

**UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
CENTRO TECNOLÓGICO
DEPARTAMENTO DE ENGENHARIA ELÉTRICA
PROJETO DE GRADUAÇÃO**

DANIEL JESUS RIBEIRO

**DESENVOLVIMENTO DE UM *DATASET* DE SINAIS DE
ÁUDIO E UM MODELO DE ROTULAÇÃO ORIENTADO A
APLICAÇÕES DE RECONHECIMENTO DE FALA**

VITÓRIA
2020

DANIEL JESUS RIBEIRO

**DESENVOLVIMENTO DE UM DATASET DE SINAIS DE ÁUDIO E
UM MODELO DE ROTULAÇÃO ORIENTADO A APLICAÇÕES DE
RECONHECIMENTO DE FALA**

Parte manuscrita do Projeto de Graduação do aluno Daniel Jesus Ribeiro, apresentado ao Departamento de Engenharia Elétrica do Centro Tecnológico da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Engenheiro Eletricista.

Orientador: Prof. Dr. Jorge Leonid Aching Samatelo

VITÓRIA
2020

DANIEL JESUS RIBEIRO

**DESENVOLVIMENTO DE UM DATASET DE SINAIS DE ÁUDIO E UM
MODELO DE ROTULAÇÃO ORIENTADO A APLICAÇÕES DE
RECONHECIMENTO DE FALA**

Parte manuscrita do Projeto de Graduação do aluno Daniel Jesus Ribeiro, apresentado ao Departamento de Engenharia Elétrica do Centro Tecnológico da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Engenheiro Eletricista.

Aprovada em 30 de novembro de 2020.

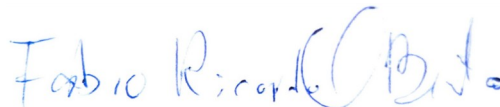
COMISSÃO EXAMINADORA:



Prof. Dr. Jorge Leonid Aching Samatelo
Universidade Federal do Espírito Santo
Orientador



Profa. Dr. Raquel Frizera Vassallo
Universidade Federal do Espírito Santo
Examinadora



Prof. MSc. Fabio Ricardo Oliveira Bento
Instituto Federal do Espírito Santo
Examinador

A Deus, meus pais e familiares, razões da minha dedicação.

AGRADECIMENTOS

Agradeço primeiramente a Deus pelas infinitas bênçãos e misericórdias derramadas sobre mim. Agradeço a meus pais, que durante toda a minha vida se fizeram presentes e dispostos a me auxiliar em tudo que fosse necessário. Agradeço a minha família pelas constantes orações e desejos de sucesso. Agradeço aos meus amigos que nesses últimos cinco anos foram os companheiros mais leais que eu poderia ter.

Cada um de vocês tem uma contribuição especial na minha chegada até aqui e afirmo com convicção que serei sempre grato a todos vocês.

Agradeço ao professor Jorge por aceitar o desafio de me orientar e embarcar comigo nesse desafio. A disponibilidade e empenho na troca de conhecimento foram marcantes e essenciais para o sucesso desse trabalho.

Agradeço ainda à Engenharia de Ferrovia Sudeste da Vale, através dos amigos Henrique Andrade, Eric Cabral e André Soares que acreditaram e me confiaram o desenvolvimento desse trabalho. O apoio e a segurança que me foi passada foram essenciais para alcançar os resultados obtidos.

Agradeço aos companheiros nesse projeto. Em especial ao João, Gabriel e Lucas, cujo apoio se mostraram fundamentais para os resultados obtidos.

RESUMO

O presente trabalho é um projeto de graduação do Curso de Engenharia Elétrica da Universidade Federal do Espírito Santo (UFES) e consiste na criação de um *dataset* de sinais de áudio e um modelo de rotulação com foco em aplicações de reconhecimento de fala em língua portuguesa. Este trabalho é parte de um projeto de parceria entre a Vale S.A. e a UFES no qual pretende-se desenvolver uma ferramenta de auxílio às auditorias dos canais de comunicação de áudio empregados na rotina operacional das ferrovias sob concessão da Vale. Embora existam vários *datasets* abertos disponíveis na internet, estes não apresentam necessariamente as mesmas características de áudio, bem como não compartilham do vocabulário próprio daquele empregado no contexto de comunicação ferroviária. Para tanto, um conjunto de dados que acolha o vocabulário do meio é essencial para o bom funcionamento de uma rede neural profunda voltada para o reconhecimento de fala. Utilizando os meios de comunicação empregados na ferrovia, foi desenvolvido um *dataset* cujos sinais de áudio compartilham das mesmas características construtivas daqueles que serão os dados de entrada para o *software* que se pretende desenvolver dentro do projeto de parceria entre a empresa e a universidade. Através de seu emprego, foi possível melhorar os resultados da rede neural do projeto em desenvolvimento em até 42%.

Palavras-chave: *Dataset*. Reconhecimento de fala. Redes Neurais Profundas.

ABSTRACT

This assignment consists in a project of graduation of the Electric Engineering Course of the Universidade Federal do Espírito Santo (UFES) and basis itself on the creation of a dataset of audio signals and a model of data labeling focused in Brazilian Portuguese speech recognition applications. This task is part of a partnership project between Vale S.A. and UFES, in which is intended to develop an auditing aiding tool to the communication audio channels used in the operational routine of the railways under concession of Vale. Although there are several open and available datasets online, these do not possess, necessarily the same audio characteristics or share the same vocabulary used within the railway communication field. Therefore, a dataset that holds the required vocabulary is vital to the proper functioning of a deep neural network applied to speech recognition. Making use of the tools employed in the railway communication system, it has been developed a dataset which audio signals share the same constructive characteristics of those that shall be the input data to the software that is being built in this partnership project between the enterprise and the university. Through its usage, it has been possible to improve the under development neural network results in 42%.

Keywords: Dataset. Speech recognition. Deep Neural Network.

LISTA DE FIGURAS

Figura 1 – Revoluções industriais.....	16
Figura 2 – Fluxo dos rótulos de áudio	40
Figura 3 – Fluxograma da revisão de rótulos	42
Figura 4 – Fluxograma de funcionamento da FGR	44
Figura 5 – Distribuição dos rótulos e arquivos de áudio	48
Figura 6 – Fluxograma do algoritmo de supressão de ruído	49
Figura 7 – Histogramas de duração, pace e quantidade de palavras	65
Figura 9 – <i>Wordcloud</i> das 100 palavras mais comuns	66
Figura 10 – Painel de acompanhamento no PowerBI.....	71
Figura 11 – Comparativo entre as <i>baselines</i>	73
Figura 12 – Testes envolvendo BRSDv1 e ACV <i>raw</i>	74
Figura 13 – Testes envolvendo BRSDv2 e ACV <i>raw</i>	75
Figura 14 – Testes envolvendo BRSDv1 e ACV <i>clean</i>	75
Figura 15 – Comparativos entre testes mistos	76
Figura 16 – Comparativo entre <i>datasets</i> próprios.....	77

