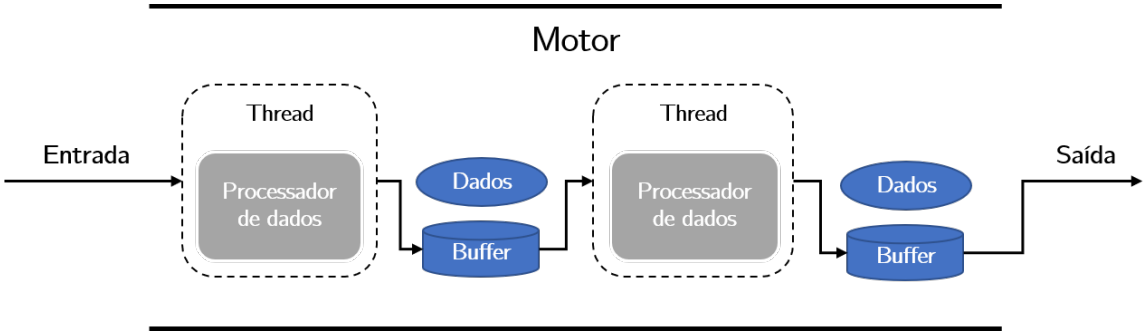


Fonte: Adaptado de Li (2005).

modelos de um ASR serem treinados para um subconjunto particular de características que são apresentadas para o sistema.

No trabalho publicado por Penagarikano e Bordel (2005) é introduzido o Sautrela, um *framework* modular de código aberto para processamento de áudio digital e reconhecimento automático de fala. O Sautrela foi desenvolvido utilizando a linguagem de programação Java e utiliza de uma abstração denominada processadores de dados, ou no inglês *Data-Processors*. Um processador de dados do Sautrela consiste de um conjunto de dados de entrada, uma função de processamento em cima desses dados e um conjunto de dados de saída. Essa abstração pode ser utilizada para qualquer função de processamento de sinais nesse *Framework* e cada processador de dados pode rodar em uma *thread* separada. Um conjunto de processadores de dados conectados como uma lista encadeada, juntamente com *buffers* de dados entre cada processador, é conhecido como um motor, ou no inglês *engine*. Esse motor é a base de funcionamento do Sautrela, pois ele define todo o processo de manipulação do sinal digital, com os seus valores de entrada, saída e todos os passos de processamento intermediários. A Figura 3 ilustra o motor de processamento do *framework*, contendo dois processadores de dados.

Figura 3 – Motor do Sautrela.



Fonte: Adaptado de Penagarikano e Bordel (2005).

Em Abushariah et al. (2010) é apresentado um sistema de reconhecimento de áudio para locutores árabes. No referido trabalho, os autores utilizaram o conjunto de ferramentas denominado *Sphinx* (LEE, 1988), que utiliza os Coeficientes Mel-Cepstrais e o Modelo Oculto de *Markov* como vetor de características e classificador, respectivamente. Esse sistema de reconhecimento de voz obteve acurácia de até 96,29% e para tal, utilizou uma base de treino com sete horas de áudio foneticamente diversificados e uma base de teste com mais uma hora de áudio.

Em Kalantarian e Sarrafzadeh (2015) os autores propõem um método de monitoramento de hábitos alimentares através do uso dos microfones embarcados em *smartwatches*. No referido trabalho, o microfone continuamente armazena trechos de áudio em um *buffer*, para que suas características sejam extraídas quando esse *buffer* atingir um determinado limite. Posteriormente elas são analisadas utilizando a Floresta Aleatória, do inglês *Random Forest* (RF) (HO, 1995), para identificar padrões de mordida e ingestão de líquidos e sólidos. Uma vez que um padrão é detectado, ele é enviado para o *smartphone* para fins de armazenamento e análise futura. O método proposto obteve 94,5% de acurácia a partir de 250 casos de teste.

Este presente trabalho apresenta como principais contribuições um processo de filtragem para remoção de ruídos usando redes neurais adversárias, conforme apresentado por Pascual, Bonafonte e Serrà (2017), além de uma metodologia de segmentação automática de áudios com base na detecção de trechos de silêncio e picos de energia no sinal de áudio.

3 Fundamentação teórica

Este capítulo trata da fundamentação teórica que expõe conceitos importantes para compreensão da metodologia proposta. Abordam-se os conceitos de som e processamento de áudio, reconhecimento de áudio, redes neurais e redes geradoras adversárias.

3.1 Áudio Digital

O som pode ser definido como uma vibração mecânica audível que se propaga através de algum meio material, como o ar (KNOBEL, ; GELFAND; LEVITT, 1998). Toda onda sonora, assim como outros tipos de onda, possuem uma frequência que normalmente é medida em Hertz (Hz) e uma energia, ou amplitude, que é normalmente medida em decibéis (db), sendo que o ouvido humano consegue perceber, em média, o som com frequências de $20Hz$ até $20kHz$ (GELFAND; LEVITT, 1998).

O som nesse estado natural é conhecido como som analógico, e para que um computador consiga realizar cálculos em cima dessa grandeza física, torna-se necessária a existência de um processo de amostragem, onde esse sinal é medido em intervalos regulares e cada amostra é convertida em valores discretos, nesse caso, em bits.

O processo de amostragem é feito por um dispositivo conhecido como conversor analógico-digital, do inglês *Analog-Digital Converter (ADC)*, em conjunto com um microfone. O microfone registra a pressão acústica do som e a transforma em uma tensão elétrica analógica, que é então quantizada pelo ADC e posteriormente armazenada em sua versão digital (SHANNON, 1949). A Figura 4 ilustra o processo de registro e amostragem de um sinal analógico para a sua versão digital.

Figura 4 – Diagrama de uma amostragem de áudio.

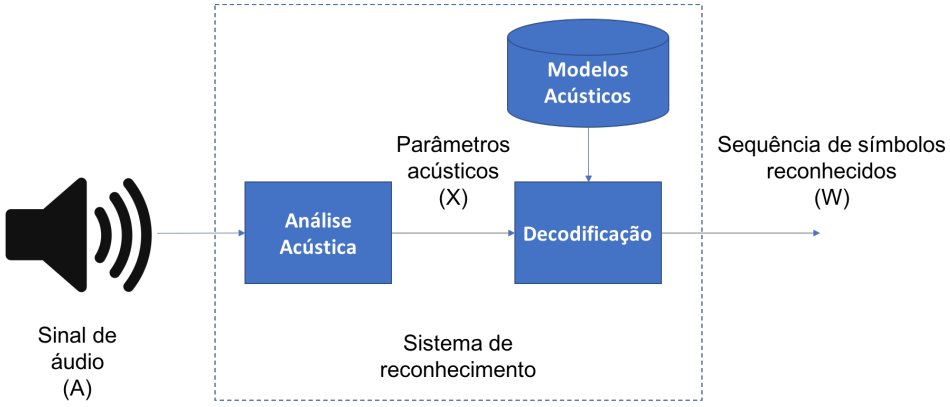
Fonte: Elaborado pelo autor.

Segundo o Teorema de Nyquist–Shannon, para que um sinal analógico possua uma boa representação em sua versão digital, é necessário que a taxa de amostragem do sinal original seja o dobro da frequência mais alta presente nessa onda, por exemplo, para que um áudio digital consiga reproduzir fielmente a qualidade do sinal analógico, é necessário que ele possua pelo menos 40.000 amostras por segundo, visto que esse valor é o dobro da maior frequência que o ouvido humano consegue perceber (GELFAND; LEVITT, 1998; SHANNON, 1949).

3.2 Sistemas Automáticos de Reconhecimento de Voz

Os sistemas automáticos de reconhecimento de voz, do inglês *Automatic Speech Recognition* (ASR), podem ser resumidos em duas etapas, conforme ilustrado na Figura 5.

Figura 5 – Diagrama de um ASR.



Fonte: Adaptado de Hyoung-Gook, Moreau e Sikora (2005).

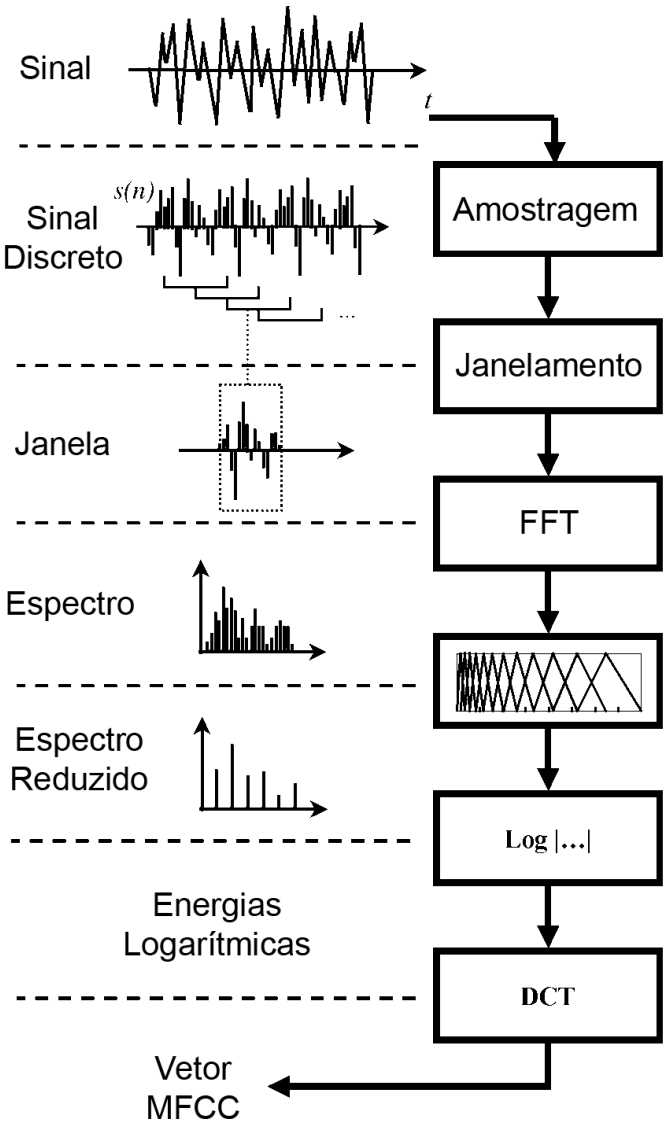
A primeira etapa, conhecida como análise acústica, consiste da extração de um vetor de coeficientes X do sinal de áudio A, esse vetor é uma representação das características de um quadro do sinal A. A segunda etapa, chamada de decodificação, consiste em encontrar as melhores correspondências para o vetor de coeficientes X com base em modelos acústicos previamente treinados. Cada um desses modelos corresponde a um símbolo que descreve a língua falada, como palavras, fonemas ou sílabas. O vetor resultante W consiste das melhores correspondências entre o sinal observado e os modelos existentes (HYOUNG-GOOK; MOREAU; SIKORA, 2005; LI, 2005).

A análise acústica pode ser dividida em seis passos. O primeiro passo constitui-se da digitalização do sinal analógico. O segundo passo consiste em aplicar um filtro passa-alta no sinal digitalizado, de forma a enfatizar as frequências mais altas. O terceiro passo é

a segmentação do áudio em pequenas janelas de tamanho fixo, que podem se sobrepor. O quarto passo aplica uma função de janelamento em cada segmento. No quinto passo, o espectro de frequência é extraído de cada janela aplicando a Transformada Rápida de Fourier, do inglês *Fast Fourier Transform* (FFT)(WALKER, 1996). Por fim, no sexto passo, é extraído um vetor de coeficientes X, sendo então as características propriamente ditas do sinal de áudio.

Esse processo é ilustrado conforme a Figura 6, sendo que o vetor de coeficientes gerado no último passo são os Coeficientes Mel-Cepstrais (MFCC) (LOGAN et al., 2000; MUDA; BEGAM; ELAMVAZUTHI, 2010). Esse vetor de características é utilizado como entrada para a segunda etapa do ASR (HYOUNG-GOOK; MOREAU; SIKORA, 2005; LI, 2005).

Figura 6 – Extração do vetor de características MFCC.



Fonte: Traduzido de Hyoung-Gook, Moreau e Sikora (2005).

O processo de decodificação de um ASR probabilístico tem como objetivo determinar a sequência de símbolos W mais provável a partir do vetor de características X obtido na etapa de análise acústica.

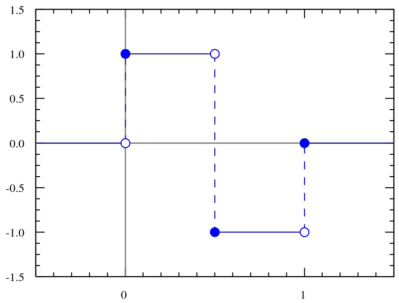
3.3 Wavelets

As *Wavelets* são funções matemáticas utilizadas para a decomposição hierárquica de outras funções ou séries de dados, esse processo de decomposição é conhecido como transformada *wavelet* e ela pode ser feita de forma discreta ou contínua. Suas aplicações abrangem diversas áreas das ciências e engenharias, sendo utilizadas para compressão de dados, remoção de ruídos, detecção de bordas, entre outras.

A análise *wavelet* de um sinal utiliza pequenas ondas conhecidas como *wavelets*, onde o sinal original pode ser decomposto em componentes constituídos por essas bases de *wavelets* e ser reconstruído como uma superposição dessas funções base (DAUBECHIES, 1992; BÉNÉTEAU; FLEET, 2011).

Como exemplos, podemos citar a *wavelet* de *Haar*, que é a primeira e mais simples *wavelet* existente (HAAR, 1910; STOLLNITZ; DEROSE; SALESIN, 1995), e a *wavelet Symlet* introduzida por Ingrid Daubechies (DAUBECHIES, 1992), exemplos de ambas podem ser visualizadas conforme as Figuras 7 e 8.

Figura 7 – *Wavelet* de Haar.

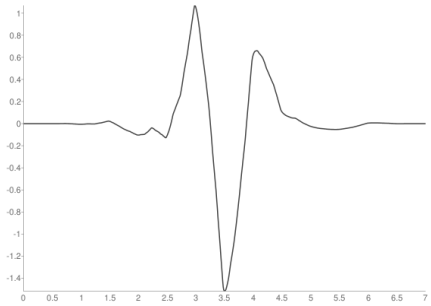


Fonte: Elaborado pelo autor.

3.3.1 Transformada de Haar de uma dimensão

A decomposição de sinais finitos de uma dimensão consiste em processar um sinal c^0 e suas N amostras $c_0^0, c_1^0, ..., c_N^0 \in c^0$, sendo N uma potência de 2. Primeiramente aplica-se

Figura 8 – *Wavelet Symlet* de quarta ordem.



Fonte: Elaborado pelo autor.

um filtro passa baixa H em c^0, seguido de um redimensionamento (*downscale*) em fatores de 2, resultando em um sinal c^1 de tamanho $N/2$ que representa uma aproximação do sinal original. Posteriormente, aplica-se um filtro passa alta G no sinal c^0, seguido de um novo redimensionamento em fatores de 2, resultando em um sinal de detalhes d^1, com tamanho $N/2$.

Após essa etapa, o sinal c^1 é processado conforme o sinal c^0, resultando em um novo sinal de aproximação c^2 e um sinal de detalhe d^2, ambos de tamanho $N/4$. Esse processo é repetido M vezes até obter-se os sinais C^M e d^M de tamanho 1. Cada passo desse processo é conhecido como ciclo de decomposição. Os sinais de aproximação e de detalhe são obtidos conforme as Equações 1 e 2.

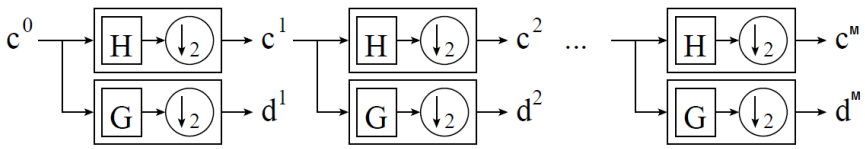
$$c_k^{m+1} = \sum_{i=1}^{M} h_i - 2k * c_i^m \tag{1}$$

$$d_k^{m+1} = \sum_{i=1}^{M} g_i - 2k * c_i^m \tag{2}$$

onde h_i e g_i são os coeficientes dos filtros H e G.

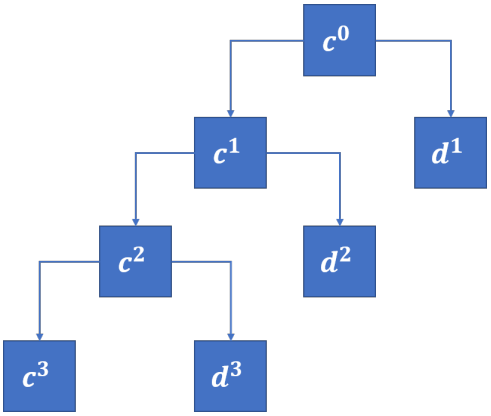
As Figuras 9 e 10 demonstram a execução dos ciclos de decomposição até se obter o sinal c^M e seus M sinais de detalhe $d^1, d^2, ..., d^M$.

Figura 9 – Transformada de Haar com M ciclos.



Fonte: Adaptado de Mello (2013)

Figura 10 – Árvore de decomposição.



Fonte: Adaptado de Mello (2013)

Para a reconstrução do sinal original, utiliza-se da transformada de *Haar* inversa, inicialmente aplicando um (*upscale*) de fator 2 nos sinais c^M e d^M, resultando no sinal c^{M-1}. Esse processo de síntese é repetido até se obter aproximadamente o sinal original c^0.

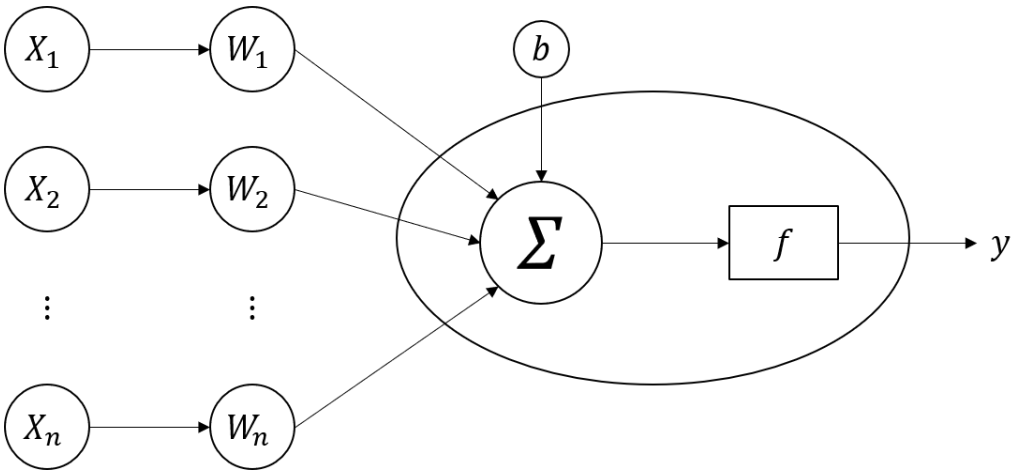
3.4 Redes Neurais Artificiais

As redes neurais artificiais (RNA) são técnicas computacionais baseadas em modelos matemáticos e são fortemente inspiradas por processos biológicos, mais especificamente, no funcionamento do cérebro humano, abstraindo os conceitos de neurônios biológicos, sinapses e axônios de forma a representar o aprendizado. Uma RNA pode ser definida como uma rede de unidades processadoras, denominadas neurônios artificiais, que possuem diversas interconexões conhecidas como sinapses artificiais (KRÖSE et al., 1993).

As RNA são utilizadas nos mais diversos campos da ciência da computação, para resolver problemas simples ou complexos, como reconhecimento de voz, detecção de objetos em imagens, reconhecimento de escrita, entre outras aplicações. A preferência pelo uso de RNA nessas aplicações vem da sua versatilidade, robustez, eficiência, e grande capacidade de generalização (HERTZ, 2018).

As RNA tem como unidade principal os neurônios, que são representados por um ou mais sinais de entrada, uma função matemática de processamento, também conhecido como função de ativação, e um valor de saída (KRÖSE et al., 1993; HAYKIN, 2007). A Figura 11 ilustra um neurônio artificial.

Figura 11 – Neurônio artificial.



Fonte: Elaborado pelo autor.

Esse modelo de rede neural com apenas um neurônio é conhecido como Rede *Perceptron* (WIDROW; LEHR, 1990; HAYKIN, 2007), e pode ser definido conforme a Equação 3.

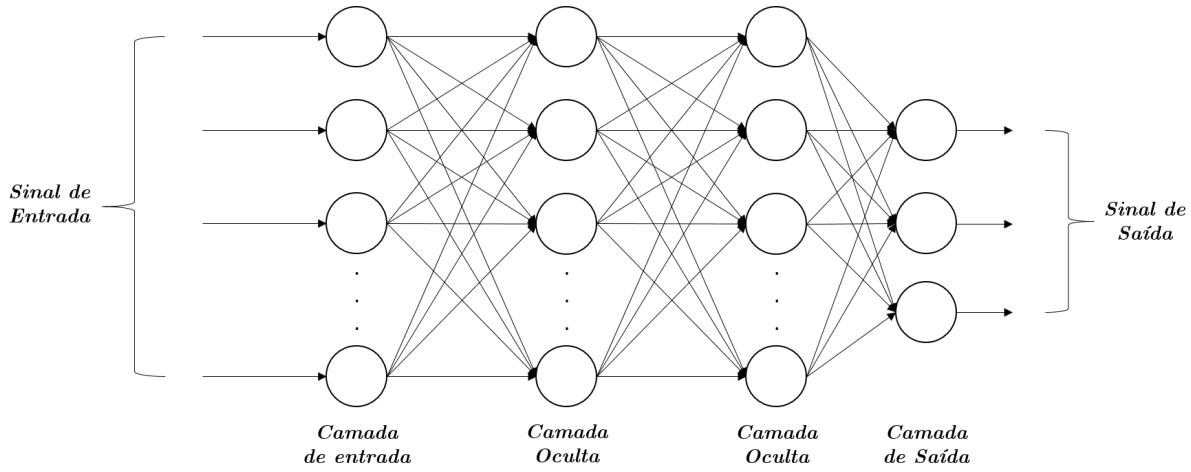
$$y = g(\sum_{i=1}^{n} x_i w_i + b) \tag{3}$$

onde $x_1, x_2, ...x_n$ são os sinais de entrada do neurônio, $w_1, w_2, ..., w_n$ são os pesos sinápticos, b é o termo bias, g é a função de ativação e y é a saída, ou resultado, gerado pelo neurônio.

O *Perceptron* é conhecido na literatura como sendo um classificador para problemas linearmente separáveis, o que torna essa rede muito limitada e incapaz de resolver diversos problemas do mundo real de forma satisfatória, entretanto é possível resolver esse problema adicionando mais neurônios e mudando a arquitetura da RNA (WIDROW; LEHR, 1990; HAYKIN, 2007).

Dentre as diversas arquiteturas de RNA existentes, a *Perceptron* de múltiplas camadas, do inglês *Multilayer Perceptron* (MLP), é uma das mais utilizadas pela literatura de redes neurais (KRÖSE et al., 1993; HAYKIN, 2007; HERTZ, 2018). Uma rede MLP é caracterizada pela presença de uma ou mais camadas intermediarias entre o sinal de entrada e o sinal de saída, denominadas camadas ocultas (HAYKIN, 2007). A presença da camada oculta possibilita à RNA extrair características mais significativas do sinal de entrada, se comparadas a redes mais simples (HAYKIN, 2007). A Figura 12 ilustra uma rede MLP com duas camadas ocultas.

Figura 12 – Rede MLP de duas camadas.



Fonte: Elaborado pelo autor.

O extenso uso das redes MLP se dá, em grande parte, em decorrência do algoritmo de correção de erros denominado *backpropagation*, introduzido por Rumelhart et al. (1987). Esse algoritmo consiste em comparar o sinal de saída gerado pela rede MLP com uma saída desejada a partir de valores de entrada conhecidos, e então propagar esse erro da camada de saída para as camadas ocultas até a camada de entrada da rede, modificando os pesos sinápticos das conexões entre as camadas na medida em que o erro é retropropagado. Podemos resumir esse algoritmo de aprendizagem conforme os passos abaixo:

1. Os pesos sinápticos são inicializados com valores aleatórios.
2. A rede MLP recebe os valores de entrada em um vetor de características $x_1, x_2, ...x_n$. Para esse vetor de entrada, deve-se especificar previamente qual a saída desejada em um vetor $d = d_1, d_2, ..., d_n$.
3. A rede produz um vetor de saída $y = y_1, y_2, ..., y_n$, calculados através da Equação 3.
4. É calculado o erro E da rede, comparando os vetores d e y, onde $E = d - y$.
5. Os pesos sinápticos são então reajustados, camada por camada, a começar pela camada de saída, através da Equação 4

$$w_ij(t+1) = w_ij(t) + \eta\delta_j * x_i \tag{4}$$

onde w_ij é o peso sináptico i do neurônio j em um determinado tempo t, x_i é um neurônio de entrada ou saída, δ_j é a denotação do erro da rede, e η é o valor arbitrário da taxa de aprendizagem da rede.

6. O algoritmo se repete a partir do passo 2 até que a rede produza uma saída satisfatória.

O calculo do erro na etapa de reajuste dos pesos se da através da Equação 5.

$$\delta_j(x) = \begin{cases} y_j(1-y_j)(d_j-y_j) & \text{se } j \text{ for um neurônio de saída} \\ x_j(1-x_j)\sum_k \delta_k w_{jk} & \text{se } j \text{ for um neurônio oculto} \end{cases} \tag{5}$$

onde d_j é a saída esperada, y_j é a saída real da rede, e k representa os neurônios acima do neurônio j.

Por fim, entende-se por taxa de aprendizagem η a amplitude das mudanças dos pesos sinápticos, influenciando diretamente no aprendizado da RNA, pois determina o quão rápido os pesos sinápticos são alterados.

Essa taxa normalmente encontra-se no intervalo de $[0, 1]$ e ao utilizar um valor pequeno de η, a rede sofre pequenas variações, o que torna o processo de treinamento mais lento e aumentando as chances da rede parar em mínimos locais. Enquanto ao se utilizar uma alta taxa η, a rede aprende mais rapidamente devido a maior variação dos pesos, porém aumentam as oscilações em torno do mínimo global (HAYKIN, 2007; SILVA; SPATTI; FLAUZINO, 2010).

3.5 Redes Geradoras Adversárias

As RNA tradicionalmente funcionam com base em um mapeamento de dados de entrada, como textos, imagens ou áudios, formados por características, que são utilizadas posteriormente como critérios de classificação de instâncias através da detecção de padrões nesses atributos por parte do classificador utilizado, sejam esses selecionadas e fornecidas na fase de treinamento (SILVA; SPATTI; FLAUZINO, 2010), ou extraídos pela própria rede neural, como acontece por exemplo com as Redes Neurais Convolucionais (LIANG et al., 2017; LECUN et al., 2010).

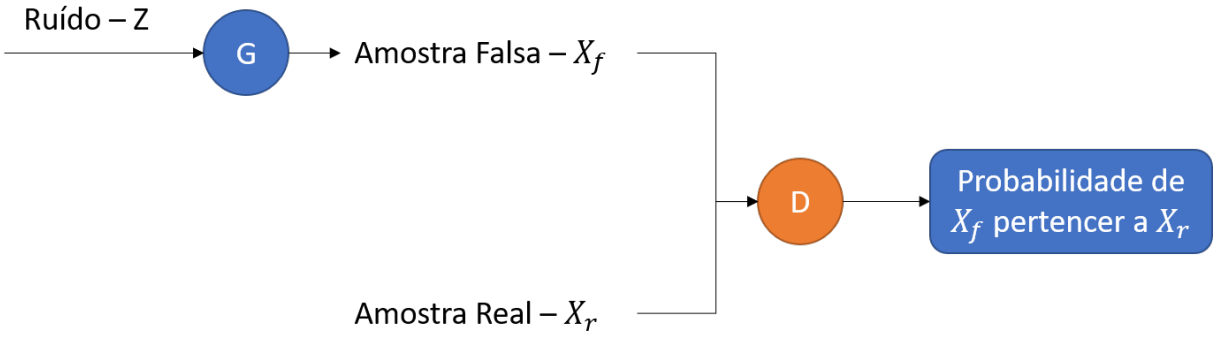
As Redes Geradoras Adversárias, do inglês *Generative Adversarial Networks* (GAN), por sua vez, são modelos generativos que também possuem a capacidade de extrair características automaticamente, e classificar novas instâncias, porém o seu principal diferencial reside na capacidade de gerar novas instâncias não conhecidas anteriormente com base nos padrões aprendidos na etapa de treinamento (GOODFELLOW et al., 2014).

A GAN foi introduzida no trabalho de Goodfellow et al. (2014), abordando um novo paradigma que ficou conhecido como treinamento adversário. Goodfellow descreveu a rede neural como um jogo entre dois jogadores. Um desses jogadores é denominado de gerador (G). Esse jogador é responsável por criar amostras que supostamente deveriam sair da base de treinamento. O segundo jogador é chamado de discriminante (D). Esse jogador tem a responsabilidade de analisar as amostras e determinar se elas são verdadeiras (pertencentes a base de treinamento) ou falsas (criadas por G).

O autor compara esse processo adversário com G sendo um falsificador de dinheiro, enquanto D seria a polícia, que tenta identificar quais cédulas são verdadeiras e quais são falsas. Dessa forma, ambos os agentes ficam continuamente aprimorando suas técnicas de falsificação e averiguação (GOODFELLOW et al., 2014).

Para ter sucesso nesse jogo, o falsificador de dinheiro precisa aprender a criar cédulas que são indistinguíveis das originais, ou seja, a rede G tem que aprender a criar amostras que são retiradas da mesma distribuição da base de treinamento, buscando sempre maximizar a probabilidade de D classificar as amostras artificiais como verdadeiras, a rede D por sua vez busca minimizar a probabilidade de uma amostra falsa ser erroneamente classificada como verdadeira (GOODFELLOW et al., 2014). A Figura 13 ilustra esse processo adversário.

Figura 13 – Redes geradora e discriminante.



Fonte: Elaborado pelo autor.

Pode-se definir uma GAN como um modelo probabilístico estruturado, que mapeia variáveis latentes z de uma distribuição Z para amostras x de uma outra distribuição X, que é uma das amostras de treino, como áudios, imagens, etc. O componente responsável por esse mapeamento é o gerador, e o seu aprendizado é feito através do treinamento

adversário em conjunto com o discriminante (GOODFELLOW et al., 2014; PASCUAL; BONAFONTE; SERRÀ, 2017).

Considerando ambas as redes G e D como funções diferenciáveis, podemos dizer que ao utilizar uma amostra $z \in Z$, a função $G(z)$ tem o objetivo de produzir como resultado uma amostra $\hat{x} \in X$. Enquanto a função $D(\hat{x})$ deve retornar a probabilidade de $\hat{x}$ pertencer a X. Em outras palavras, a saída do discriminante para $D(x)$ deve ser próxima de 1, enquanto a sua saída para $D(G(z))$ deve ser próxima de 0 (GOODFELLOW, 2016).

Esse jogo é descrito como uma função min max conforme a Equação 6.

$$\min_G \max_D V(D,G) = E_{xp_{data}(x)}[logD(x)] + E_{zp_z(z)}[log(1 - D(G(z)))] \tag{6}$$

onde p_z é a distribuição de amostras de entrada, G é a rede geradora, D é a rede discriminante. $D(x)$ representa a probabilidade da amostra ser real, e esse valor busca ser maximizado, enquanto $\log(1 - D(G(z)))$ busca ser minimizado.

Esse processo de treinamento pode ser resumido conforme os passos abaixo.

1. A rede D, usualmente um classificador binário, recebe um conjunto de amostras reais e aprende a classificá-las como verdadeiras (1) através de *backpropagation*;
2. A rede D recebe um novo conjunto de amostras, agora gerados pela rede G, e aprende a classificá-las como falsas (0);
3. O treino da rede D é suspenso, e a rede G aprende a gerar saídas que D classifica como verdadeiras.