LISTA DE GRÁFICOS

Gráfico 1 – Distribuição de áudios aptos.....	38
Gráfico 2 – Gráfico de amplitude do áudio 1.	51
Gráfico 3 – Foco do gráfico de amplitude do áudio 1.	51
Gráfico 4 – Gráfico de amplitude do áudio 2.	52
Gráfico 5 – FFT do áudio 1.	53
Gráfico 6 – FFT do áudio 2.	54
Gráfico 7 – MFCC do áudio 1.	55
Gráfico 8 – MFCC do áudio 2.	55
Gráfico 9 – MFCC limpos e gráficos e amplitude dos sinais.	56
Gráfico 10 – Distribuição dos áudios coletados.	58
Gráfico 11 – Distribuição dos áudios coletados por canal.	60
Gráfico 12 – Duração total e média por canal.	61
Gráfico 13 – Contagem de palavras e média por canal.	62
Gráfico 14 – <i>Pace</i> médio e distribuição dos canais nas categorias de <i>pace</i>	64
Gráfico 15 – Gráfico das maiores recorrências no dataset para palavras com mais de duas letras.....	67
Gráfico 16 – Quantidade de palavras por frequência.	68
Gráfico 17 – Relação da duração e palavras como vocabulário.	70

LISTA DE TABELAS

Tabela 1 – Áudios e percentual de longos por data.	36
Tabela 2 – Áudios e percentual de longos por canal.	38
Tabela 3 – Relação de áudios duvidosos, ruidosos, descartados e recuperados.....	46
Tabela 4 – Classificação dos áudios coletados.	59

LISTA DE QUADROS

Quadro 1 – Prós e contras das técnicas de rotulação de dados	22
Quadro 2 – <i>Datasets</i> de áudio em português brasileiro	24
Quadro 3 – Áudios gerados no console de gravação	33
Quadro 4 – Canais de comunicação analisados	37
Quadro 5 – Áudios manipulados na FGR	45
Quadro 6 – Arquivos de áudio usados na análise de ruído	50
Quadro 7 – Limites das categorias de <i>pace</i>	64
Quadro 8 – Matriz de correlação de características médias	66
Quadro 9 – Métricas dos <i>datasets</i> abertos	69
Quadro 10 – Cenários de teste	72

LISTA DE ABREVIATURAS E SIGLAS

ACV	Analisador de Comunicação de Voz
ASD	<i>Automatic Speech Diarization</i>
ASI	<i>Automatic Speech Identification</i>
ASR	<i>Automatic Speech Recognition</i>
CCE	Centro de Controle de Emergência
CCM	Centro de Controle de Manutenção
CCO	Centro de Controle de Operação
CCP	Centro de Controle de Pátio
CER	<i>Character Error Rate</i>
CSLU	<i>Center for Spoken Language Understanding</i>
EFVM	Estrada de Ferro Vitória a Minas
FFT	<i>Fast Fourier Transform</i>
FGR	Ferramenta de Gerenciamento de Rótulos
GSM	<i>Global System for Mobile Communication</i>
LDL	Liberação e Devolução de Linha
LVCSR	<i>Large Vocabulary Continuous Speech Recognition</i>
MFCC	<i>Mel-Frequency Cepstral Coefficients</i>
NG	<i>Noise Gating</i>
OOF	Oficial de Operação Ferroviária
PBI	<i>Phoneme-Based Indexing</i>
PCM	<i>Pulse-Code Modulation</i>
PR	Perfil de Ruído
RNA	Redes Neurais Artificiais
SS	<i>Spectral Subtraction</i>
TC	Torre de Controle
UFES	Universidade Federal do Espírito Santo
WPM	<i>Words Per Minute</i>

SUMÁRIO

1	INTRODUÇÃO	15
1.1	Motivação Pessoal	15
1.2	Indústria 4.0	16
1.3	Sistemas de Comunicação Ferroviária	17
1.4	Justificativa	18
1.5	Objetivos	19
1.5.1	Objetivo Geral.....	19
1.5.2	Objetivos Específicos.....	19
1.6	Escopo	19
2	FUNDAMENTAÇÃO TEÓRICA	20
2.1	Introdução	20
2.2	Redes Neurais Artificiais	21
2.3	Rotulação de Dados	21
2.3.1	Rotulação Interna	22
2.3.2	Outsourcing	23
2.3.3	Crowdsourcing	23
2.3.4	Empresas Especializadas de Outsourcing	23
2.3.5	Rotulação Sintética	24
2.3.6	Programação de Dados.....	24
2.4	Modelos de Bancos de Dados	24
2.4.1	CSLU: Spoltech Brazilian Portuguese	25
2.4.2	Sid	25
2.4.3	VoxForge	25
2.4.4	LapsBM1.4.....	26
2.4.5	CETUC.....	26
2.4.6	Compatibilidade dos Modelos apresentados	26
2.5	Técnicas de Supressão de Ruído	27
3	COMUNICAÇÕES DE RÁDIO	29
3.1	Postos de Trabalho	29
3.1.1	Centro de Controle de Pátio	29

3.1.2	Centro de Controle Operacional	29
3.1.3	Centro de Controle de Manutenção	30
3.1.4	Centro de Controle de Emergência	30
3.1.5	Torres de Controle	30
3.1.6	Oficial de Operação Ferroviária.....	30
3.1.7	Oficina.....	31
3.1.8	Maquinista.....	31
3.2	Dispositivos de Comunicação.....	31
3.2.1	Características do Áudio	32
3.2.2	Console de Gravação	33
4	METODOLOGIA E ETAPAS DE DESENVOLVIMENTO.....	34
5	DADOS DE ÁUDIO.....	36
5.1	Coleta de dados.....	36
5.1.1	Canais de Comunicação	37
5.2	Escolha do Método de Rotulação.....	39
5.3	Rotulação dos Dados.....	40
5.4	Rotina de Revisão de Rótulos em Lotes	41
5.5	Ferramenta de Gerenciamento de Rótulos.....	42
5.5.1	FGR Aplicada aos Rótulos.....	43
5.5.2	FGR Aplicado aos Áudios	44
5.6	Descarte de Áudios.....	45
6	FERRAMENTA DE SUPRESSÃO DE RUÍDO	47
6.1	Método de Supressão de Ruído.....	47
6.2	Algoritmo de Supressão de Ruído Aplicado a Ficheiros	48
6.3	Análise Qualidade de Remoção de Ruído	49
6.3.1	Amplitude.....	50
6.3.2	FFT.....	52
6.3.3	MFCC.....	54
6.3.4	Audição	56
6.4	Aplicabilidade.....	57
7	ANÁLISE DOS DADOS.....	58

7.1	Métricas.....	60
7.1.1	Duração	61
7.1.2	Contagem de palavras	62
7.1.3	Speech Pace.....	63
7.1.4	Matriz de Correlação.....	65
7.2	Vocabulário.....	66
7.3	Comparativo Outros <i>Datasets</i>	69
7.4	<i>PowerBI</i>.....	70
8	APLICAÇÃO EM UMA REDE NEURAL	72
9	CRIAÇÃO DE UM ROTEIRO	78
10	CONCLUSÕES E PROJETOS FUTUROS	79
	REFERÊNCIAS BIBLIOGRÁFICAS.....	81
	APÊNDICE A – ROTEIRO DE ENUNCIÇÃO.....	85

1 INTRODUÇÃO

1.1 Motivação Pessoal

Durante o processo de escolha de tema para o projeto de graduação foram observados vários fatores, de forma a escolher e estudar algo que pudesse não apenas agregar conhecimento, mas suprir carências e trabalhar habilidades importantes no meio industrial brasileiro.