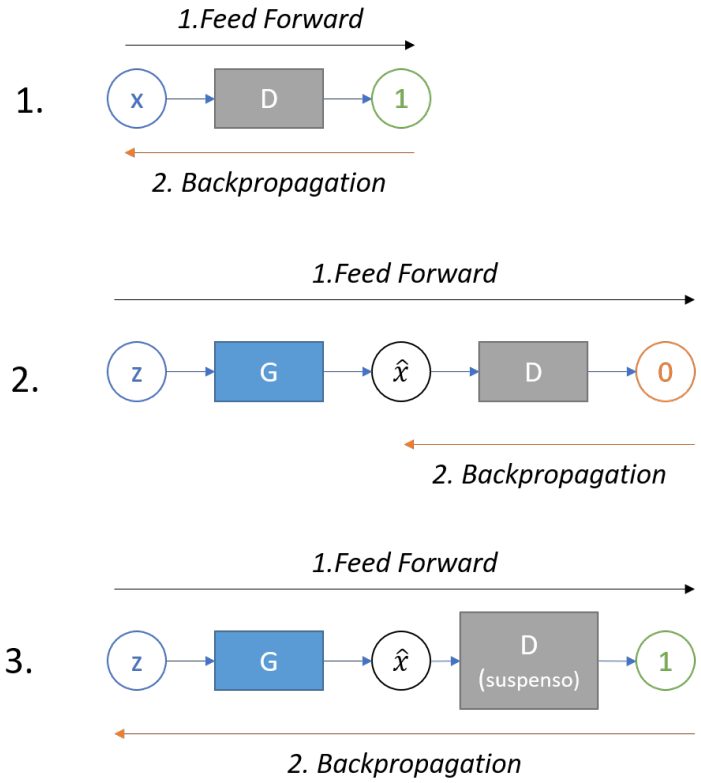
Por fim, a Figura 14 ilustra o treinamento da rede.

3.5.1 SEGAN

Segundo Pascual, Bonafonte e Serrà (2017), o processo de melhoramento da qualidade de um sinal é definido de forma que existe um sinal ruidoso $\tilde{x}$ que deve ser aprimorado de forma a se obter um sinal melhorado $\hat{x}$. Nesse contexto, foi apresentada a rede adversária para melhoramento de fala, do inglês *Speech Enhancement Generative Adversarial Network* (SEGAN).

A SEGAN é uma GAN especializada em aprender as características de áudios sem a presença de ruídos, podendo então gerar sinais limpos a partir de amostras ruidosas. Sendo

Figura 14 – Processo de treinamento da GAN.



Fonte: Elaborado pelo autor.

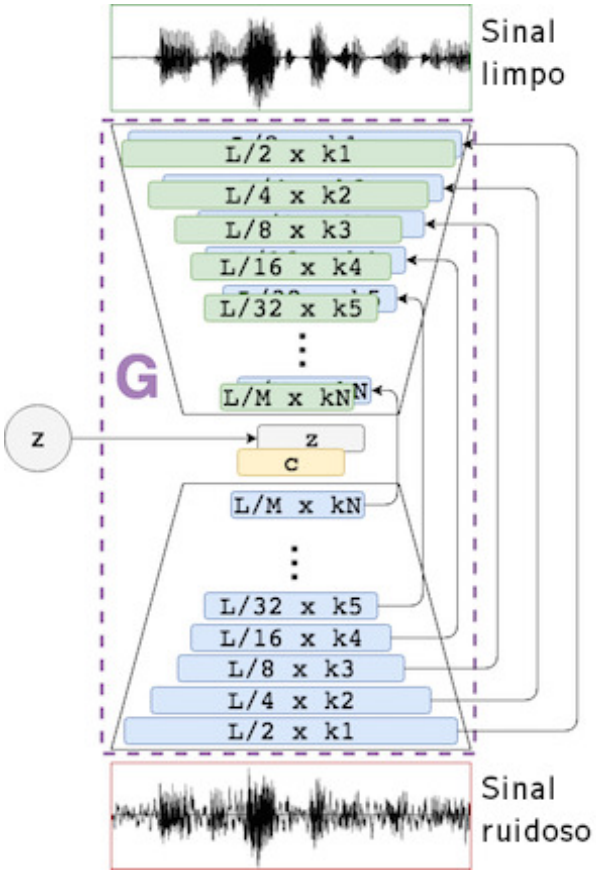
então a rede G responsável por calcular a função $G(\tilde{x}) = \hat{x}$ (PASCUAL; BONAFONTE; SERRÀ, 2017). A arquitetura dessa rede é ilustrada conforme a Figura 15.

O autor atenta para o fato de que a rede G da SEGAN possui estrutura similar à das redes *Autoencoders* (LIOU et al., 2014), pois possui um estágio de codificação e subamostragem, seguido de um segundo estágio de decodificação. Porém, ao contrário dos *Autoencoders* que possuem uma conexão direta entre todas as camadas ocultas da rede de forma sequencial, a rede G pula essa conexão e liga as camadas ocultas do estágio de codificação diretamente com as camadas ocultas análogas do estágio de decodificação (PASCUAL; BONAFONTE; SERRÀ, 2017).

Outra característica importante da rede G da SEGAN vem de sua estrutura ponto-a-ponto, ou seja, ela processa o sinal de áudio diretamente, efetivamente descartando a necessidade de realizar transformações intermediárias ou extrair características manualmente (PASCUAL; BONAFONTE; SERRÀ, 2017).

O processo de treinamento da SEGAN segue conforme a Figura 16, onde as amostras de treino são fornecidas em pares, em duas combinações de pares diferentes. O primeiro par, conhecido como par real, é composto pelo sinal original limpo e por um sinal ruidoso

Figura 15 – Arquitetura da rede G de uma SEGAN.



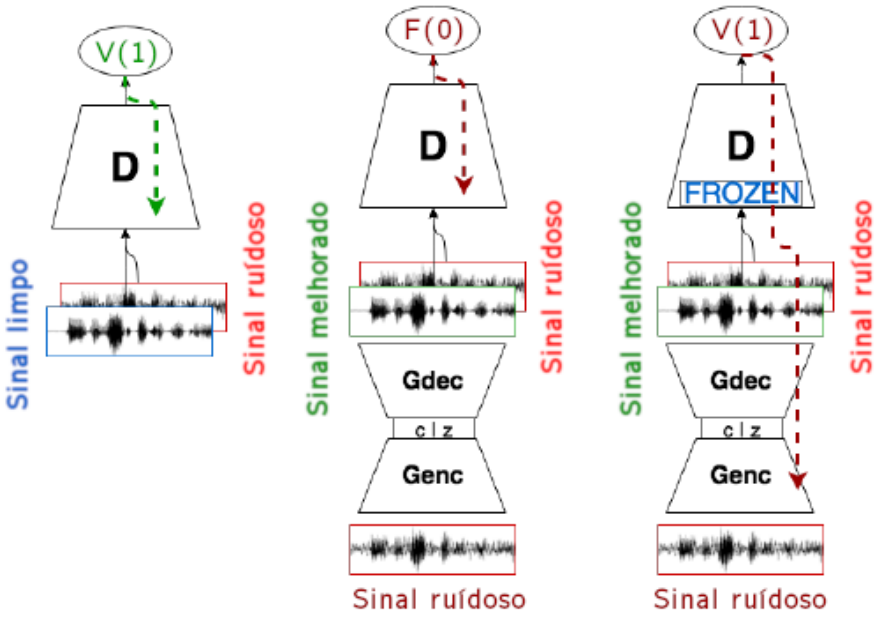
Fonte: Traduzido de Pascual, Bonafonte e Serrà (2017).

(x e $\tilde{x}$). O segundo par é denominado par falso, sendo composto por um sinal melhorado e por um sinal ruidoso ($\hat{x}$ e $\tilde{x}$).

Esse treinamento funciona de forma análoga à de uma GAN tradicional, onde inicialmente a rede D é treinada para classificar corretamente as amostras verdadeiras, seguido de um treino para classificar corretamente as amostras falsas, e por fim, treinando a rede G para conseguir gerar amostras que sejam classificadas pela rede D como verdadeiras (PASCUAL; BONAFONTE; SERRÀ, 2017).

Em termos práticos, a rede G recebe um sinal de áudio ruidoso como entrada, e modifica diretamente o formato da onda de som de forma a eliminar o ruído presente. Em seguida, esse sinal gerado por G é fornecido como entrada para D classificar como verdadeiro ou falso, tendo os erros retropropagados para que a rede G consiga cada vez mais gerar sinais de maior qualidade (PASCUAL; BONAFONTE; SERRÀ, 2017). Após essa etapa de treinamento, desconecta-se a rede D da arquitetura, a fim da SEGAN ter como saída a onda de áudio melhorado pela rede G.

Figura 16 – Treinamento adversário na SEGAN.



Fonte: Traduzido de Pascual, Bonafonte e Serrà (2017).

3.6 Estudo duplo-cego

Os estudos abertos e estudos cegos, são de forma geral, duas maneiras de se conduzir e analisar resultados de experimentos científicos. Segundo Robertson e Kesselheim (2016) em um estudo aberto, os testes são realizados de forma que a identidade dos autores, revisores, métodos aplicados e sujeitos de teste são divulgadas abertamente. Por sua vez, no estudo cego os experimentos são realizados de maneira que os sujeitos de teste, quais métodos são empregados, e o autor dos experimentos se desconhecem.

O estudo cego, em especifico o duplo-cego, do inglês *Double-blind trial*, consiste na realização de experimentos onde o autor do experimento, e os sujeitos de teste estão alheios aos grupos de controle a qual eles pertencem. Em alguns casos, ambos desconhecem quais métodos são empregados no estudo, assim como as identidades dos envolvidos. Para Robertson e Kesselheim (2016), esta é uma forma de evitar resultados tendenciosos por parte dos sujeitos de teste e dos condutores do experimento.

Outro fator de suma importância nesse tipo de estudo é a aleatorização do processo, onde os sujeitos de teste são escolhidos de forma aleatória, assim como quais métodos serão apresentados para cada sujeito, ou em que ordem essas são apresentadas. A identidade dos sujeitos, assim como dos métodos e a sua ordem de aplicação são mantidos por terceiros

ao estudo, e são divulgados aos pesquisadores somente após o fim da experimentação (ROBERTSON; KESSELHEIM, 2016).

Essa metodologia de aplicação de testes foi utilizado pela primeira vez em 1907 pelo psiquiatra W. H. R. Rivers, enquanto conduzia um estudo que analisava os efeitos da cafeína no organismo humano (RIVERS; WEBBER, 1907), e desde então tem sido utilizado nas mais diversas áreas da ciência como meio de aumentar a confiabilidade dos resultados obtidos nos ensaios clínicos, revisões e experimentações (ROBERTSON; KESSELHEIM, 2016).

Como exemplo, no contexto de avaliação de submissões de artigos científicos, ao aplicar-se o estudo duplo-cego, o autor do artigo desconhece quem são os revisores, assim como os revisores desconhecem quem é o autor do trabalho, diminuindo as chances de um revisor favorecer ou prejudicar um autor em decorrência de diversas influências do consciente humano, como preconcepções ou conflitos de interesse (ROBERTSON; KESSELHEIM, 2016).

4 Remoção de ruídos aditivos e segmentação de palavras-chave em áudios

Neste capítulo é apresentada a metodologia proposta neste trabalho. Primeiramente são apresentadas as bases de áudio utilizadas e posteriormente são introduzidas os métodos de remoção de ruídos e de segmentação de áudio, sendo detalhados os seus passos.

4.1 Bases e amostras de áudio

Nesse trabalho foram utilizadas duas bases de áudios, sendo ambas públicas e um conjunto de amostras de áudio compilada pelo autor deste trabalho, cada uma delas sendo utilizada para etapas diferentes da metodologia aplicada. A configuração dessas bases é exposta nas subseções abaixo.

4.1.1 Base DEMAND

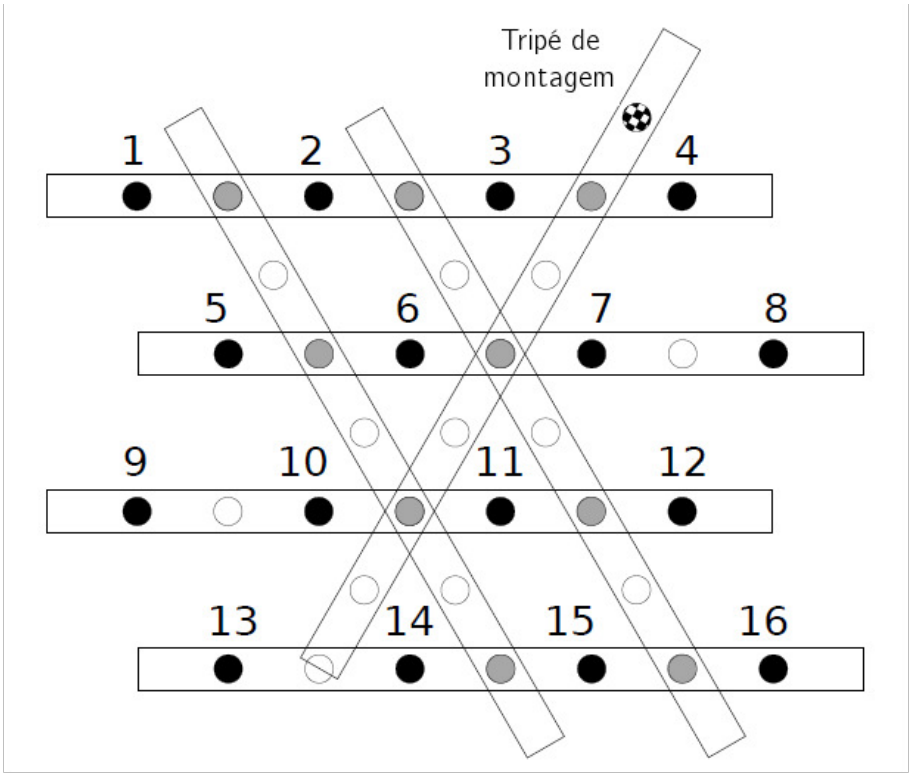
A base DEMAND, do inglês *Diverse Environments Multichannel Acoustic Noise Database*, é uma coleção de arquivos de gravações de áudio disponibilizada gratuitamente na internet e foi criada com o intuito de facilitar testes de algoritmos de remoção de ruídos (THIEMANN; ITO; VINCENT, 2013).

Os áudios da base DEMAND foram gravados em diversos ambientes utilizando um conjunto de 16 microfones distanciados entre si de 5 cm a 21,8 cm, sempre paralelos ao solo, e foram esquematizados conforme a Figura 17.

As gravações foram feitas com a taxa de amostragem de $48kHz$ durante longos períodos de tempo. Os áudios tiveram os primeiros e últimos 300 segundos removidos propositalmente, de forma a evitar ruídos provenientes da instalação e desinstalação do conjunto de microfones. Os áudios também estão disponibilizados com a taxa de amostragem em $16kHz$ e todos encontram-se no formato ".wav".

A base DEMAND está dividida em seis categorias, onde cada uma representa um ambiente diferente. Dentre essas, quatro são ambientes internos e dois são ambientes externos. Os ambientes internos são classificados como ambiente doméstico, escritório, público e transportes. Os ambientes externos são classificados como natureza e ruas. Cada uma dessas categorias possui três localizações distintas onde foram realizadas as gravações de áudio.

Figura 17 – Visão esquemática do conjunto de microfones. Os círculos pretos indicam a posição dos microfones, os círculos cinzas indicam os parafusos de conexão entre as barras, círculos brancos são buracos não utilizados nas barras de montagem. Cada número representa o canal de cada microfone.



Fonte: Traduzido de Thiemann, Ito e Vincent (2013).

4.1.2 Base de Valentini

A base publicada por Valentini-Botinhao et al. (2016) é constituída por áudios de 28 locutores distintos, todos residentes na Inglaterra, sendo 14 homens e 14 mulheres. Essa base possui aproximadamente 400 sentenças gravadas para cada locutor, além de suas respectivas transcrições, e pode ser acessada gratuitamente[1]. Além disso, cada áudio possui sua versão paralela com ruídos adicionados artificialmente.

Os ruídos utilizados nessa base foram extraídos da base DEMAND (THIEMANN; ITO; VINCENT, 2013). As categorias de ruído escolhidas foram: Ruídos domésticos (barulhos de cozinha); sons de escritório (sala de conferência); ruídos de espaços públicos (café, restaurante, metrô); ruídos de meios de transporte (carros e trens); e ruídos de trânsito (VALENTINI-BOTINHAO et al., 2016).

Todos os áudios dessa base foram gravados com a taxa de amostragem de $48kHz$ e no formato ".wav", tendo sido normalizados após a gravação. Também foram removidos

[1] https://datashare.is.ed.ac.uk/handle/10283/2791

trechos de silêncio superior a 200 milissegundos do início e do final de cada arquivo de áudio (VALENTINI-BOTINHAO et al., 2016).