Por não haver optado por uma ênfase em computação, surgiu um interesse inicial na área. Percebendo uma escassez de disciplinas envolvendo algoritmos e linguagens de programação na matriz curricular obrigatória do curso, a área de computação se tornou cada vez mais atrativa para a área de estudo do projeto de pesquisa.

Durante o período de estágio realizado na Vale S.A., foi possível perceber a demanda do mercado por profissionais capazes de lidar com grandes quantidades de dados e desenvolver sistemas inteligentes. Com a recente difusão do conceito de Indústria 4.0 (MINISTÉRIO DA INDÚSTRIA, COMÉRCIO E SERVIÇOS, 2019), um novo termo ganha certo destaque no mercado: Cientista de Dados. Embora a proposta não aborde diretamente as atribuições do referido profissional, a linha de pesquisa e ferramentas de trabalho elencados neste documento sugerem que será criada certa familiaridade com o ambiente e ferramentas utilizadas. Assim ocorrendo, existe maior facilidade de assimilação de futuros conhecimentos ligados à área.

Em reuniões da equipe da Engenharia de Operação da Engenharia Ferroviária da Estrada de Ferro Vitória a Minas (EFVM) manifestou-se a necessidade de desenvolver uma ferramenta de suporte à auditoria das comunicações de voz. Uma vez familiarizado com a temática de Redes Neurais, tema abordado no projeto apresentado na disciplina de Projeto Orientado (docência do Prof. Dr. Moisés Renato Nunes Ribeiro), surgiu a ideia de uma parceria entre a empresa e a universidade para trabalhar na ferramenta especificada.

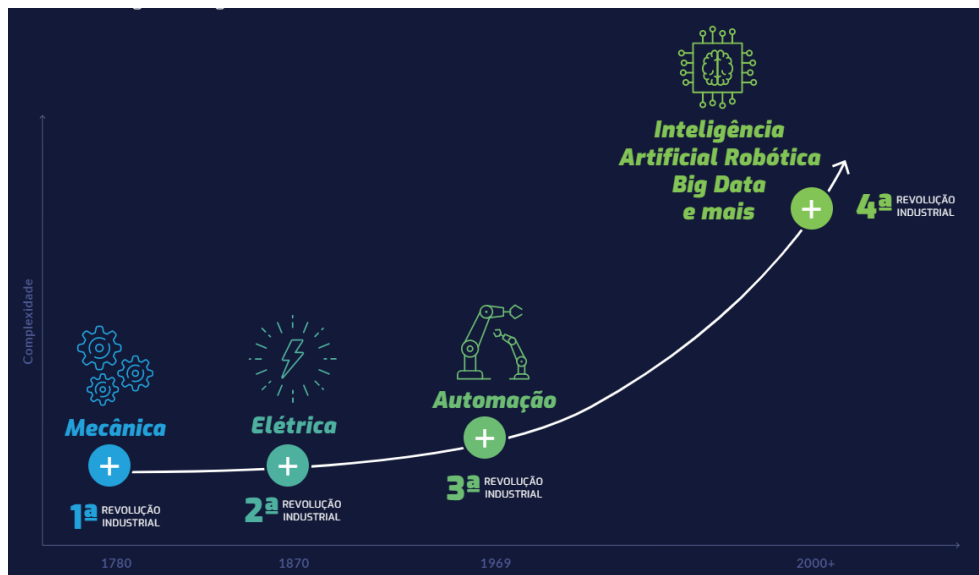
A escolha do tema do projeto, após avaliar a relevância dos fatos citados, tornou-se clara. A linha de pesquisa selecionada contempla todos os requisitos propostos: supre carências em relação à formação básica em Engenharia; agrega conhecimentos, uma vez que permite o aprofundamento em áreas distintas das ênfases escolhidas pelo autor; desenvolve habilidades

importantes para os profissionais atuantes no cenário industrial brasileiro, ao mesmo tempo que permite o desenvolvimento de uma demanda oriunda de seu próprio ambiente de atuação.

1.2 Indústria 4.0

Para uma definição do termo, de acordo com Anderl e Strang (2014, apud NETO et al., 2018, p. 1381), “essencialmente, o termo Indústria 4.0 tem por objetivo identificar iniciativas e tecnologias que buscam melhorar as cadeias de valor em todas as diversas etapas de fabricação de um produto” e “a quarta revolução industrial, [...], se caracteriza, por um conjunto de tecnologias que permitem a fusão do mundo físico, digital e biológico” (MINISTÉRIO DA INDÚSTRIA, COMÉRCIO E SERVIÇOS, 2019). Isso é possível por meio da implementação de diversos sensores em diferentes pontos de uma malha de controle, por exemplo, mas também pode ser alcançado com uma análise coerente e exaustiva de um grande volume de dados gerados por sistemas convencionais (que não se enquadram na definição de Indústria 4.0). A Figura 1 traça uma relação de complexidade com o momento de início de cada revolução tecnológica já registrada, avaliando ainda sua principal inovação.

Figura 1 – Revoluções industriais



Fonte: Ministério da Indústria, Comércio e Serviços (2019).

1.3 Sistemas de Comunicação Ferroviária

A segurança de uma ferrovia é baseada, entre outros elementos, na comunicação entre os maquinistas e os controladores de tráfego. Eles são responsáveis, respectivamente, pela condução da composição na malha férrea e da reserva dos circuitos (menor unidade de uma ferrovia que corresponde a um trecho de aproximadamente sete quilômetros) para circulação de cada trem, observando a densidade de fluxo, restrições operacionais de velocidade, bem como condições de segurança. Para que haja segurança na operação de todos os mais de 900 quilômetros de ferrovia sob concessão da Vale S.A no trecho de Vitória a Minas Gerais (BRASIL, 2018), é necessário que os controladores de tráfego estejam sempre ao alcance dos maquinistas. Suprindo essa necessidade, foi implantado um sistema de rádio que permite a comunicação entre os operadores e estabelece uma série de protocolos, como cortejo e palavras para indicar o fim de uma mensagem ou a espera por uma resposta.

Devido à necessidade de garantir a integridade das comunicações realizadas na ferrovia e cumprindo determinações internas, todas as conversas efetivadas nesse sistema são gravadas. Em posse dessas gravações, existe uma equipe de auditores responsável por avaliar se todos os protocolos estão sendo cumpridos e se há algum ponto de melhora quanto à intensidade vocal, dicção e vocabulário empregado.

Assim como utilizado na Vale S.A., as demais empresas do ramo seguem uma estrutura similar nos procedimentos de comunicação, como pôde ser averiguado pelos equipamentos de rádio embarcados nas locomotivas que fazem rotas em ferrovias sob concessão de outras empresas.

Buscando reduzir a quantidade de pessoal empregado nos procedimentos de auditoria, surgiu demanda por um sistema para auxiliar o auditor. Dentre as características idealizadas pela equipe de Engenharia Ferroviária da Vale, podem-se ressaltar: capacidade de identificação de interlocutores e capacidade de transcrição da gravação em texto (gerando um texto onde é indicado o falante, tal como a frase dita). Por definição, tais necessidades implicam no desenvolvimento dos seguintes sistemas: Reconhecimento Automático de Fala (ASR, do inglês *Automatic Speech Recognition*), Identificação Automática de Fala (ASI, do inglês *Automatic Speech Identification*) e Segmentação Automática de Fala (ASD, do inglês *Automatic Speech Diarization*).

Para o desenvolvimento da ferramenta de auditoria denominada Analisador de Comunicação de Voz (ACV), foi elaborada uma parceria entre a empresa Vale S.A. e a UFES, por meio do Departamento de Engenharia Elétrica. A iniciativa sugere a utilização de Redes Neurais Profundas e tecnologias compatíveis com a Indústria 4.0 para resolução do problema apresentado, sendo este o primeiro sistema de suporte a auditoria aplicado a ferrovias a nível nacional e mundial. Tendo em vista a grande complexidade do projeto, este foi dividido em projetos de mestrado, graduação e iniciação científica, respeitando o escopo de cada um dos níveis de estudo, bem como o grau de dificuldade proposto a cada uma das partes envolvidas no projeto. Este trabalho é um dos desdobramentos dessa parceria.

1.4 Justificativa

Apesar da eficácia do atual modelo de auditoria de comunicações utilizado na EFVM, é sempre de interesse da empresa garantir que os procedimentos nos quais há manuseio de equipamentos pesados sejam tão seguros quanto possível. Dessa forma, o ACV se apresenta como uma ferramenta que pode não apenas elevar o nível de confiabilidade da ferrovia (por meio de uma auditoria ininterrupta), mas também melhorar a competitividade da empresa ao possibilitar uma melhor distribuição da sua força de trabalho em atividades de maior interesse da companhia.

A necessidade de construir um *dataset* de áudios justifica o presente trabalho. Uma vez que os dados estejam rotulados apenas com as informações de interesse à aplicação, o treinamento dos modelos de aprendizado será factível e mais susceptível ao sucesso. Embora seja possível obter *datasets* públicos disponíveis *online* que podem ser usados para o treinamento de redes neurais, não existe certeza que estes sinais de áudio possuem as mesmas características construtivas e contemplam o mesmo vocabulário utilizado no meio ferroviário. Além disso, fazendo uso de áudios oriundos do sistema que receberá a aplicação final, é garantida a compatibilidade de características dos sinais de áudio e vocabulário, bem como ruído próprio do canal.

1.5 Objetivos

1.5.1 Objetivo Geral

O objetivo deste trabalho é a construção de um *dataset* de sinais de áudio de rádio da EFVM. Tal *dataset* será aplicado nas etapas de treinamento e teste de modelos de aprendizado baseados em redes neurais profundas, orientadas nas tarefas de reconhecimento de fala e identificação de diálogos. O modelo de rotulação de dados empregado na construção deste *dataset* será a base utilizada nas demais etapas do projeto ACV.

1.5.2 Objetivos Específicos

Para que o objetivo geral pudesse ser atingido, alguns objetivos específicos precisaram ser alcançados. Dentre estes, destacam-se:

- Desenvolver um *dataset* representativo e compatível com os modelos públicos;
- Definir e implantar um método de rotulação de dados aplicável ao contexto do projeto;
- Obter a melhor representatividade das grandezas dos sinais de áudio analisadas pela rede neural.

1.6 Escopo

O escopo deste trabalho consiste em desenvolver um *dataset* de áudio baseado nas comunicações de rádio da EFVM, bem como desenvolver as ferramentas para gerenciamento, supervisão e análise dos rótulos.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Introdução

Redes Neurais vêm sendo aplicadas cotidianamente na resolução de problemas da indústria moderna, como mostram Monteiro, Carneiro e Moreira (2012), Antoneli e Neitzel (2016) e Torres Júnior, Machado e Barreti (2006). De acordo com Sen, Dutta e Dey (2019, p. viii), na atual era de informação tecnológica, existe um crescimento exponencial do volume de dados gerados pelas mais variadas fontes. Para uma pesquisa eficiente e metodológica, é interessante que a informação possa ser facilmente obtida dentro de um processo de tomada de decisão. Seguindo a linha de raciocínio proposta pelos autores supracitados, é importante perceber que parte do volume de dados criado a cada dia é em forma de áudio ou vídeo.

No tocante à análise das informações contidas em um arquivo de áudio, Sen, Dutta e Dey (2019, p. 1) salientam que as duas principais abordagens para indexação de um sinal de áudio são dadas por meio do uso de um reconhecimento de fala de grande vocabulário (LVCSR, do inglês *large vocabulary continuous speech recognition*) ou de uma indexação orientada a fonemas (PBI, do inglês *phoneme-based indexing*). Para ambos os casos, no entanto, é necessário que o arquivo de áudio seja o mais claro possível. Para o LVCSR é ainda mais evidente a necessidade de tratamento do sinal de áudio, uma vez que neste modelo de indexação, a informação em áudio é transcrita, e, a partir desse ponto, esta é indexada como um texto. Numa indexação orientada a fonemas, a clareza e nitidez do áudio são de maior representatividade, devido ao uso de um dicionário de fonemas predeterminado que funciona como referencial para a conversão de textos em voz.

Assim, pode-se argumentar que os dois modelos trabalham com propostas opostas: o LVCSR é processado no âmbito textual, enquanto o PBI é avaliado no domínio de áudio. Seguindo a linha de pesquisa à qual este trabalho reporta conhecimento, é mais valoroso estudar e produzir dados para o primeiro modelo de indexação, uma vez que o projeto lidará também com a transcrição dos sinais de áudio.

2.2 Redes Neurais Artificiais

Haykin (2009, p. 2) diz que uma Rede Neural Artificial (RNA) é um processador altamente paralelizado e distribuído, constituído de unidades de processamento simples com uma propensão natural para armazenar experiências e torná-las disponíveis para uso, assemelhando-se ao cérebro humano nos seguintes aspectos: capacidade de aprendizado do ambiente por meio de processos de assimilação e pelos pesos sinápticos que são usados para armazenar tais experiências. Segundo Fischler e Firschein (1987 apud HAYKIN, 2009, p. 24), conhecimento é uma informação ou modelo armazenado e utilizado por uma pessoa ou máquina para interpretar, prever e interagir apropriadamente com o mundo externo. Unindo estes dois conceitos, uma RNA é um modelo algorítmico cuja finalidade é aprender a interagir com uma série de entradas de informação fornecendo saídas (respostas) coerentes.

Ainda segundo Haykin (2009, p. 3), uma das vantagens de utilizar uma RNA é poder trabalhar com mapeamento de entradas e saídas. Tal mapeamento faz uso de um processo de aprendizado supervisionado, através do qual são atribuídos valores aos pesos de uma RNA por meio de um algoritmo de otimização que utiliza um determinado conjunto de dados rotulados como exemplos de treinamento.

Em todo processo de aprendizagem humana, a interação social e a mediação do outro têm fundamental importância. Na escola, pode-se dizer que a interação professor-aluno é imprescindível para que ocorra o sucesso no processo ensino-aprendizagem (LOPES, 2017, p. 4).

Logo, é notável uma similaridade dos processos de aprendizagem humanos aos procedimentos de aprendizagem de uma máquina.