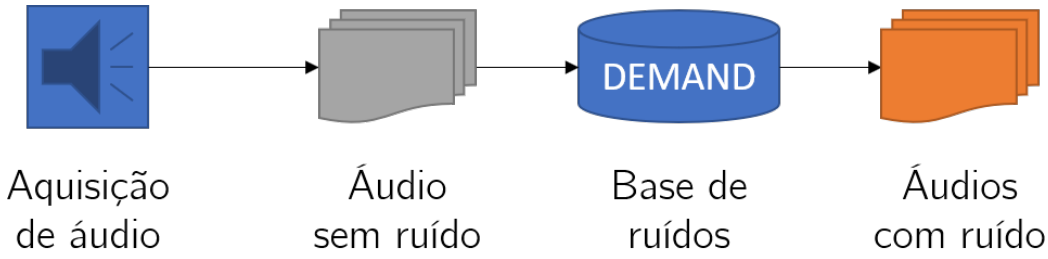
4.1.3 Amostras de áudio para remoção de ruído e segmentação

Para avaliar o desempenho dos métodos de remoção de ruído, são necessárias amostras de áudios que contenha ambos os áudios com ruído e livres de ruído. Para tal, foi feita a gravação de 192 frases de acordo com o roteiro presente no Anexo C. A gravação foi realizada na Universidade Federal do Maranhão, dentro do laboratório VIP, e contou com a presença de 9 locutores, sendo 6 homens e 3 mulheres. Cada locutor gravou as 192 frases presentes no roteiro, em uma média de 20 minutos por seção de gravação para cada locutor, resultando em uma coleção com 1728 sentenças livre de ruídos, totalizando aproximadamente de duas horas de áudio contínuo.

Posteriormente, foram selecionados cinco tipos de ruído distintos, 2 gerados artificialmente e 3 da base DEMAND. As categorias de ruído escolhidas da base DEMAND para a montagem dessa coleção de áudios são diferentes das que estão presentes na base de Valentini-Botinhao et al. (2016). Sendo essas categorias: Ruídos da floresta; Ruídos de um parque público; e ruídos gerados pela água corrente de um rio.

Esses ruídos foram inseridos artificialmente em cada arquivo de áudio com as 192 sentenças, acrescentando mais 1728 sentenças, agora com ruídos diversos, à coleção de amostras de áudios. Esse processo é ilustrado pela Figura 18.

Figura 18 – Processo de aquisição de áudios e geração de amostras de áudios.

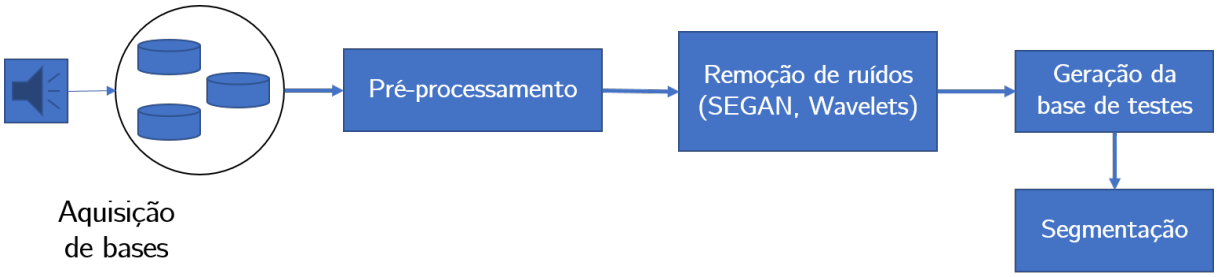


Fonte: Elaborado pelo autor.

4.2 Metodologia proposta

Esta seção apresenta a metodologia proposta para a remoção de ruídos em áudios e a segmentação automatizada de BC. A sequência de passos dessa metodologia é feita de acordo com a Figura 19, e cada passo é descrito em detalhes nas subseções seguintes.

Figura 19 – Diagrama da metodologia proposta.



Fonte: Elaborado pelo autor.

4.2.1 Aquisição de bases

Segundo ChiÑu e Rothkrantz (2007), o processo de montagem de uma boa base de áudios demanda muito tempo e esforço. Em virtude disso, essa primeira etapa consistiu em fazer um levantamento de bases de áudio relevantes ao contexto dessa pesquisa, buscando as que mais se adequassem às necessidades de treino da SEGAN. Esse processo resultou na adoção da base de Valentini, devido à quantidade de arquivos de áudio transcritos, com versões limpas e ruidosas. Outro fator decisivo na adoção dessa base foi a sua disponibilização gratuita através da internet.

Entretanto, para o contexto desta pesquisa, existia a necessidade de realizar testes utilizando amostras de áudios com locutores nativos do português brasileiro. Portanto, foi feita a montagem de uma coleção de áudios de testes para a remoção de ruídos e segmentação, conforme descrito na Subseção 4.1.3.

4.2.2 Pré-processamento

Nessa etapa, o primeiro passo consiste em normalizar a amplitude do áudio através da Equação 7, de forma a deixar o volume mais uniforme em todo o áudio.

$$Norm(A_i) = \frac{A_i}{max(abs(A))} \quad (1 \leq i \leq N) \tag{7}$$

onde A é o vetor com as amostras do sinal de áudio, A_i é a *i-ésima* amostra do sinal, $max()$ é a função que retorna o maior valor do vetor, e $abs(A)$ é a função que retorna os valores absolutos do vetor A.

Em seguida, o áudio tem sua taxa de amostragem reduzida para $16kHz$, pois essa configuração fornece informações suficientes para extrair boas características do espectro de áudio, reduzindo também o seu custo computacional (HYOUNG-GOOK; MOREAU; SIKORA, 2005; COOK, 2002; RAMANA; LAXMINARAYANA; MYTHILISHARAN, 2012).

4.2.3 Remoção de ruídos

Para realizar a etapa de remoção de ruídos, foi treinada uma rede SEGAN utilizando a base de Valentini como base de treino. A rede foi configurada para rodar durante 150 épocas, com a taxa de aprendizado de $0,0002$ para ambas as redes G e D. Foram utilizadas amostras de áudio com 70 segundos de duração. O uso desse tamanho de amostra pequeno ocorreu em decorrência da limitação da memória da placa gráfica, e valores superiores a 70 segundos resultavam em um estouro de pilha.

O processo de treinamento da SEGAN durou aproximadamente 72 horas, e o modelo resultante ocupa por volta de 1,5 GB de espaço em disco. Após essa etapa de treinamento, desconectou-se a rede D da SEGAN, para que fosse possível realizar os testes de remoção de ruído. A SEGAN treinada levou aproximadamente 50 segundos para realizar a remoção de ruídos para cada minuto de áudio.

Adicionalmente foram desenvolvidas outras duas abordagens de redução de ruído multivariada com *wavelets*, a critério de comparação com o método proposto, por serem amplamente adotadas na literatura (WALKER, 2002; BAHOURA; ROUAT, 2006; YU; MALLAT; BACRY, 2008). O desenvolvimento desses métodos foi feito de acordo com

Aminghafari, Cheze e Poggi (2006), sendo utilizadas as *wavelets* de *Haar* e a *wavelet symlet* de quarta ordem.

4.2.4 Segmentação

O método proposto para a segmentação de áudio, em específico de BC, funciona através da detecção de trechos de silêncio e picos de energia no sinal de áudio. Portanto, antes de realizar o recorte dos segmentos de áudio, foi realizada a aplicação de uma função de atenuação, conforme a Equação 8. Posteriormente, foi aplicada a Equação 9 para detectar os picos de amplitude e silenciar os vales, com base no desvio padrão do sinal de áudio e seu escore padronizado (PUKELSHEIM, 1994).

$$S(A_i) = A_i * \left(\frac{|A_i|}{|max(A)|} \right) \quad (1 \leq i \leq N) \tag{8}$$

onde $S(A_i)$ é a função de atenuação, N é o tamanho do vetor de amostras, A_i é a *i-ésima* amostra do sinal de áudio e $max(A)$ é o maior valor encontrado no conjunto de amostras A.

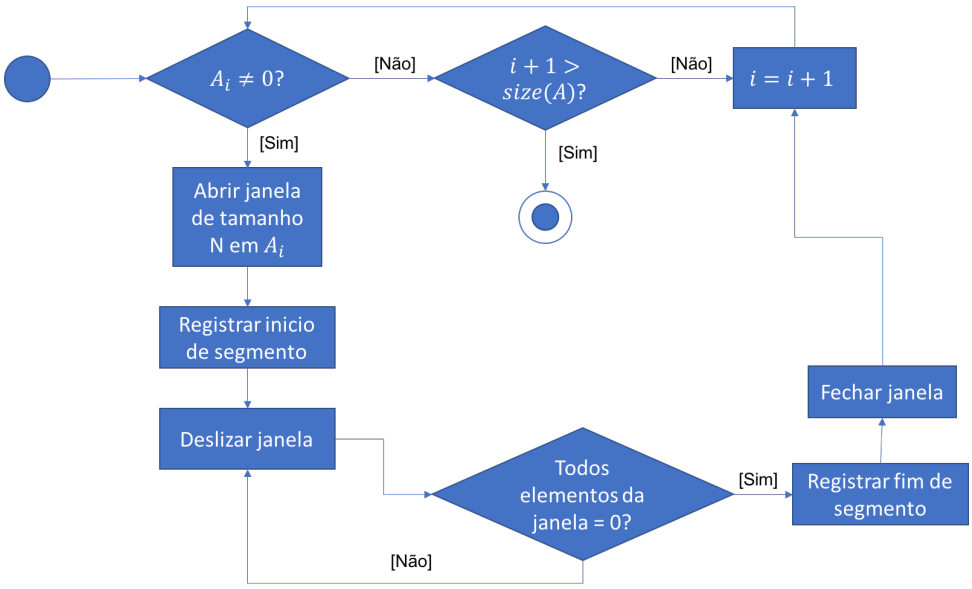
$$Z(A_i) = \begin{cases} A_i, & \text{se } \left| A_i \left(\frac{1}{N} \sum_{i=1}^{N} A_i \right) \right| \geq \frac{SD(A)}{4} \quad (1 \leq i \leq N) \\ zero, & \text{caso contrário} \end{cases} \tag{9}$$

onde $Z(A_i)$ é a função de silenciamento, $SD(A)$ é a função de desvio padrão, N é o tamanho do vetor de amostras e A_i é a *i-ésima* amostra do sinal de áudio.

Após essa etapa, é aplicada uma função de mapeamento conforme ilustrado pela Figura 20. Essa função irá identificar os segmentos de áudio e suas respectivas durações com base em uma janela deslizante com sobreposições, e nos silêncios gerados nos passos anteriores.

O algoritmo inicia verificando amostra por amostra pela primeira ocorrência diferente de zero, e ao encontrá-la, inicia uma função de janelamento a partir da amostra A_i e com o tamanho N. Nesse momento, é registrado o inicio do segmento de áudio.

Figura 20 – Diagrama do algoritmo de mapeamento.



Fonte: Elaborado pelo autor.

A janela desliza até que todos os seus elementos sejam zero, ou seja, até encontrar um segmento de silêncio. A partir desse instante é registrado o fim do segmento de áudio e sua duração é calculada a partir da amostra inicial e final.

Posteriormente, o processo se repete até que a última amostra do sinal A seja analisada. Por fim, utiliza-se o mapa gerado pelo algoritmo para realizar os recortes dos segmentos no sinal de áudio original, resultando em um segmento com um BC.

5 Experimentos

Neste capítulo são apresentados os materiais e métricas utilizados neste trabalho. Primeiramente são apresentadas as configurações das máquinas onde foram executados os testes da metodologia. Posteriormente são apresentadas as coleções de arquivos de áudio utilizados para testar a metodologia. Por fim, são introduzidas as métricas de desempenho utilizadas para avaliar a metodologia proposta neste trabalho.

5.1 Ferramentas

O computador utilizado para implementar e testar os métodos propostos possui a seguinte configuração: Processador Intel Core i5-4440 Quad-Core; 16GB de memória RAM; Placa gráfica Nvidia GeForce 1060ti com 6GB de memória dedicada; Sistema operacional Windows 10.

A linguagem de programação utilizada foi Python (versão 3.6.2). Também foi utilizada a API do CUDA (Versão 9.0), em conjunto com a biblioteca cuDNN (versão 7.0) (CHETLUR et al., 2014), ambos como pré-requisitos para instalar e configurar o Tensorflow (versão 1.4.0) (ABADI et al., 2016). A implementação da SEGAN foi feita utilizando o Tensorflow, e foi configurado de forma a rodar as etapas de treino e teste utilizando a placa gráfica do sistema.

A gravação dos arquivos de áudio foram realizadas através do gravador de áudio padrão do Windows. Os recortes de áudio, adição de ruídos, marcação de segmentação e algumas etapas do pré-processamento foram feitas utilizando a ferramenta *Audacity* (versão 2.0) (TEAM, 2012).

5.2 Geração de amostras de teste

Para a realização dos testes de remoção de ruído e segmentação, foram montadas duas coleções de áudio para serem utilizados com a metodologia proposta, utilizando como referência a coleção descrita na Seção 4.1.3. O processo de geração dessas duas coleções de teste é descrito a seguir.

5.2.0.1 Amostras de teste de remoção de ruídos

No contexto dessa pesquisa, foram feitos testes de percepção como forma de validar o desempenho das técnicas de remoção de ruído empregadas nos arquivos de áudio. Para tal, foram selecionadas 15 frases entre as 192 presentes na coleção de amostras de teste descrita na Seção 4.1.3, extraindo também as suas respectivas versões com ruído. Essa seleção resultou em um arquivo com 45 segundos de áudio limpo, e outro com 45 segundos de áudio ruidoso. A duração desses áudios equivale à soma da duração das 15 frases selecionadas ao serem reproduzidas sequencialmente. Esses dois arquivos são denominados de áudios de controle.

A seleção das frases foi feita através de sorteio, levando em consideração as classes de ruído, resultando então em 3 frases para cada tipo de ruído. A Tabela 1 mostra quais frases foram selecionadas, assim como o tipo de ruído presente em cada uma.

Tabela 1 – Transcrição dos áudios utilizados no teste de eliminação de ruídos.

#	Frase	Tipo de ruído
1	A escalada é tudo que existe	Branco
2	Eu estou pensando em você	Branco
3	Preciso ir buscar o carro agora	Branco
4	Todo o poder emana do povo	Conversa de fundo
5	Vivo na liberdade do dia a dia	Conversa de fundo
6	Queremos uma revolução em nós mesmos	Conversa de fundo
7	Qual é a data do seu nascimento?	Ruídos da floresta
8	Eu quero ser a rainha	Ruídos da floresta
9	Em que ano estamos?	Ruídos da floresta
10	Robu	Ruídos de parque público
11	A menina não o produzia com frequência	Ruídos de parque público
12	Você parece surpresa	Ruídos de parque público
13	Vaisol	Ruídos de rio
14	Estou numa montanha-russa que só vai para cima	Ruídos de rio
15	O porto é o lugar mais seguro para um barco	Ruídos de rio

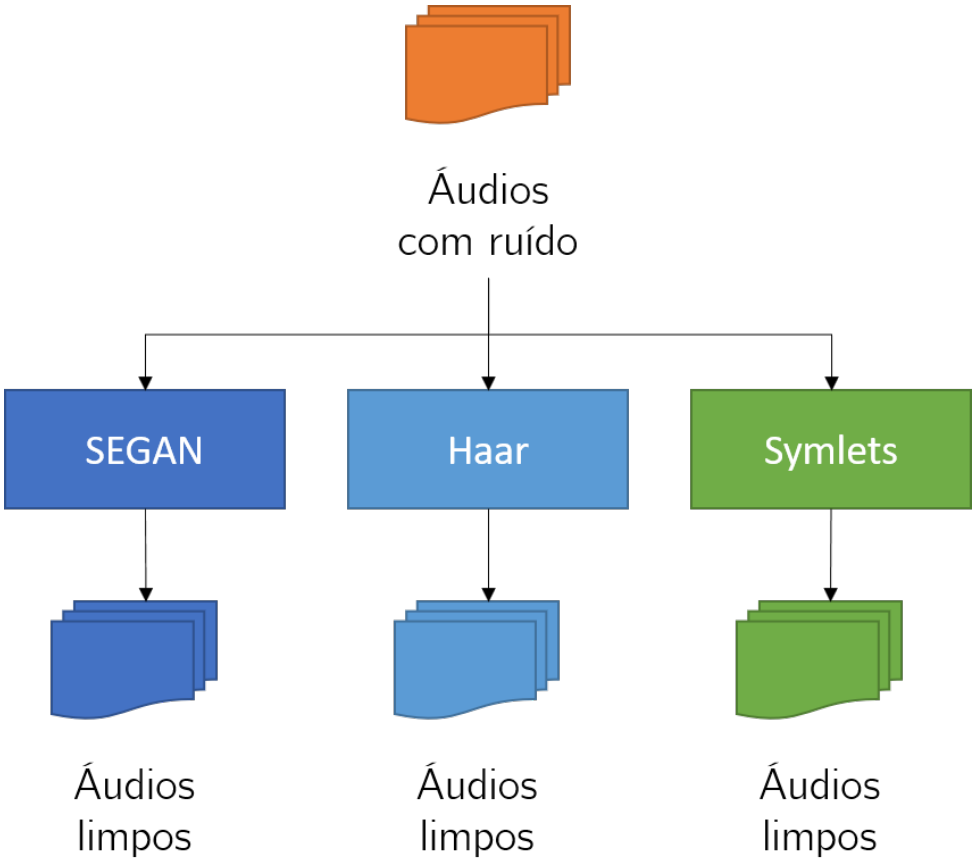
Fonte: Elaborado pelo autor.