Evidencia-se, portanto, para o uso de uma rede neural, a necessidade de um conjunto de dados para treinamento da rede com aprendizado supervisionado.

2.3 Rotulação de Dados

Rotulação de dados (do inglês *data labeling*) é o procedimento de relacionar dados de entrada e de saída para que o modelo tenha exemplos que permitam definir os valores para os pesos do modelo em tempo de treinamento. “O principal desafio para uma equipe de cientistas de dados é decidir quem será responsável pela associação dos rótulos, estimar o tempo necessário e quais

ferramentas são mais adequadas” (ALTEXSOFT, 2018, tradução nossa). Além disso, de acordo com Rochman (2019), 80% do tempo de um projeto de Aprendizado de Máquinas é empregado em procedimentos de rotulação, limpeza e outras atividades de melhoria do conjunto de dados, de modo que apenas 20% do tempo seja empregado no desenvolvimento e treinamento do modelo.

A técnica de rotulação de dados a ser empregada varia, naturalmente, com a complexidade do problema e com a quantidade de recursos pessoais, computacionais e financeiros disponíveis. O Quadro 1 elenca as técnicas disponíveis para rotulação de dados, bem como seus prós e contras.

Quadro 1 – Prós e contras das técnicas de rotulação de dados

Técnica de Rotulação	Prós	Contras
Rotulação Interna	<ul style="list-style-type: none"> • Resultados previsíveis • Alta acurácia dos dados rotulados • Possibilidade de acompanhar o progresso 	<ul style="list-style-type: none"> • Procedimento demorado
<i>Outsourcing</i>	<ul style="list-style-type: none"> • Possibilidade de avaliar as habilidades dos envolvidos 	<ul style="list-style-type: none"> • Necessidade de organização do fluxo de trabalho
<i>Crowdsourcing</i>	<ul style="list-style-type: none"> • Redução de custos • Resultados rápidos 	<ul style="list-style-type: none"> • Pode reduzir a qualidade do trabalho
Empresas Especializadas de <i>Outsourcing</i>	<ul style="list-style-type: none"> • Garantia de qualidade 	<ul style="list-style-type: none"> • Alto valor de investimento
Rotulação Sintética	<ul style="list-style-type: none"> • Menos restrições no uso de dados sensíveis • Dados de treinamento sem lacunas e descasamentos • Eficácia de custo e tempo 	<ul style="list-style-type: none"> • Alto poder de processamento requerido
Programação de Dados	<ul style="list-style-type: none"> • Automatização • Resultados rápidos 	<ul style="list-style-type: none"> • <i>Dataset</i> de qualidade inferior

Fonte: AltexSoft (2018).

Nota: Tradução nossa.

2.3.1 Rotulação Interna

A rotulação interna é a técnica mais indicada para situações nas quais haja disponibilidade de pessoal e de tempo. Trazendo resultados sólidos e alta acurácia dos dados, a RNA em questão poderá alcançar resultados bem satisfatórios, uma vez que a quantidade de *mismatches* é praticamente quase nula. A possibilidade de acompanhar o progresso do projeto é interessante, visto que, no geral, existem prazos a serem cumpridos.

Buscando reduzir o tempo empregado nesta técnica, surgem práticas que tornam o procedimento mais rápido, embora possam prejudicar a qualidade do trabalho. Segundo AltexSoft (2018), é possível fazer uso de um sistema semi-supervisionado para agilizar a construção do *dataset*. Em outra vertente, Ochi, Dias e Soares (2004) propõem o uso de técnicas de *clustering* de dados, que seria uma forma de agrupar dados similares entre si e, em seguida, lidar com os subconjuntos de dados.

2.3.2 Outsourcing

“Uma das maneiras de acelerar o processo de rotulação de dados é procurar empregados esporádicos em diversos *sites* de recrutamento, *freelance* e redes sociais” (ALTEXSOFT, 2018, tradução nossa). Embora seja mais barato que alocar uma equipe de especialistas e engenheiros, pode comprometer a qualidade da rotulação, além de oferecer riscos no caso de informações sensíveis.

2.3.3 Crowdsourcing

Crowdsourcing é uma técnica bem semelhante à *outsourcing*. Seu diferencial consiste na força de trabalho empregada e no tempo de execução. AltexSoft (2018) destaca duas empresas no cenário mundial que prestam este tipo de serviço: *Amazon Mechanical Turk* e *Clickworker*. Ambas as empresas atuam de forma bem similar e baseiam seus serviços na utilização de milhares de pessoas espalhadas pelo mundo focadas na rotulação de dados de seus contratantes. O emprego de diversas pessoas na rotulação de dados, no entanto, pode trazer uma margem de subjetividade nos rótulos, prejudicando a confiabilidade do *dataset*.

2.3.4 Empresas Especializadas de Outsourcing

Empresas especializadas surgem no mercado mundial como uma alternativa para soluções de *crowdsourcing*. Embora seja levemente mais demorada, a qualidade da rotulação de dados é perceptivelmente mais garantida. Enquanto no *crowdsourcing* é contratada uma empresa para desempenhar a tarefa, em empresas especializadas de *outsourcing* o cliente define a quantidade de especialistas que serão envolvidos na rotulação de dados, segundo AltexSoft (2018).

2.3.5 Rotulação Sintética

“Rotulação sintética consiste na geração de dados similares a dados reais no tocante a parâmetros essenciais definidos pelo usuário. As informações sintéticas são produzidas por um modelo gerador que é treinado e validado em um conjunto de dados originais” (ALTEXSOFT, 2018, tradução nossa). Tal modelo é aplicável em situações em que o comportamento é de interesse do estudo, mas a confidencialidade dos dados é protegida. *Fintechs* e planos de saúde são dois exemplos de aplicação da técnica. Uma das dificuldades encontradas é a necessidade de alto poder computacional para geração dos dados.

2.3.6 Programação de Dados

Enquanto as técnicas já citadas dependem da ação humana, a programação de dados permite que os rótulos sejam associados sem que haja nenhum tipo de interação humana, como aborda AltexSoft (2018). Embora seja sabido que a acurácia é prejudicada, ainda assim é vista como uma alternativa válida, principalmente em aplicações que aceitam uma fraca supervisão de modelos finais de alta qualidade.

2.4 Modelos de Bancos de Dados

“Para a construção de um sistema de reconhecimento de fala é necessário o uso de um *dataset* grande e diversificado. Essa tarefa, no entanto, se apresenta desafiadora pelo fato de não existirem muitas opções de *dataset* disponíveis para uso” (QUINTANILHA, 2017, p. 73, tradução nossa). Além dessa dificuldade, têm-se ainda dois modelos de *dataset* aplicáveis para esta tarefa: baseados em palavras (LVCSR) e baseados em fonemas (PBI). Embora haja a possibilidade de uma rede neural ser treinada para trabalhar com ambos modelos de dados, reduz-se o esforço de treinamento e melhora-se a acurácia quando se utiliza apenas um deles.

Nesta Seção serão apresentados quatro *datasets* voltados para o português brasileiro. Algumas de suas características de interesse são sintetizadas no Quadro 2.

Quadro 2 – Datasets de áudio em português brasileiro

<i>Dataset</i>	<i>Distribuição</i>	<i>Falantes</i>	<i>Enunciados</i>	<i>Indexação</i>
CSLU	Pago	477	8080	LVCSR/PBI *
Sid	Gratuito	72	5777	LVCSR
VoxForge	Gratuito	+111	4090	LVCSR
LapsBM1.4	Gratuito	35	700	LVCSR
CETUC	Gratuito	100	100000	LVCSR

Fonte: Quintanilha, Biscainho e Netto (2020) e Quintanilha (2017).

Nota: * Nem todos os áudios possuem as duas versões de indexação.

2.4.1 CSLU: Spoltech Brazilian Portuguese

Segundo Quintanilha (2017), o *dataset* do Centro para Compreensão de Língua Falada (CSLU, do inglês *Center for Spoken Language Understanding*) compreende 8080 enunciados de 477 falantes diferentes, contemplando diversas regiões e sotaques do Brasil. No entanto, seu uso se torna inviável devido aos altos valores de investimentos necessários para a sua utilização.

2.4.2 Sid

O *dataset* Sid é baseado em 72 falantes de faixa etária entre 17 e 59 anos. Dentre esses falantes, 20 são mulheres. O texto enunciado aborda local de nascimento, idade, gênero, educação e profissão, conforme detalha Quintanilha (2017). Uma particularidade desse *dataset* é a existência de dígitos nos enunciados. Dentro do contexto das comunicações ferroviárias é constante o uso de números e dígitos, indicando uma vantagem no seu estudo.

O *dataset* é baseado no modelo LVCSR e todos os áudios possuem taxa de amostragem de 22,05 kHz em ambiente não controlado. Pouco mais de 5770 enunciados são transcritos palavra por palavra, mas sem indicação de início e fim.

2.4.3 VoxForge

O *dataset* apresentado pela VoxForge é fruto de um projeto “criado para coletar transcrições de fala para uso com programas de reconhecimento de voz livres e baseados em código aberto” (VOXFORGE, 2006). De acordo com Quintanilha (2017), este é o *dataset* mais heterogêneo. Tal característica é oriunda na própria concepção do banco de dados: qualquer pessoa pode

colaborar com o *dataset*. Ao acessar a página da VoxForge, é possível fazer a leitura de textos selecionados, e tal gravação já se incorpora ao conjunto de dados. A utilização de diferentes equipamentos para efetuar essa gravação, no entanto, insere diferentes taxas de amostragem, variando entre 16 kHz e 44 kHz.

Devido à existência de áudios de baixa frequência e gravações em ambientes não controlados, existe a similaridade com as comunicações ferroviárias ao passo que se verifica uma certa quantidade de ruído no ambiente do maquinista. O *dataset* conta com 4130 enunciados transcritos em concordância com o modelo LVCSR.

2.4.4 *LapsBM1.4*

O *dataset* LapsBM1.4 é utilizado pelo grupo Fala Brasil, da Universidade Federal do Pará. Sua estrutura é compatível com o modelo LVCSR. De acordo com Quintanilha (2017), o *dataset* é baseado nos enunciados de 35 pessoas diferentes, sendo 10 mulheres. Cada falante contribuiu com 20 enunciados, totalizando 700 enunciados gravados a 22,05 kHz em um ambiente não controlado.

2.4.5 *CETUC*

O *dataset* CETUC possui cerca de 145 horas de áudio enunciado por 50 homens e 50 mulheres. Cada um dos falantes pronunciou cerca de mil frases foneticamente balanceadas em um ambiente controlado com taxa de amostragem de 16 kHz (QUINTANILHA; BISCAINHO; NETTO, 2020).

2.4.6 *Compatibilidade dos Modelos apresentados*

Uma vez que o *dataset* CSLU: *Spoltech Brazilian Portuguese* é de distribuição paga, e até mesmo pelos fins deste trabalho, fica vetada a sua utilização neste. Os demais *datasets* compartilham características referentes ao *dataset* a ser produzido neste trabalho. Possuem uma variada taxa de amostragem entre si, excursando desde 16 kHz até 44 kHz, são obtidos em gravações em ambiente não controlado e possuem números dentro dos enunciados propostos.

Tais semelhanças agregam valor e possibilitam um estudo mais aprofundado deles a fim de determinar características de interesse no *dataset* a ser construído.

Para efeitos de teste, Quintanilha, Biscainho e Netto (2020) definem dois grupos principais de *dataset*: BRSDv1 e BRSDv2. Enquanto o segundo é constituído pelos conjuntos de dados listados no Quadro 2, o primeiro possui os mesmos conjuntos à exceção do CETUC. Tal divergência se dá pelo fato de que, embora Quintanilha, Biscainho e Netto (2020) façam uso de todos estes *datasets*, Quintanilha (2017) não tem a sua disposição o conjunto de dados CETUC. Optou-se por manter a nomenclatura adotada por Quintanilha, Biscainho e Netto (2020).

2.5 Técnicas de Supressão de Ruído

“A abordagem mais simples para redução de ruído ambiente em gravações de áudio é o uso de filtros de baixa e alta frequência” (BROWN; GARG; MONTGOMERY, 2018, p. 5011, tradução nossa). Tal processamento aplicado aos sinais de áudio garante que, uma vez sabida as frequências onde o sinal de interesse se encontra, pode-se suprimir as demais componentes de frequência do sinal. De acordo com Sotero Filho (2017), a voz humana possui componentes espectrais que alcançam até 12 kHz, entretanto, apenas as componentes de até 4 kHz possuem energia significativa para a formação do sinal de voz. Para o uso em sinais de áudio onde o ruído está alocado acima da marca de 4 kHz, um simples filtro do tipo passa alta pode ser suficiente para limpar o áudio.

Para situações nas quais o ruído se encontra na mesma faixa de frequência que o sinal de interesse, é necessário investigar outros métodos mais complexos de filtragem de sinais. Uma técnica apresentada por Boll (1979, apud BROWN; GARG; MONTGOMERY, 2018) é a subtração espectral de sinais (SS, do inglês *spectral subtraction*). Essa técnica serviu de base para um dos primeiros algoritmos para redução de ruído ambiente, o qual consistiu no uso de um perfil (um sinal contendo exclusivamente ruído) cujas frequências são analisadas, permitindo a subtração dessas componentes no áudio original. Dessa forma, obtém-se exclusivamente o sinal de interesse.

Brown, Gard e Montgomery (2018) salientam ainda um outro método similar ao supracitado. O Bloqueador de Ruído (NG do inglês, *noise gating*) que também trabalha com o conceito de

perfil de ruído, mas traz uma abordagem diferente, onde é possível estimar a intensidade do ruído e reduzir o volume de trechos ruidosos. “A eficácia dos métodos SS e NG é observada pelo seu emprego em editores de áudio amplamente difundidos como SoX e Audacity, respectivamente” (BROWN; GARG; MONTGOMERY, 2018, p. 5011, tradução nossa).

3 COMUNICAÇÕES DE RÁDIO

3.1 Postos de Trabalho

Como já esclarecido, as comunicações de rádio na EFVM são de caráter vital para o bom funcionamento da ferrovia bem como para a segurança dos empregados envolvidos ao longo dos mais de 900 km de linha férrea sob concessão da Vale S.A. Nesse contexto, existem alguns postos de trabalho cuja atividade, quando esclarecida, podem auxiliar na escolha do conjunto de dados de forma a trazer uma maior representatividade da realidade cotidiana do operador ferroviário. Para obtenção dessas informações, o autor, na condição de estagiário da empresa, esteve acompanhando a rotina destes postos de trabalho e angariando as informações sintetizadas.