Posteriormente, foram aplicadas três técnicas distintas de remoção de ruído, sendo elas: redução de ruído multivariada com *wavelets*, usando a *wavelet* de *Haar*, e a *symlet* de quarta ordem; e o método proposto, com o melhoramento de sinal através da rede SEGAN.

Esse passo resulta em mais três arquivos de áudio com os mesmos 45 segundos de duração dos áudio originais, conforme ilustrado pela Figura 21. Esses arquivos de áudio,

em conjunto com os dois arquivos de controle, compõem a coleção de amostras de testes para a remoção de ruídos.

Figura 21 – Diagrama da produção da coleção amostras de testes.



Fonte: Elaborado pelo autor.

5.2.0.2 Amostras de teste para segmentação

Para a realização dos testes de segmentação, foram extraídos da coleção de amostras descrita na Seção 4.1.3 os trechos correspondentes aos 33 códigos da BCG. Esse processo resultou em uma nova coleção de amostras de testes contendo 9 arquivos de áudio sem ruído e suas respectivas versões com ruído. Cada arquivo representa um locutor individual e possui uma média de duração de 1 minuto cada. O tempo de cada arquivo de áudio nessa coleção de amostras de testes é diretamente relacionado ao tempo que cada locutor levou para pronunciar os 33 BC.

As versões ruidosas dos áudios passaram então pelo mesmo processo de remoção de ruídos ilustrado pela Figura 21, gerando mais 27 arquivos de áudio. Essa coleção de

amostras de teste de segmentação é composta por 45 arquivos de áudio, em um total de 48 minutos de áudio corrido.

5.3 Métricas de Validação dos Resultados

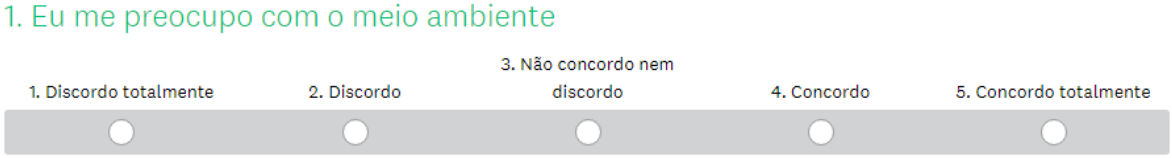
Nessa seção são apresentadas as métricas de desempenho utilizadas para avaliar os resultados gerados pela metodologia proposta e as que podem ser utilizadas para avaliar os trabalhos futuros.

5.3.1 Escala de Likert

A escala de *Likert* é um tipo de escala psicométrica[1] por intervalo, normalmente utilizada em questionários de pesquisa de opinião. Essa escala consiste em medir uma determinada característica com base em um intervalo numérico equidistante, onde cada item na escala representa o grau de concordância do entrevistado em relação à alguma pergunta ou afirmação sobre a característica pesquisada. O nome dessa escala é dado devido ao seu autor e criador, o psicólogo Renis Likert (LIKERT, 1932; MALHOTRA, 2012).

Essa escala possui 5 ou mais pontos, onde o ponto central representa um valor neutro, enquanto os demais pontos podem representar qualidades positivas ou negativas. Por exemplo, dada a afirmação "Eu me preocupo com o meio ambiente", na escala o primeiro valor representa "Discordo totalmente", o valor intermediário, ou neutro, representa "Não concordo, nem discordo", enquanto o último valor representa "Concordo totalmente". Esse formato é conhecido como Escala de *Likert* bipolar (LIKERT, 1932). A Figura 22 ilustra essa afirmação utilizada em um questionário.

Figura 22 – Exemplo de afirmação com escala de *Likert* bipolar.

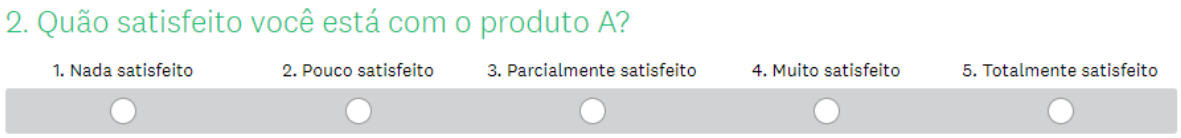


Fonte: Elaborado pelo autor.

[1] Entende-se por psicometria as medidas associadas a processos psicológicos, como percepção e opinião (PASQUALI, 2009).

Também podemos expressar a escala de forma unipolar, ou seja, ao invés de dividir a característica em pontos positivos e negativos, ela pode medir a ausência ou presença de uma determinada qualidade, com intervalos intermediários crescentes. Por exemplo, dada a pergunta "Quão satisfeito você está com o produto A?", o primeiro valor na escala corresponderia a "Nada satisfeito", um valor intermediário representaria "Parcialmente satisfeito", enquanto o último valor representaria "Totalmente satisfeito"(MALHOTRA, 2012). A Figura 23 ilustra essa pergunta utilizando a escala unipolar.

Figura 23 – Exemplo de pergunta com escala de *Likert* unipolar.



Fonte: Elaborado pelo autor.

5.3.2 Desempenho de segmentação de áudio

No contexto de segmentação de áudio, os testes de desempenho envolvem medir a detecção dos segmentos que contém as informações relevantes, como frases, palavras, ou fonemas, de forma a avaliar quão bem a técnica aplicada consegue discriminar a presença ou ausência desses segmentos. No contexto deste trabalho, considera-se informação relevante um BC individual. Dito isso, pode-se identificar as seguintes situações possíveis:

- O segmento extraído pela técnica contém apenas um BC, ou seja, é uma das respostas esperadas - Segmento Corretamente Identificado (SCI);
- O segmento extraído pela técnica não contém BC, ou contém mais de um BC, em outras palavras, não pertence ao grupo de respostas esperadas - Segmento Incorretamente Identificado (SII);
- O número total de segmentos encontrados pela técnica - Segmentos Detectados (SD), onde $SD = SCI + SII$;

Com base nessas três situações, utiliza-se o Coeficiente de Similaridade de *Dice* (CSD), do inglês *Dice Score* (DICE, 1945; HYOUNG-GOOK; MOREAU; SIKORA, 2005) como métrica de desempenho para validação do método de segmentação proposto. O CSD

é amplamente utilizado na literatura para avaliar os métodos de segmentação de áudio e imagem, onde a saída do método avaliado é comparado com a marcação manual feita por um especialista(ZIJDENBOS et al., 1994; HYOUNG-GOOK; MOREAU; SIKORA, 2005), e é definido conforme a Equação 10.

$$CSD = \frac{2 * PRE * SEN}{PRE + SEN} \tag{10}$$

onde PRE e SEN são, respectivamente, a Precisão e a Sensibilidade do método.

A Precisão de um sistema de segmentação de áudio pode ser definido conforme a Equação 11, enquanto a Sensibilidade é definida pela Equação 12.

$$PRE = \frac{SCI}{SD} \tag{11}$$

$$SEN = \frac{SCI}{N} \tag{12}$$

onde N é o número total de segmentos a serem identificados no áudio.

O *Dice Score* é delimitado entre os valores $[0, 100]$, onde $CSD = 100$ representa um processo de segmentação perfeito, enquanto $CSD = 0$ implica que o processo de segmentação está completamente errado.

Ainda nas métricas de desempenho de segmentação, a Precisão indica a capacidade do método de retornar segmentos relevantes ao contexto, enquanto a Sensibilidade diz respeito a proporção em que o método retorna segmentos relevantes dentre os segmentos existentes no áudio.

6 Resultados e discussões

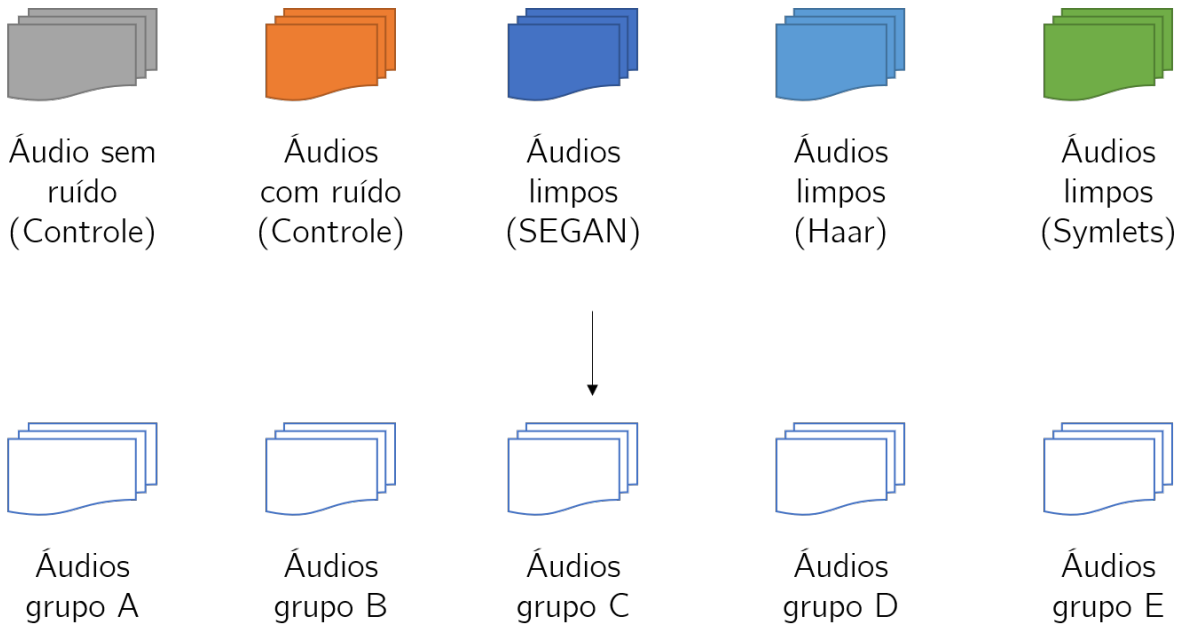
Neste capítulo, serão apresentados os modelos gerados na fase de treinamento da metodologia proposta e os resultados alcançados na fase de teste a partir destes modelos.

6.1 Teste de remoção de ruídos

A avaliação de desempenho das técnicas de remoção de ruído foi feita através de um estudo duplo-cego com 7 voluntários com a faixa etária de 21 a 26 anos. A base utilizada foi a base de testes de ruído, elaborada conforme descrito na Subseção 5.2. Nesse ensaio, considera-se a existência de três atores: o pesquisador; o sorteador; e o aplicador do teste.

O pesquisador entrega para o sorteador o questionário presente no Anexo B, em conjunto com a base de testes. O sorteador, por sua vez, fica responsável por ofuscar os nomes dos métodos presentes na base, além de aleatorizar a ordem em que os áudios serão apresentados para os voluntários. A Figura 24 ilustra o processo de ofuscação dos métodos.

Figura 24 – Processo de ofuscação dos métodos.

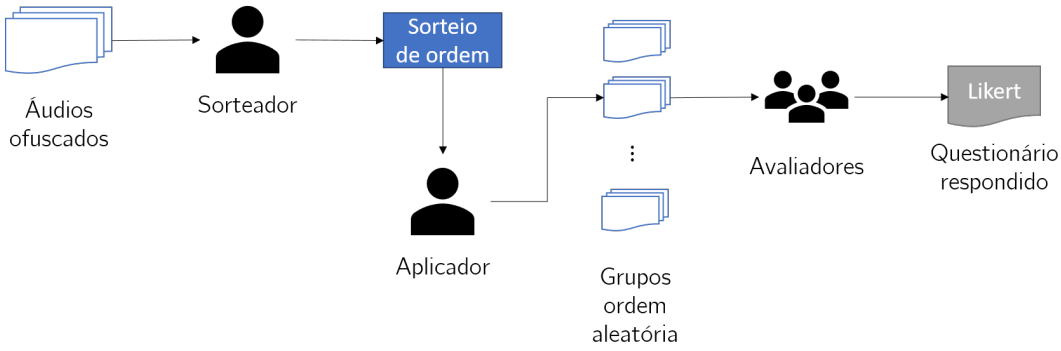


Fonte: Elaborado pelo autor.

Posteriormente, o sorteador entrega os questionários ordenados aleatoriamente, assim como os áudios ofuscados, para o aplicador do teste. Em seguida, o aplicador coleta

as assinaturas do termo de participação presente no Anexo A, e aplica os questionários. Esse processo é ilustrado pela Figura 25.

Figura 25 – Aplicação de questionário na metodologia duplo-cego.



Fonte: Elaborado pelo autor.

Durante a execução desses testes, o pesquisador, o aplicador, e os avaliadores desconheciam o nome dos métodos e as suas ordens de execução. Após os questionários terem sido respondidos, o aplicador entregou os resultados para o pesquisador, que os tabelou e realizou os cálculos necessários de validação. Somente então, o sorteador revelou o nome dos métodos.

Os avaliadores foram convidados a responder três categorias de pergunta. A primeira categoria referia-se à presença de ruídos nos áudios, onde o avaliador escutou todas as frases presentes nos arquivos de áudio, e respondeu em uma escala de *Likert* de 5 pontos o quão ruidoso ele considerava cada frase escutada. Nessa escala, o ponto 1 representava um áudio muito ruidoso, enquanto o ponto 5 representava um áudio livre de ruídos.

A segunda categoria referia-se a qualidade geral do áudio, onde o avaliador analisou a clareza do que estava sendo dito, além de atentar para a presença de distorções nos áudios. Assim como na primeira categoria, existia uma pergunta para cada frase escutada, com respostas em uma escala de *Likert* de 5 pontos. Nessa segunda escala, o ponto 1 representava um áudio que não era possível compreender o que estava sendo dito, enquanto o ponto 5 indicava que foi possível compreender a frase perfeitamente.

Vale ressaltar que nessas duas primeiras categorias, constavam no conjunto de áudios as versões de controle com ruído e sem ruído. Dessa forma, obteve-se uma indicação segundo a percepção dos avaliadores do que eles consideravam um áudio de qualidade.

A terceira e última categoria tange à escolha do melhor método. Os avaliadores foram solicitados a eleger um dos métodos como o mais eficiente. Nessa pergunta, foram

excluídos os áudios de controle, de forma a forçar os avaliadores a escolherem somente os áudios processados pelos métodos de remoção de ruído.

A Tabela 2 apresenta os resultados da coleta de dados do questionário para a primeira categoria de perguntas. Nessa tabela, temos a média aritmética de todos os itens, separados por método, além da média dos valores mínimos e máximos que foram marcados durante os testes. Destacam-se o SEGAN e a remoção de ruídos com a *wavelet symlet*, houve pouca diferença da eficácia das duas e ambas apresentaram melhorias significativas em relação à base de controle com ruídos.

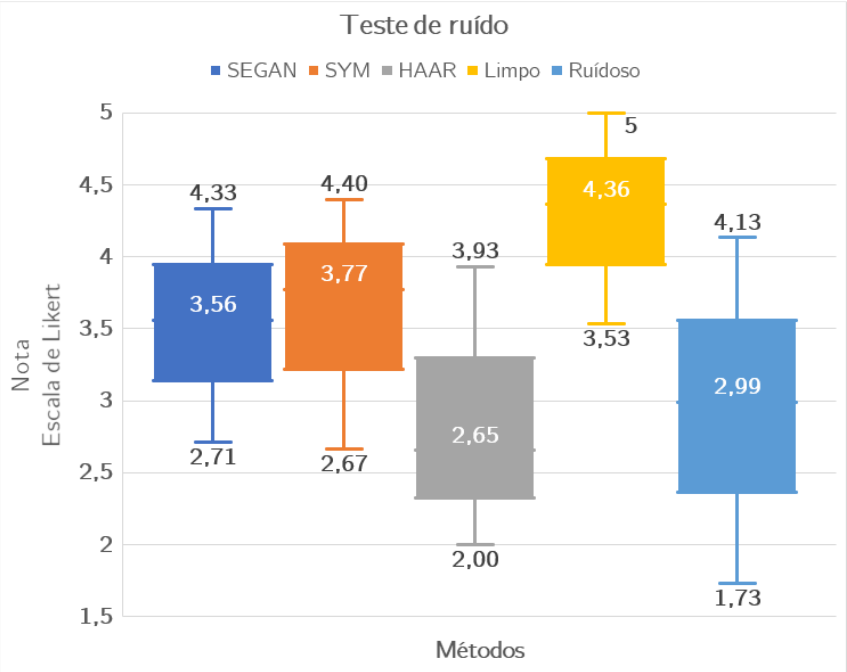
Tabela 2 – Resultados do questionário para presença de ruído.

Método	Média	Min	Max
SEGAN	3,56	2,71	4,33
SYM	3,77	2,67	4,40
HAAR	2,65	2	3,93
Limpo	4,36	3,53	5
Ruidoso	2,99	1,73	4,13

Fonte: Elaborado pelo autor.

A Figura 26 ilustra os resultados dessa primeira categoria de pergunta em forma de gráfico de caixa. Nesse gráfico torna-se perceptível a equivalência entre a SEGAN e a *wavelet symlet*.

Figura 26 – Resultados do teste de presença de ruído.



Fonte: Elaborado pelo autor.

A Tabela 3, por sua vez, apresenta os resultados para a segunda categoria de perguntas, onde a qualidade geral do som é levada em consideração. Também são apresentadas as mesmas médias da tabela anterior. Nesse caso, nota-se que o SEGAN obteve uma média maior que a da *symlet*, assim como a variação de pontos mínimos e máximos teve menor amplitude nesse método, apresentando resultados mais consistentes.

Tabela 3 – Resultados do questionário para qualidade do áudio.

Método	Média	Min	Max
SEGAN	4,14	2,73	4,73
SYM	3,92	2,13	4,67
HAAR	3,66	2,07	4,53
Limpo	4,77	4,2	5
Ruidoso	4,64	3,6	5

Fonte: Elaborado pelo autor.

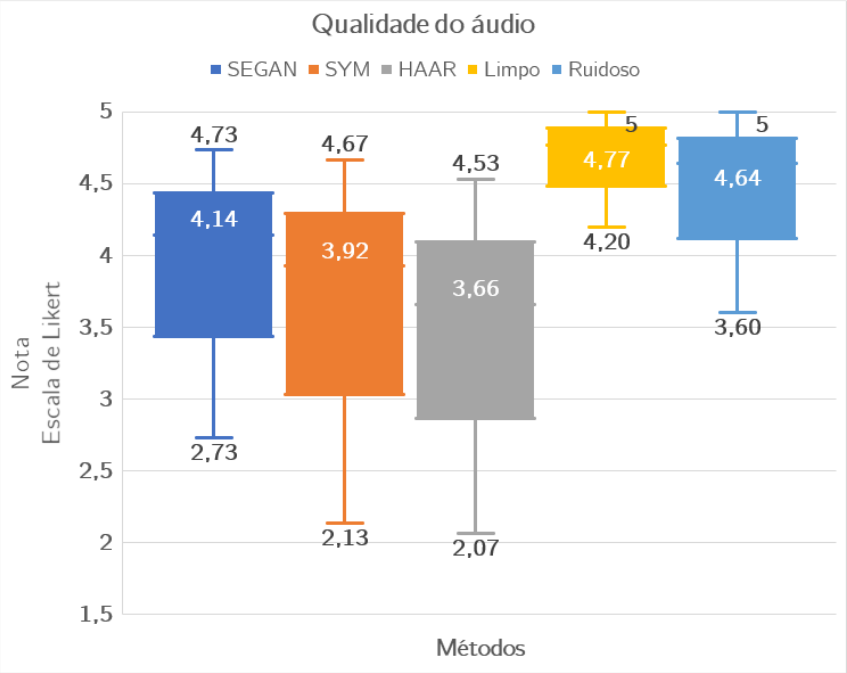
Na Figura 27 os resultados também foram apresentados em gráfico de caixa. É possível notar que nessa categoria os avaliadores atribuíram notas mais altas para a base de controle com ruídos em relação à categoria anterior. Ao comparar a qualidade do áudio das bases de controle com as bases geradas através dos três métodos, percebe-se que houve uma queda na qualidade geral do áudio.

Em outras palavras, segundo a percepção dos avaliadores, os métodos aplicados conseguiram efetivamente remover o ruído no sinal de áudio, mas acabaram por prejudicar um pouco a qualidade do sinal original, dificultando a sua compreensão.

Por fim, a Figura 28 demonstra os resultados para a última categoria de perguntas, onde os avaliadores foram solicitados a escolherem um entre os três métodos de remoção de ruído, de acordo com o que eles consideraram mais eficiente. É possível notar que novamente a SEGAN e a *symlet* dividiram opiniões, havendo somente um voto de diferença a favor da *symlet*, devido ao seu desempenho similar.

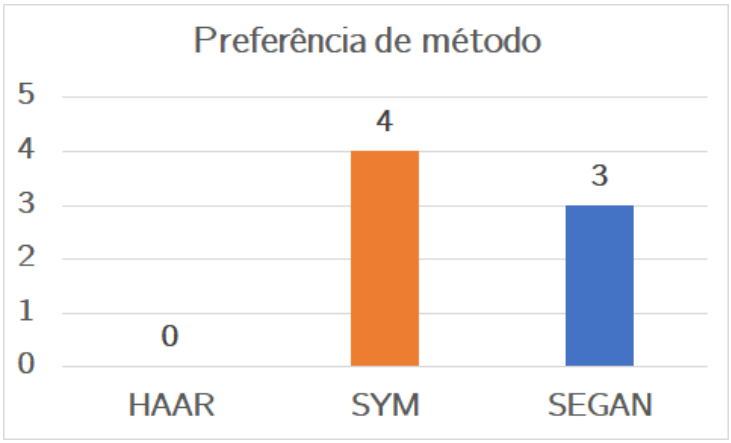
Vale ressaltar que apesar desses experimentos terem gerado resultados similares entre os métodos usando a SEGAN e a *wavelet symlet*, a rede neural foi treinada utilizando locutores nativos da língua inglesa. A SEGAN obteve resultados satisfatórios, e muito similares ao outro método, apesar de ter sido testada com uma base que possui locutores de um idioma que não fez parte da etapa de treino. Portanto, supõe-se que com uma base de treino adequada, e com locutores que falem a mesma língua, a eficácia desse método pode melhorar e superar o método de remoção de ruídos com *wavelets*.

Figura 27 – Resultados do teste de qualidade do áudio.



Fonte: Elaborado pelo autor.

Figura 28 – Resultados do teste de preferência de método.



Fonte: Elaborado pelo autor.

6.2 Teste de segmentação de áudio

A validação do método de segmentação proposto foi feita através do uso das métricas de Precisão, Sensibilidade e o Coeficiente de Similaridade de *Dice* (CSD). A base utilizada para os testes de segmentação foi elaborada conforme descrito na Subseção 5.2. Todos os áudios presentes na base de testes foram segmentados através da metodologia proposta, e a validação foi feita de forma manual, analisando os segmentos individualmente.

Levou-se em consideração para a validação três fatores: clareza; quantidade de códigos segmentados; e duração do segmento. O primeiro fator remete à capacidade do avaliador compreender qual foi o BC pronunciado no áudio. O segundo fator remete à quantos BC foram segmentados. Enquanto o último fator remete à duração total do segmento de áudio.

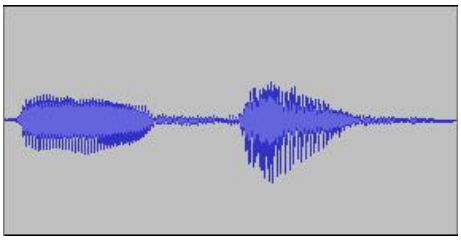
Um áudio somente foi considerado como corretamente segmentado se todos os critérios abaixo foram atendidos.

- É possível compreender qual BC foi pronunciado no segmento;
- O segmento extraído pela técnica contém apenas um BC;
- A duração do segmento, somada à duração de eventuais silêncios incluídos na segmentação, não superou o dobro da duração do segmento marcado manualmente.

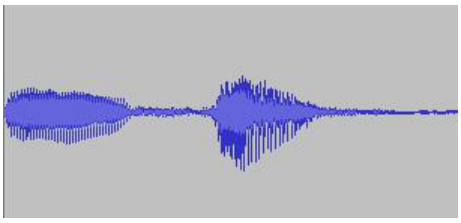
A Figura 29 apresenta um exemplo da aplicação do método de segmentação de áudio, e a sua respectiva marcação manual. Nesse caso, a segmentação é classificada como correta.

Figura 29 – Exemplo de segmentação de BC - “Oba”.

(a) Marcação manual.



(b) Marcação gerada.



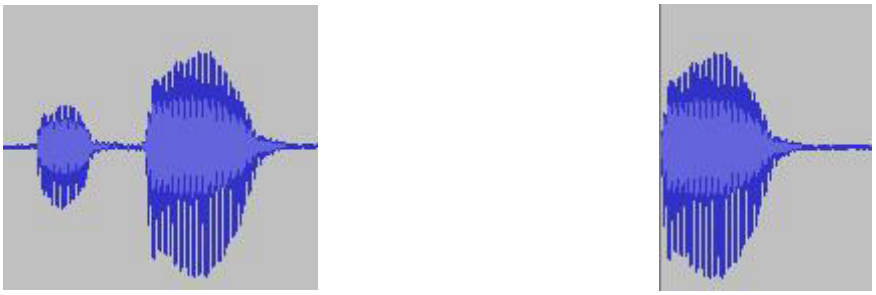
Fonte: Elaborado pelo autor.

Na Figura 30 é ilustrado um exemplo de uma segmentação que é classificada como incorreta. É possível notar na imagem que o método proposto gerou um segmento menor do que o esperado, efetivamente segmentando somente o som “ba”.

Dito isso, os resultados dos testes de segmentação são apresentados na Tabela 4. Os resultados individuais de Segmentos Corretamente Identificados (SCI), de Segmentos Incorretamente Identificados (SII), assim como a precisão, sensibilidade, e CSD (DICE) são apresentados para cada variação da base de testes, incluindo as versões de controle. Os valores percentuais encontrados nessa tabela representam a média aritmética dessas métricas para os 9 arquivos de áudio em cada base.

Figura 30 – Exemplo de segmentação incorreta de BC - “Oba”.

(a) Marcação manual. (b) Marcação gerada.



Fonte: Elaborado pelo autor.

Tabela 4 – Resultados dos testes de segmentação.

	N	SCI	SII	PRE	SEN	DICE
Limpo	297	257	50	83,71% (12,67)	86,53% (10,82)	85,10% (11,58)
Ruidoso	297	228	60	79,17% (15,57)	76,77% (24,62)	77,95% (21,47)
Haar	297	219	74	74,74% (15,09)	73,74% (24,28)	74,24% (21,07)
Symlets	297	227	64	78,01% (14,25)	76,43% (24,23)	77,21% (20,76)
Método proposto	297	212	48	81,54% (12,92)	71,38% (18,99)	76,12% (15,47)

Fonte: Elaborado pelo autor.

Os valores de *Dice score* obtidos foram, em média, satisfatórios. Entretanto, percebe-se uma queda de aproximadamente 10% ao se comparar os resultados obtidos com a base sem ruído e as demais. Tal perda de desempenho se deve ao fato do ruído aditivo dificultar a localização de trechos de silêncio. Além disso, os métodos de remoção de ruído acabaram por reduzir a qualidade geral dos arquivos de áudio, contribuindo ainda mais para essa elevada variação de desempenho.

Ao analisar as bases onde foram empregadas técnicas de remoção de ruído, foi possível observar que a precisão e sensibilidade mantiveram-se estáveis entre essas bases, contudo, o desvio padrão de todas as métricas foi elevado, chegando até a 21,07% em um dos casos.

O desempenho da segmentação em cima da base sem ruídos foi consideravelmente melhor do que nos demais casos de teste, havendo também um desvio padrão muito menor. Além disso, em dois casos de teste, o *Dice score* chegou a 100%. Contudo, foi possível observar um caso onde o CSD foi de apenas 67,61%. Ao analisar manualmente esses casos, observou-se que nos melhores casos os locutores pronunciavam os BC com um tom de voz mais elevado e sem pausas desnecessárias entre as sílabas. Enquanto o locutor do pior caso pronunciou os BC com um tom de voz mais baixo e com um espaçamento de tempo mais

prolongado entre as sílabas dos BC. Supõe-se que se o processo de gravação dos áudios seguir um protocolo mais rígido, será possível eliminar esses casos com CSD baixos.

Por fim, foi observado que o áudio de um dos locutores que obteve CSD de 100% teve uma queda de desempenho para aproximadamente 21% nas bases com ruído, exceto a SEGAN, onde esse valor foi de 93,94%. Esse caso também foi analisado manualmente, e notou-se que o ruído desse arquivo de áudio era caracterizado por ventos fortes. Nesse caso, os algoritmos de remoção de ruído com *wavelets* não tiveram exito em remover o ruído, resultando em uma segmentação muito falha. Entretanto, a SEGAN se mostrou eficaz para remover esse tipo específico de ruído, o que justificou a baixa variação do CSD em relação ao resultado do áudio original.

7 Conclusão

Este trabalho apresentou uma metodologia para segmentação automática de BC em arquivos de áudio. Para tal, foram utilizadas técnicas de remoção de ruídos baseadas em redes generativas adversárias e *wavelets*.

Para desenvolver esse trabalho, foi montada uma base de áudios em português, contendo 192 frases distintas, gravadas com 9 locutores diferentes. Cada arquivo de áudio gerado também possui a sua versão paralela com ruídos, o que torna essa base um instrumento importante na validação de diversos sistemas de processamento e reconhecimento de áudio. Além disso, foi feito um levantamento de outras bases que foram utilizadas no treinamento uma SEGAN para a remoção de ruídos em áudios. Apesar do desempenho dessa rede não ter superado outras técnicas de remoção de ruído, ainda apresentou resultados equivalentes.

Quanto à etapa de segmentação de áudio, utilizou-se técnicas estatísticas para atenuar o sinal de áudio e facilitar a detecção de trechos de fala, além de um algoritmo de janela deslizante que segmentou os áudios com uma precisão média de aproximadamente 83% e sensibilidade média de 86%, obtendo-se um DSC acima de 85% em áudios com pouco ruído. Em situações de muito ruído, o método proposto apresentou, respectivamente, médias de precisão, sensibilidade e DSC de 79,19%, 76,77% e 77,95%.

Portanto, conclui-se que o objetivo geral deste trabalho foi alcançado, pois o método de segmentação utilizado nessa pesquisa foi testado e validado, assim como as técnicas de remoção de ruído. Ambos os métodos forneceram resultados satisfatórios, entretanto torna-se necessário um estudo mais aprofundado de outras técnicas de segmentação, assim como a realização do treinamento da SEGAN utilizando-se uma base completamente em português, de forma a melhorar os resultados obtidos e tornar o método ainda mais confiável.

7.1 Trabalhos futuros

Como trabalhos futuros, sugere-se explorar novas técnicas de aquisição de áudio, preferencialmente em estúdios onde a presença de ruídos é virtualmente nula, pois dessa forma a base de áudio terá instâncias de maior qualidade para serem utilizadas em treinamentos de modelos acústicos e de remoção de ruídos. Para a etapa de pré-

processamento, recomenda-se a aplicação de técnicas de atenuação do sinal que não sejam destrutivas, de forma que o sinal original não perca características importantes na montagem de novos modelos. Para a etapa de remoção de ruídos através de redes SEGAN, aconselha-se o treinamento utilizando bases de áudio no mesmo idioma no qual serão realizados os testes. Alternativamente, considera-se a aplicação de técnicas de transferência de aprendizagem, do inglês *Transfer Learning*, para melhorar o desempenho da remoção de ruídos, similar ao realizado por (PASCUAL et al., 2017). Para a etapa de segmentação de áudio, recomenda-se o uso de descritores de áudio extras além da simples detecção de picos e vales. Por fim, como forma de automatizar o processo de detecção de BC, recomenda-se o uso de técnicas de ASR, como o modelo oculto de *markov* ou utilizando redes neurais com aprendizagem profunda.