3.1.1 Centro de Controle de Pátio

O Centro de Controle de Pátio (CCP) é responsável pelo trânsito de trens dentro dos pátios. No contexto aqui apresentado, é entendido como o pátio o conjunto de linhas férreas que se localizam no interior de uma unidade industrial da Vale S.A. As principais atividades do CCP são manobras de formação, desmembramento, classificação de vagões, posicionamento para descarga, além de entrega e retirada de vagões e locomotivas para as oficinas.

3.1.2 Centro de Controle Operacional

Ao Centro de Controle Operacional (CCO) é reservado o controle de circulação dos trens ao longo da linha férrea que não se encontra dentro de uma unidade industrial. É de sua responsabilidade administrar o fluxo dos trens, o cruzamento de veículos, observação de restrições de velocidade, orientações quanto à mudança de linha, acompanhamento de equipes trabalhando em campo através da abertura de LDLs (liberação e devolução de linha), sendo essa uma atividade de alta criticidade por envolver pessoas trabalhando sobre a linha férrea. Existem 4 postos no CCO que trabalham concomitantemente.

3.1.3 Centro de Controle de Manutenção

O Centro de Controle de Manutenção (CCM) é a unidade responsável pela manutenção da linha férrea. Sua atuação é imprescindível para o bom funcionamento da malha ferroviária. Sua principal atuação é no tocante às manutenções agendadas para via permanente, *housings* e equipamentos de sinalização. Sua principal (porém não única) interlocução é com os controladores do CCO.

3.1.4 Centro de Controle de Emergência

O Centro de Controle de Emergência (CCE) é responsável por atender as emergências que surgem ao longo da EFVM. Descarrilamento, abalroamentos atropelamentos e tombamentos são situações que, embora não sejam comuns, são atendidas por esse posto. Cabe a este centro ainda o acompanhamento, registro e provisionamento de medidas de controle para quaisquer situações anormais na linha férrea ou no seu entorno. Sua atividade é essencial e deve ser o mais eficiente possível, tendo em vista os possíveis agravamentos da demora no atendimento a uma emergência ou situações de risco. Sua solicitação é baixa, mas é importante que este seja percebido nos estudos aqui desenvolvidos.

3.1.5 Torres de Controle

As Torres de Controle (TC) são responsáveis pela formação da composição. É sua tarefa realizar o agrupamento de vagões e locomotivas que fazem parte do trem que irá circular na linha tronco. É importante que a TC observe algumas características dos veículos como, por exemplo, a integridade do sistema de freio, prevenindo assim acidentes ou falhas. Basicamente, as TC são subdivisões do pátio de acordo com o tipo de manobra e características da operação da região. As TC compõem o CCP.

3.1.6 Oficial de Operação Ferroviária

O Oficial de Operação Ferroviária (OOF) é o responsável em campo pela integridade da composição que é preparada pelas TC. Sua atividade foca na correta disposição dos veículos, bem como numa avaliação de sua integridade e no bom funcionamento através de testes de

eficácia que são performados durante a montagem da composição. Sua atividade é diretamente ligada ao maquinista e aos controladores.

3.1.7 Oficina

As oficinas como um todo são as áreas responsáveis pela manutenção dos veículos, sejam eles vagões ou locomotivas. Sua atividade na comunicação de rádio é mais voltada para sinalização e informações de ativos quando num contexto interno, e dados sobre manobras que acontecem nas linhas sob sua tutela, quando em comunicação com postos de trabalho de torres e centros.

3.1.8 Maquinista

O maquinista de viagem é o responsável imediato pela condução de uma composição. Sua atividade consiste em operar o trem de forma a obedecer às restrições impostas pela unidade de controle que gerencia o trecho onde esse se encontra (seja esse CCO, CCP ou CCE). Além disso, é de sua alçada permanecer em contato com os centros de controle, avisando-os caso haja alguma situação incomum ou diferente do esperado no trecho.

O maquinista de manobra atua dentro dos pátios e, no geral, se reporta ao CCP e às TC. Sua atividade é auxiliada pelo OOF na montagem da composição, bem como nas manobras de veículos, sejam esses vagões ou locomotivas. Suas comunicações possuem algumas particularidades que são importantes de serem notadas.

3.2 Dispositivos de Comunicação

Para efetuar a comunicação entre os diversos postos espalhados pela ferrovia, são utilizados equipamentos de rádio. Em situações corriqueiras, os controladores (pessoas que trabalham no CCP, CCO, CCE e TC) têm à sua disposição equipamentos de rádio fixos e pertencentes ao posto de trabalho havendo, portanto, para um mesmo aparelho, a rotatividade de controladores. Os maquinistas – sejam de viagem, sejam de manobra – têm uma situação similar. O equipamento de rádio por estes utilizado é embarcado no locomotiva. Dessa forma, uma mesma locomotiva pode receber diversos maquinistas e, portanto, diversos locutores. Em geral, é

comum que uma composição de viagem seja controlada por 4 maquinistas distintos ao longo de seu trajeto pela ferrovia.

3.2.1 Características do Áudio

Tipicamente, o *bit rate* de um arquivo de áudio é uma medida relativa à qualidade da gravação (BRIAN; SHI, 2009). Em termos gerais, determina a taxa de bits codificados a cada segundo e possui valor de referência de 13 kbps para conversas telefônicas e 64 a 128 kbps para faixas de música (MICROPYRAMID, 2017).

Um das principais aplicações das gravações de áudio é, sem dúvida, a distribuição e consumo de música. No âmbito musical, a experiência ao se ouvir uma canção é algo muito valorizado. Uma das ferramentas que se utiliza para aprimorar essa experiência é o uso de duas ou mais faixas de áudio sendo reproduzidas simultaneamente. Tal recurso permite a criação de uma sensação de movimento e de espaço através do balanço de volume de cada instrumento em cada uma das faixas. Classicamente, numa configuração estéreo, uma faixa é denominada esquerda e a outra direita. A essas faixas dá-se o nome de canal.

Sample rate é basicamente a taxa de amostragem do sinal. Vive-se um mundo analógico onde as grandezas são contínuas. No contexto da computação, entretanto, as grandezas são discretas. Para que o computador possa assimilar e processar informações oriundas do mundo real é necessário que haja uma discretização da grandeza através de uma amostragem periódica. Segundo Nyquist (1928), é necessário que a taxa de amostragem de um sinal contínuo seja de, no mínimo, o dobro da frequência máxima deste.

Uma importante característica dos sinais de áudio é a forma como as amostras são armazenadas e compactadas no arquivo de áudio. *Sample encoding*, ou simplesmente codificação de amostras, indica qual o *codec* que é utilizado para codificar e comprimir o arquivo de áudio. É válido ressaltar que a extensão do arquivo não traz essa informação intrinsecamente. Alguns dos *codecs* mais utilizados atualmente incluem *linear PCM*, *floating point*, *μ-law*, *A-law*, *DVI* (BAGWELL, 2013).