Referências

ABADI, M.; BARHAM, P.; CHEN, J.; CHEN, Z.; DAVIS, A.; DEAN, J.; DEVIN, M.; GHEMAWAT, S.; IRVING, G.; ISARD, M. et al. Tensorflow: a system for large-scale machine learning. In: *OSDI*. Savannah - USA: usenix, 2016. v. 16, p. 265–283.

ABUSHARIAH, M. A.; AINON, R. N.; ZAINUDDIN, R.; ELSHAFEI, M.; KHALIFA, O. O. Natural speaker-independent arabic speech recognition system based on hidden markov models using sphinx tools. In: IEEE. *Computer and Communication Engineering (ICCCE), 2010 International Conference on*. [S.l.], 2010. p. 1–6.

AGREIL, C.; MEURET, M. An improved method for quantifying intake rate and ingestive behaviour of ruminants in diverse and variable habitats using direct observation. *Small Ruminant Research*, Elsevier, v. 54, n. 1, p. 99–113, 2004.

AMINGHAFARI, M.; CHEZE, N.; POGGI, J.-M. Multivariate denoising using wavelets and principal component analysis. *Computational Statistics & Data Analysis*, Elsevier, v. 50, n. 9, p. 2381–2398, 2006.

BAHOURA, M.; ROUAT, J. Wavelet speech enhancement based on time–scale adaptation. *Speech Communication*, Elsevier, v. 48, n. 12, p. 1620–1637, 2006.

BÉNÉTEAU, C.; FLEET, P. J. V. Discrete wavelet transformations and undergraduate education. *Notices of the AMS*, v. 58, n. 05, 2011.

CHETLUR, S.; WOOLLEY, C.; VANDERMERSCH, P.; COHEN, J.; TRAN, J.; CATANZARO, B.; SHELHAMER, E. cudnn: Efficient primitives for deep learning. *arXiv preprint arXiv:1410.0759*, 2014.

CHIŃU, A. G.; ROTHKRANTZ, L. J. Building a data corpus for audio-visual speech recognition. In: *Euromedia*. [S.l.: s.n.], 2007. p. 88–92.

CODEN, A. R.; BROWN, E. W.; SRINIVASAN, S. *Information Retrieval Techniques for Speech Applications*. [S.l.]: Springer Science & Business Media, 2002. v. 2273.

COOK, S. Speech recognition howto. *The Linux Documentation Project*, 2002.

DAUBECHIES, I. *Ten lectures on wavelets*. [S.l.]: Siam, 1992. v. 61.

DICE, L. R. Measures of the amount of ecologic association between species. *Ecology*, Wiley Online Library, v. 26, n. 3, p. 297–302, 1945.

DIJKSTRA, J.; FORBES, J. M.; FRANCE, J. *Quantitative aspects of ruminant digestion and metabolism*. [S.l.]: CABI, 2005.

GELFAND, S. A.; LEVITT, H. *Hearing: An introduction to psychological and physiological acoustics*. [S.l.]: Marcel Dekker New York, 1998. v. 4.

GOODFELLOW, I. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

GOODFELLOW, I.; POUGET-ABADIE, J.; MIRZA, M.; XU, B.; WARDE-FARLEY, D.; OZAIR, S.; COURVILLE, A.; BENGIO, Y. Generative adversarial nets. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2014. p. 2672–2680.

HAAR, A. Zur theorie der orthogonalen funktionensysteme. *Mathematische Annalen*, Springer, v. 69, n. 3, p. 331–371, 1910.

HAYKIN, S. *Redes neurais: princípios e prática*. [S.l.]: Bookman Editora, 2007.

HERTZ, J. A. *Introduction to the theory of neural computation*. [S.l.]: CRC Press, 2018.

HO, T. K. Random decision forests. In: IEEE. *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*. [S.l.], 1995. v. 1, p. 278–282.

HYOUNG-GOOK, K.; MOREAU, N.; SIKORA, T. *MPEG-7 Audio and Beyond*. [S.l.]: John Wiley & Sons Ltd, 2005.

IBGE. *Produção da pecuária municipal*. IBGE, 2015. v. 43. Disponível em: ⟨http://biblioteca.ibge.gov.br/visualizacao/periodicos/84/ppm_2015_v43_br.pdf⟩.

JOHNSON, D. H. Signal-to-noise ratio. *Scholarpedia*, v. 1, n. 12, p. 2088, 2006.

KALANTARIAN, H.; SARRAFZADEH, M. Audio-based detection and evaluation of eating behavior using the smartwatch platform. *Computers in biology and medicine*, Elsevier, v. 65, p. 1–9, 2015.

KNOBEL, M. Física da fala e da audição. *Campinas: Instituto de Física da UNICAMP*.

KRÖSE, B.; KROSE, B.; SMAGT, P. van der; SMAGT, P. An introduction to neural networks. Citeseer, 1993.

LECUN, Y.; KAVUKCUOGLU, K.; FARABET, C. et al. Convolutional networks and applications in vision. In: *ISCAS*. [S.l.: s.n.], 2010. v. 2010, p. 253–256.

LEE, K.-F. On large-vocabulary speaker-independent continuous speech recognition. *Speech communication*, Elsevier, v. 7, n. 4, p. 375–379, 1988.

LI, X. *Combination and generation of parallel feature streams for improved speech recognition*. Tese (Doutorado) — Carnegie Mellon University Pittsburgh, PA, 2005.

LIANG, H.; SUN, X.; SUN, Y.; GAO, Y. Text feature extraction based on deep learning: a review. *EURASIP journal on wireless communications and networking*, Nature Publishing Group, v. 2017, n. 1, p. 211, 2017.

LIKERT, R. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

LIOU, C.-Y.; CHENG, W.-C.; LIOU, J.-W.; LIOU, D.-R. Autoencoder for words. *Neurocomputing*, Elsevier, v. 139, p. 84–96, 2014.

LOGAN, B. et al. Mel frequency cepstral coefficients for music modeling. In: *ISMIR*. [S.l.: s.n.], 2000.

MALHOTRA, N. K. *Pesquisa de marketing: uma orientação aplicada*. [S.l.]: Bookman Editora, 2012.

MELLO, C. A. Processamento digital de sinais. *Centro de Informática UFPE. Disponível em¡ http://www.cin.ufpe.br/˜abm/pds/PDS.pdf¿. Acesso em 01/08/2018.*, v. 30, n. 09, 2013.

MORENO, P. J. *Speech recognition in noisy environments*. Tese (Doutorado) — Carnegie Mellon University Pittsburgh, 1996.

MUDA, L.; BEGAM, M.; ELAMVAZUTHI, I. Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques. *arXiv preprint arXiv:1003.4083*, 2010.

PASCUAL, S.; BONAFONTE, A.; SERRÀ, J. Segan: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452*, 2017.

PASCUAL, S.; PARK, M.; SERRÀ, J.; BONAFONTE, A.; AHN, K. Language and noise transfer in speech enhancement generative adversarial network. *CoRR*, abs/1712.06340, 2017. Disponível em: ⟨http://arxiv.org/abs/1712.06340⟩.

PASQUALI, L. Psychometrics. *Revista da Escola de Enfermagem da USP*, SciELO Brasil, v. 43, n. SPE, p. 992–999, 2009.

PENAGARIKANO, M.; BORDEL, G. Sautrela: a highly modular open source speech recognition framework. In: IEEE. *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on*. University of the Basque Country, 48940 Leioa, Spain: IEEE, 2005. p. 386–391.

PUKELSHEIM, F. The three sigma rule. *The American Statistician*, Taylor & Francis, v. 48, n. 2, p. 88–91, 1994.

RAMANA, A.; LAXMINARAYANA, P.; MYTHILISHARAN, P. Investigation of speech coding effects on different speech sounds in automatic speech recognition. In: *Perception and Machine Intelligence*. [S.l.]: Springer, 2012. p. 367–377.

RIVERS, W.; WEBBER, H. The action of caffeine on the capacity for muscular work. *The Journal of physiology*, Wiley Online Library, v. 36, n. 1, p. 33–47, 1907.

ROBERTSON, C. T.; KESSELHEIM, A. S. *Blinding as a solution to bias: Strengthening biomedical science, forensic science, and law*. [S.l.]: Academic Press, 2016.

RUMELHART, D. E.; MCCLELLAND, J. L.; GROUP, P. R. et al. *Parallel distributed processing*. [S.l.]: MIT press Cambridge, MA, 1987. v. 1.

SHANNON, C. E. Communication in the presence of noise. *Proceedings of the IRE*, IEEE, v. 37, n. 1, p. 10–21, 1949.

SHRAWANKAR, U.; THAKARE, V. Noise estimation and noise removal techniques for speech recognition in adverse environment. In: SPRINGER. *International Conference on Intelligent Information Processing*. [S.l.], 2010. p. 336–342.

SILVA, I. d.; SPATTI, D. H.; FLAUZINO, R. A. Redes neurais artificiais para engenharia e ciências aplicadas. *São Paulo: Artliber*, v. 23, n. 5, p. 33–111, 2010.

SOEST, P. J. V. *Nutritional ecology of the ruminant*. [S.l.]: Cornell University Press, 1994.

STOLLNITZ, E. J.; DEROSE, A. D.; SALESIN, D. H. Wavelets for computer graphics: a primer. 1. *IEEE Computer Graphics and Applications*, IEEE, v. 15, n. 3, p. 76–84, 1995.

TEAM, A. Audacity (version 2.0.0). *Audio editor and recorder*, 2012.

THIEMANN, J.; ITO, N.; VINCENT, E. *DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments*. 2013. Supported by Inria under the Associate Team Program VERSAMUS. Disponível em: ⟨https://doi.org/10.5281/zenodo.1227121⟩.

TORRES-ACOSTA, J.; CASTRO, C. A. S. Adapting a bite coding grid for small ruminants browsing a deciduous tropical forest. *Tropical and Subtropical Agroecosystems*, v. 17, n. 1, 2014.

VALENTINI-BOTINHAO, C.; WANG, X.; TAKAKI, S.; YAMAGISHI, J. Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks. In: *Interspeech*. [S.l.: s.n.], 2016. p. 352–356.

WALKER, J. S. *Fast fourier transforms*. [S.l.]: CRC press, 1996. v. 24.

WALKER, J. S. *A primer on wavelets and their scientific applications*. [S.l.]: CRC press, 2002.

WIDROW, B.; LEHR, M. A. 30 years of adaptive neural networks: perceptron, madaline, and backpropagation. *Proceedings of the IEEE*, IEEE, v. 78, n. 9, p. 1415–1442, 1990.

YU, G.; MALLAT, S.; BACRY, E. Audio denoising by time-frequency block thresholding. *IEEE Transactions on Signal processing*, IEEE, v. 56, n. 5, p. 1830–1839, 2008.

ZEN, S. D.; SANTOS, M.; MONTEIRO, C. M. Evolução da caprino e ovinocultura. *Ativos da Pecuária de Caprino e Ovinocultura 1p*, 2014.

ZIJDENBOS, A. P.; DAWANT, B. M.; MARGOLIN, R. A.; PALMER, A. C. Morphometric analysis of white matter lesions in mr images: method and validation. *IEEE transactions on medical imaging*, IEEE, v. 13, n. 4, p. 716–724, 1994.

Anexo A – Termo de consentimento

Termo de participação

Nós, abaixo assinado, autorizamos o uso, com fins acadêmicos, das informações fornecidas durante a sessão de teste qualitativo das ferramentas de remoção de ruído utilizadas na pesquisa do mestrando Maurício César Pinto Pessoa, realizadas na Universidade Federal do Maranhão, no dia ___/___/2018, a partir das ___ horas.

Estamos cientes de que: (1) Nossa participação é voluntária; (2) a sessão será registrada com anotações e fotografias; (3) esta sessão também visa prover informações a uma pesquisa acadêmica relacionada à técnicas de remoção de ruído em arquivos de áudio digital, desenvolvido por Maurício César Pinto Pessoa, sob orientação do professor doutor Tiago Bonini Borchartt, na Universidade Federal do Maranhão; (4) no uso das informações geradas nesta sessão de avaliação, será garantido o anonimato referente às contribuições pontuais.

( ) Autorizamos o uso das fotos no contexto da pesquisa supracitada.

( ) Não autorizamos o uso das fotos no contexto da pesquisa supracitada.

Ordinal	Nome	CPF	Assinatura
1.			
2.			
3.			
4.			
5.			
6.			
7.			
8.			
9.			
10.			

Anexo B – Questionário para teste de remoção de ruídos

Questionário qualitativo de remoção de ruídos em áudio digital

Aos voluntários:

Esse questionário foi montado com o objetivo de medir a qualidade de diversas técnicas de remoção de ruído em arquivos de áudio digital, com base na percepção auditiva humana.

Pedimos que leiam atentamente as instruções abaixo antes de iniciar a resposta do questionário.

- O aplicador do questionário fornecerá uma pasta com diversos arquivos de áudio.
- Cada arquivo de áudio contém 15 frases.
- Cada arquivo de áudio foi processado utilizando técnicas distintas de remoção de ruídos.
- Pedimos que você escute todos os arquivos de áudio atentamente, em especial pedimos atenção:
    a. Na presença de ruídos no áudio, tais como – vento, animais, ruído branco, etc.
    b. Na clareza do que está sendo dito em cada frase, ou seja, se é possível entender o que está sendo dito, ou se houve dificuldade na compreensão, se o som estava muito distorcido, etc.
- Responda o questionário com base no que você escutou.
- Você poderá escutar os áudios quantas vezes julgar necessário.
    a. Caso já tenha começado a responder o questionário, e estiver em dúvida, você poderá escutar novamente qualquer áudio ou frase.
    b. Responda uma pergunta de cada vez, escutando cada frase individualmente.

Antes de começar, por favor informe a sua idade:

- Idade: _____ Anos.

Parte 1: Presença de ruídos no áudio

Segundo a sua percepção, quão ruidoso estão os áudios que você escutou?

Considerando que na escala **1 = Muito ruidoso** e **5 = Sem ruídos**, marque uma opção para cada metodologia apresentada.

- **Frase 1**

Metodologia	1	2	3	4	5
A	( )	( )	( )	( )	( )
B	( )	( )	( )	( )	( )
C	( )	( )	( )	( )	( )
D	( )	( )	( )	( )	( )
E	( )	( )	( )	( )	( )

- **Frase 2**

Metodologia	1	2	3	4	5
A	( )	( )	( )	( )	( )
B	( )	( )	( )	( )	( )
C	( )	( )	( )	( )	( )
D	( )	( )	( )	( )	( )
E	( )	( )	( )	( )	( )

- **Frase 3**

Metodologia	1	2	3	4	5
A	( )	( )	( )	( )	( )
B	( )	( )	( )	( )	( )
C	( )	( )	( )	( )	( )
D	( )	( )	( )	( )	( )
E	( )	( )	( )	( )	( )

- **Frase 4**

Metodologia	1	2	3	4	5
A	( )	( )	( )	( )	( )
B	( )	( )	( )	( )	( )
C	( )	( )	( )	( )	( )
D	( )	( )	( )	( )	( )
E	( )	( )	( )	( )	( )

- **Frase 5**

Metodologia	1	2	3	4	5
A	( )	( )	( )	( )	( )
B	( )	( )	( )	( )	( )
C	( )	( )	( )	( )	( )
D	( )	( )	( )	( )	( )
E	( )	( )	( )	( )	( )

- **Frase 6**

Metodologia	1	2	3	4	5
A	( )	( )	( )	( )	( )
B	( )	( )	( )	( )	( )
C	( )	( )	( )	( )	( )
D	( )	( )	( )	( )	( )
E	( )	( )	( )	( )	( )

- **Frase 7**

Metodologia	1	2	3	4	5
A	( )	( )	( )	( )	( )
B	( )	( )	( )	( )	( )
C	( )	( )	( )	( )	( )
D	( )	( )	( )	( )	( )
E	( )	( )	( )	( )	( )

- **Frase 8**

Metodologia	1	2	3	4	5
A	( )	( )	( )	( )	( )
B	( )	( )	( )	( )	( )
C	( )	( )	( )	( )	( )
D	( )	( )	( )	( )	( )
E	( )	( )	( )	( )	( )

- **Frase 9**

Metodologia	1	2	3	4	5
A	( )	( )	( )	( )	( )
B	( )	( )	( )	( )	( )
C	( )	( )	( )	( )	( )
D	( )	( )	( )	( )	( )
E	( )	( )	( )	( )	( )

- **Frase 10**

Metodologia	1	2	3	4	5
A	( )	( )	( )	( )	( )
B	( )	( )	( )	( )	( )
C	( )	( )	( )	( )	( )
D	( )	( )	( )	( )	( )
E	( )	( )	( )	( )	( )

- **Frase 11**

Metodologia	1	2	3	4	5
A	( )	( )	( )	( )	( )
B	( )	( )	( )	( )	( )
C	( )	( )	( )	( )	( )
D	( )	( )	( )	( )	( )
E	( )	( )	( )	( )	( )

- **Frase 12**

Metodologia	1	2	3	4	5
A	( )	( )	( )	( )	( )
B	( )	( )	( )	( )	( )
C	( )	( )	( )	( )	( )
D	( )	( )	( )	( )	( )
E	( )	( )	( )	( )	( )

- **Frase 13**

Metodologia	1	2	3	4	5
A	( )	( )	( )	( )	( )
B	( )	( )	( )	( )	( )
C	( )	( )	( )	( )	( )
D	( )	( )	( )	( )	( )
E	( )	( )	( )	( )	( )

- **Frase 14**

Metodologia	1	2	3	4	5
A	( )	( )	( )	( )	( )
B	( )	( )	( )	( )	( )
C	( )	( )	( )	( )	( )
D	( )	( )	( )	( )	( )
E	( )	( )	( )	( )	( )

- **Frase 15**

Metodologia	1	2	3	4	5
A	( )	( )	( )	( )	( )
B	( )	( )	( )	( )	( )
C	( )	( )	( )	( )	( )
D	( )	( )	( )	( )	( )
E	( )	( )	( )	( )	( )

Parte 2: Qualidade do áudio

Você conseguiu compreender bem o que foi dito nos áudios? Leve em consideração a qualidade geral do áudio, em especial a presença ou ausência de distorções e clareza do que está sendo dito.

Considerando que na escala **1 = Não foi possível compreender as frases** e **5 = Compreendi tudo perfeitamente**, marque uma opção para cada metodologia apresentada.

- **Frase 1**

Metodologia	1	2	3	4	5
A	( )	( )	( )	( )	( )
B	( )	( )	( )	( )	( )
C	( )	( )	( )	( )	( )
D	( )	( )	( )	( )	( )
E	( )	( )	( )	( )	( )

- **Frase 2**

Metodologia	1	2	3	4	5
A	( )	( )	( )	( )	( )
B	( )	( )	( )	( )	( )
C	( )	( )	( )	( )	( )
D	( )	( )	( )	( )	( )
E	( )	( )	( )	( )	( )

- **Frase 3**

Metodologia	1	2	3	4	5
A	( )	( )	( )	( )	( )
B	( )	( )	( )	( )	( )
C	( )	( )	( )	( )	( )
D	( )	( )	( )	( )	( )
E	( )	( )	( )	( )	( )

- **Frase 4**

Metodologia	1	2	3	4	5
A	( )	( )	( )	( )	( )
B	( )	( )	( )	( )	( )
C	( )	( )	( )	( )	( )
D	( )	( )	( )	( )	( )
E	( )	( )	( )	( )	( )

- **Frase 5**

Metodologia	1	2	3	4	5
A	( )	( )	( )	( )	( )
B	( )	( )	( )	( )	( )
C	( )	( )	( )	( )	( )
D	( )	( )	( )	( )	( )
E	( )	( )	( )	( )	( )

- **Frase 6**

Metodologia	1	2	3	4	5
A	( )	( )	( )	( )	( )
B	( )	( )	( )	( )	( )
C	( )	( )	( )	( )	( )
D	( )	( )	( )	( )	( )
E	( )	( )	( )	( )	( )

- **Frase 7**

Metodologia	1	2	3	4	5
A	( )	( )	( )	( )	( )
B	( )	( )	( )	( )	( )
C	( )	( )	( )	( )	( )
D	( )	( )	( )	( )	( )
E	( )	( )	( )	( )	( )

- **Frase 8**

Metodologia	1	2	3	4	5
A	( )	( )	( )	( )	( )
B	( )	( )	( )	( )	( )
C	( )	( )	( )	( )	( )
D	( )	( )	( )	( )	( )
E	( )	( )	( )	( )	( )

- **Frase 9**

Metodologia	1	2	3	4	5
A	( )	( )	( )	( )	( )
B	( )	( )	( )	( )	( )
C	( )	( )	( )	( )	( )
D	( )	( )	( )	( )	( )
E	( )	( )	( )	( )	( )

- **Frase 10**

Metodologia	1	2	3	4	5
A	( )	( )	( )	( )	( )
B	( )	( )	( )	( )	( )
C	( )	( )	( )	( )	( )
D	( )	( )	( )	( )	( )
E	( )	( )	( )	( )	( )

- **Frase 11**

Metodologia	1	2	3	4	5
A	( )	( )	( )	( )	( )
B	( )	( )	( )	( )	( )
C	( )	( )	( )	( )	( )
D	( )	( )	( )	( )	( )
E	( )	( )	( )	( )	( )

- **Frase 12**

Metodologia	1	2	3	4	5
A	( )	( )	( )	( )	( )
B	( )	( )	( )	( )	( )
C	( )	( )	( )	( )	( )
D	( )	( )	( )	( )	( )
E	( )	( )	( )	( )	( )

- **Frase 13**

Metodologia	1	2	3	4	5
A	( )	( )	( )	( )	( )
B	( )	( )	( )	( )	( )
C	( )	( )	( )	( )	( )
D	( )	( )	( )	( )	( )
E	( )	( )	( )	( )	( )

- **Frase 14**

Metodologia	1	2	3	4	5
A	( )	( )	( )	( )	( )
B	( )	( )	( )	( )	( )
C	( )	( )	( )	( )	( )
D	( )	( )	( )	( )	( )
E	( )	( )	( )	( )	( )

- **Frase 15**

Metodologia	1	2	3	4	5
A	( )	( )	( )	( )	( )
B	( )	( )	( )	( )	( )
C	( )	( )	( )	( )	( )
D	( )	( )	( )	( )	( )
E	( )	( )	( )	( )	( )

Parte 3: Preferência de metodologia

Dentre as metodologias de remoção de ruído aplicadas nesse experimento, marque abaixo uma das três opções, considerando a que você considerou mais eficiente (que melhor removeu os ruídos de áudio).

Metodologia A	Metodologia B	Metodologia C
( )	( )	( )

Anexo C – Roteiro de gravação de áudio

Roteiro de gravação

Ao participante:

- Leia as frases normalmente, sem pressa.
- Dê uma pequena pausa entre uma frase e a seguinte.
- Pode haver interrupções entre as frases para descanso.
- Você não precisa ler e falar ao mesmo tempo.
    - Se julgar necessário, leia mentalmente a frase quantas vezes quiser antes de falar ao microfone.
- Tente empregar da melhor forma a sua dicção.
- Evite deixar o microfone muito próximo da boca, e evite falar diretamente para ele.
    - Isso ajudará a evitar sons de respiração.
- As frases são bastante aleatórias. Tente não expressar reações (risos, comentários, etc).

Frase 1: Todo o poder emana do povo.
Frase 2: prevalência dos direitos humanos.
Frase 3: é garantido o direito de herança.
Frase 4: repouso semanal remunerado, preferencialmente aos Domingos.
Frase 5: Viva na liberdade do dia a dia.
Frase 6: Viver não é vencer ou derrotar.
Frase 7 : O dia a dia se torna a sobremesa da vida.
Frase 8: Olá, meu nome é João.
Frase 9: Tirei dez na prova.
Frase 10: Eu nem queria mesmo.
Frase 11: Hoje são 30 do mês de Abril.
Frase 12: Na escola éramos felizes.
Frase 13: Quem de nós dois vai dizer que é impossível.
Frase 14: Eu estou pensando em você.
Frase 15: Queremos uma revolução em nós mesmos.
Frase 16: Hoje preciso de você com qualquer humor.
Frase 17: Preciso ir buscar o carro agora.
Frase 18: Você quer brincar na neve.
Frase 19: Queria poder te dizer que te amo.
Frase 20: Desisti de correr atrás das coisas grandes.
Frase 21: Apenas desfruto a vida na tranquilidade do dia a dia.
Frase 22: O maior presente da vida são as experiências.
Frase 23: Não queira tudo de uma só vez.
Frase 24: Você não sabe nada, Jon Snow.
Frase 25: Muitos que a tentam escalar, falham e nunca mais tentam de novo.
Frase 26: A escalada é tudo o que existe.
Frase 27: Eu quero ser a rainha.
Frase 28: Um leão não se preocupa com a opinião de uma ovelha.
Frase 29: Você sabe que horas são?
Frase 30: Em que ano estamos?
Frase 31: Qual é a data do seu nascimento?
Frase 32: Deveríamos comprar este ou aquele combo?
Frase 33: Vamos combinar: Não sabemos como fazer isso.
Frase 34: Você já fez a cópia das chaves?
Frase 35: Queremos algo que não sabemos bem o que é.
Frase 36: Você gostaria de adicionar batatas?
Frase 37: Sem você eu não sou nada.
Frase 38: A primavera está chegando, é o fim da solidão.
Frase 39: Devia ter amado mais, me importado mais.
Frase 40: Quem acredita sempre alcança.
Frase 41: Você poderia ir comigo ao local do evento?
Frase 42: Quero café com pão.
Frase 43: Fiz a minha lição de casa hoje.
Frase 44: Você vem à aula amanhã?
Frase 45: Qual é o cardápio hoje?
Frase 46: Gostaria de solicitar a minha participação no projeto.
Frase 47: Quero estar aos pés da cruz.
Frase 48: O amor de Deus é incomparável.

Frase 49: Você gostaria de uma xícara de chá?
Frase 50: A gratidão é a memória do coração.
Frase 51: Deixe pra trás o que não te leva pra frente.
Frase 52: Falo nada, só observo.
Frase 53: Insista, persista e nunca desista.
Frase 54: Aqui se faz, aqui se paga.
Frase 55: As mudanças importantes acontecem no dia a dia, um pouquinho de cada vez.
Frase 56: Despreza-se um homem que tem ciúmes da mulher.
Frase 57: Um falso amigo é mais temível que um animal selvagem.
Frase 58: A vida me ensinou a dizer adeus às pessoas que amo.
Frase 59: Gostaria de exaltar o ego do personagem.
Frase 60: Estamos diante de um real paradoxo aqui.
Frase 61: Relaxa é necessário para se viver.
Frase 62: Matamos o tempo, o tempo nos enterra.
Frase 63: Eu não sou propriamente um autor defunto, mas um defunto autor.
Frase 64: Que é a saudade, senão uma ironia do tempo e da fortuna?
Frase 65: Nem divinizar o dinheiro, nem também baní-lo.
Frase 66: Eu deixo-me estar entre o poeta e o sábio.
Frase 67: Quem me pôs no coração esse amor de vida?
Frase 68: Não há juventude sem meninice?
Frase 69: Amei a outro; que importa, se acabou?
Frase 70: Verdade é que tinha a alma decrépita.
Frase 71: Morreu sem lhe poder valer a ciência dos médicos.
Frase 72: A valsa é uma deliciosa coisa.
Frase 73: Ventilai as consciências!
Frase 74: Não há amor possível sem a oportunidade dos sujeitos.
Frase 75: Saio em busca de um Grande Talvez.
Frase 76: Nada acontecia como eu imaginava.
Frase 77: Você realmente leu todos aqueles livros em seu quarto?
Frase 78: Você precisa aprender a conviver com as pessoas.
Frase 79: Pelo menos morro inteligente.
Frase 80: Tu és linda, sabia?
Frase 81: Tudo que é construído termina por desmoronar.
Frase 82: Tínhamos de perdoar para sobreviver no labirinto.
Frase 83: Em vão tenho lutado comigo mesmo.
Frase 84: A vaidade e o orgulho são coisas diferentes.
Frase 85: Mas confesso que a mim nada disso me encanta.
Frase 86: Prefiro infinitamente um livro.
Frase 87: Não tem nenhuma compaixão pelos meus nervos.
Frase 88: O meu amor e os meus desejos permanecem inalterados.
Frase 89: A distância é curta quando se tem um bom motivo.
Frase 90: Meu horário está cheio hoje.
Frase 91: Tem espaço em sua agenda?
Frase 92: Que horas vem o próximo ônibus?
Frase 93: Você parece surpresa.
Frase 94: Quando é o seu aniversário?
Frase 95: Diga-me com quem andas e eu direi ok.
Frase 96: Você mais que eu sonhava.

Frase 97: Quando te vejo meu coração sai pela boca.
Frase 98: Se tu olhares verás algo maior.
Frase 99: Vem matar essa paixão.
Frase 100: Numa folha qualquer eu desenho um sol amarelo.
Frase 101: Você não dava nada por mim.
Frase 102: Quero sentar ali na próxima mesa.
Frase 103: Venha cá, você não está demais hoje?
Frase 104: Não posso fixar a hora ou o lugar.
Frase 105: Nada é mais enganoso do que a aparência da humildade
Frase 106: Tenho medo de ser esquecida.
Frase 107: Todo salvamento é temporário.
Frase 108: Sei que o amor é um grito no vazio.
Frase 109: E mesmo assim doía.
Frase 110: Eu tenho medo de ser esquecido.
Frase 111: Talvez o.k. venha a ser o nosso sempre.
Frase 112: Alguns infinitos são maiores que outros.
Frase 113: Afinal, sem a Dor não reconheceríamos o prazer.
Frase 114: Às vezes parece que o universo quer ser notado.
Frase 115: Estou numa montanha-russa que só vai para cima.
Frase 116: O único dom que me salva é a distração.
Frase 117: Do que os humanos são capazes?
Frase 118: Com um sorriso desses você não precisa de olhos.
Frase 119: Tudo desaparecia quando ela estava dormindo.
Frase 120: A menina não o produzia com frequência.
Frase 121: Ninguém tinha a menor chance diante dele.
Frase 122: Uma menina feita de trevas.
Frase 123: Não torne as coisas piores.
Frase 124: Pensando, o burro morreu.
Frase 125: É isso que considera de importância nesse mundo?
Frase 126: É que o amor é essencialmente perecível.
Frase 127: Tinha beijado o papel devotamente.
Frase 128: Mas vamos começar com o para sempre.
Frase 129: O porto é o lugar mais seguro para um barco.
Frase 130: Algum conceito do quanto eu te amo?
Frase 131: Algo que você ainda não tenha.
Frase 132: Então o leão se apaixonou pelo cordeiro.
Frase 133: Eu gosto da noite.
Frase 134: Você está em cada pensamento que tenho.
Frase 135: Decida o que quer.
Frase 136: O destino é um parente elegante do acaso.
Frase 137: Então, só restava uma pessoa.
Frase 138: Você é mais importante que todos os outros.
Frase 139: Posso muito bem fazer o serviço completo.
Frase 140: Seja sempre forte.
Frase 141: Se amor é cego, nunca acerta o alvo.
Frase 142: Todo mundo deveria ser aplaudido de pé.
Frase 143: Fugi do fogo para não me queimar.
Frase 144: Você é um ímã vivo.

Frase 145: Por trás de um sorriso pode haver um punhal.
Frase 146: Sou o melhor predador do mundo.
Frase 147: O certo é que eu sou e quero ser inconstante.
Frase 148: Ninguém é de fato inteiramente bom.
Frase 149: Minha vida era a meia-noite.
Frase 150: São oito e meia da manhã.
Frase 151: São duas da tarde.
Frase 152: São dez da noite.
Frase 153: Bom cabrito não berra.
Frase 154: Estou andando até aí.
Frase 155: Você me mordeu.
Frase 156: Estou me coçando muito.
Frase 157: Meu jardim possui muitas plantas.
Frase 158: Eu gosto de frutas.
Frase 159: Ele é vegetariano.
Frase 160: Oba.
Frase 161: Obe.
Frase 162: Obi.
Frase 163: Obo.
Frase 164: Obu.
Frase 165: Obus.
Frase 166: Roba.
Frase 167: Robe.
Frase 168: Robi.
Frase 169: Robo.
Frase 170: Robu.
Frase 171: Robus.
Frase 172: Ra.
Frase 173: Re.
Frase 174: Ri.
Frase 175: Ro.
Frase 176: Ru.
Frase 177: Rus.
Frase 178: Hua.
Frase 179: Hue.
Frase 180: Hui.
Frase 181: Huo.
Frase 182: Hu.
Frase 183: Hus.
Frase 184: La.
Frase 185: Le.
Frase 186: Li.
Frase 187: Lo.
Frase 188: Lu.
Frase 189: Lus.
Frase 190: Flo.
Frase 191: Fru.
Frase 192: Vaisol.