3.2.2 Console de Gravação

Por critérios internos, a Vale S.A. dispõe de dispositivos de gravação com a finalidade de auditar os processos de comunicação e gerar uma constante melhora nestes. Um ponto importante para o desenvolvimento desse projeto é o acesso à base de dados que contém tais gravações. Uma vez que foi concedido tal acesso ao autor, pôde-se analisar algumas características importantes dos arquivos de áudio gerados. O Quadro 3 apresenta as características dos áudios gerados no console de gravação.

Quadro 3 – Áudios gerados no console de gravação

Característica	Valor
<i>Bit rate</i>	13,0 kbps
Canais	1
<i>Sample rate</i>	8000 Hz
<i>Sample encoding</i>	GSM

Fonte: Produção do próprio autor.

Como pode ser percebido, o *sample encoding* dos áudios obtidos através do console de gravação não está listado entre os *codecs* mais populares. Isso se mostra como um fator dificultador, uma vez que não é garantida a capacidade de processamento desse sinal na maioria das aplicações de áudio. Outro ponto importante que vale ser ressaltado é a alta taxa de compressão imposta por esse *codec* (HUERTA; STERN, 1997). Isso dificulta o uso desse arquivo em processamento de sinais, mas, por outro lado facilita o seu armazenamento. Considerando o contexto de criação do *codec* (sistemas de telecomunicação) e o contexto de sua aplicação, é realmente válido que o arquivo seja otimizado para armazenamento, uma vez que até o presente momento, essa é a principal finalidade dos arquivos de áudio gerados.

4 METODOLOGIA E ETAPAS DE DESENVOLVIMENTO

O presente trabalho, quanto à sua natureza, trata-se de uma pesquisa aplicada, posto que os resultados obtidos serão utilizados não apenas em outros trabalhos, mas também em soluções industriais imediatas. Quanto aos seus objetivos, é classificado como uma pesquisa exploratória, uma vez que serão gerados conhecimentos sobre o assunto com o intuito de avançar no desenvolvimento de um projeto de maior abrangência. Do ponto de vista dos procedimentos técnicos, define-se como uma pesquisa experimental e também um estudo de caso. Por fim, o problema será abordado sob uma ótica qualitativa.

Como abordado em seções anteriores, este trabalho tem como objetivo desenvolver uma técnica de rotulação de dados, bem como definir um modelo de *dataset* contendo as principais informações utilizadas em RNA voltadas para o processamento de áudio. Em primeira instância, a principal atividade realizada foi dar seguimento aos estudos em Redes Neurais Artificiais em andamento, buscando absorver a maior quantidade de informação possível. Nesse contexto, a disciplina de Tópicos Especiais em Visão Computacional III, ministrada pelo orientador deste trabalho, apresentou-se como uma espécie de etapa zero na preparação para o desenvolvimento deste.

Em seguida, foi realizada uma avaliação profunda dos métodos de rotulação de dados apresentados na Seção 2.3, de forma a determinar qual das técnicas melhor se aplica ao caso estudado, levando em consideração a utilização de dados potencialmente sensíveis.

Dando prosseguimento, foi agrupado um conjunto de dados produzidos pelo próprio autor em parceria com a equipe de Auditoria do CCO/CCP da EFVM. Este conjunto de dados foi avaliado com o objetivo de perceber os ganhos obtidos na compreensão de fala por redes neurais profundas. Para estes testes, foi utilizada uma RNA com o intuito de apurar os resultados de cada modelo de *dataset* (públicos e próprios) no treinamento da rede, fazendo uma análise comparativa do resultado da transcrição. Foi também desenvolvida uma ferramenta de supressão de ruído preliminar com o intuito de melhorar a qualidade do sinal fornecido à rede neural e, a partir dessa, foi gerada uma versão alternativa do *dataset*. Dessa forma, é possível observar o comportamento dos *datasets* na utilização das RNAs que serão desenvolvidas ao

longo do projeto ao qual este trabalho reporta conhecimento e avaliar a qualidade da supressão de ruído.

Todas as etapas de trabalho foram devidamente documentadas de forma a permitir sua reprodução, além de detalhar de forma suficiente a esclarecer leitores leigos como foram obtidos estes resultados. Este documento explora todas as etapas e decisões tomadas na construção do *dataset*, além de detalhar o funcionamento dos algoritmos criados no gerenciamento e tratamento dos sinais obtidos. Elencando ainda as características necessárias para a construção do conjunto de dados, é possibilitado o desenvolvimento de *datasets* complementares a este, seja para fins de expansão de vocabulário, seja para inclusão de canais na análise. Uma outra possibilidade que é mapeada é a utilização dos métodos aqui descritos na criação de *dataset* orientado a outros contextos, até mesmo internos à Vale S.A., como comunicações portuárias, por exemplo.

5 DADOS DE ÁUDIO

5.1 Coleta de dados

O acesso à base de dados do console de gravação de áudio foi concedido pela equipe de tecnologia operacional da EFVM. Através do portal denominado AUTOVM foi possível escolher os dias e canais de interesse para a análise no presente trabalho. É válido ressaltar que os arquivos obtidos contemplam os mais diversos períodos do dia, de forma a abranger a maior quantidade possível de particularidades. Uma das estratégias que foi utilizada consistiu em focar a coleta de dados em períodos de início e término dos serviços realizados na linha férrea e trocas de turno. Nestas situações existe uma atuação mais constante dos controladores.

De acordo com a rotulação dos dados e com a relevância dos canais (representatividade e abrangência do vocabulário), em conjunto com a Engenharia de Operação da EFVM foram elencados os dias de maior interesse. A Tabela 1 exibe os dias que foram selecionados, bem como a quantidade de arquivos de áudio coletados.

Tabela 1 – Áudios e percentual de longos por data

Data de referência	Áudios	Longos	Áudios aptos	% Total
20 jan. 2020	733	196 (26,74%)	537	23,09%
16 fev. 2020	50	0 (0%)	50	2,15%
10 mar. 2020	637	146 (22,92%)	491	21,11%
12 mar. 2020	77	7 (9,09%)	70	3,01%
25 mar. 2020	147	22 (14,97%)	125	5,37%
29 mar. 2020	170	20 (11,76%)	150	6,45%
25 abr. 2020	312	43 (13,78%)	269	11,56%
12 mai. 2020	77	11 (14,29%)	66	2,84%
2 jun. 2020	175	11 (6,29%)	164	7,05%
12 jun. 2020	466	62 (13,3%)	404	17,37%
Total	2844	518 (17,02%)	2326	100,00%

Fonte: Produção do próprio autor.

Os dias que possuem uma representatividade mais elevada (acima de 15%) correspondem a dias cuja intensidade de tráfego na malha ferroviária foram significativos, ou dias em que as atividades de pátio se mostraram interessantes para registro (execução de atividades específicas). Os demais dias foram retirados de forma esporádica e sempre que os áudios disponíveis para rotulação já haviam sido rotulados e processados pelos transcritores ligados ao projeto.

Um dos critérios de exclusão aplicados aos arquivos extraídos do AUTOVM é a duração do áudio. Analisando os demais *datasets*, pode-se perceber uma predominância de enunciados curtos. Os *datasets* LapsBM1.4 e VoxForge por exemplo, apresentam faixas com duração variando, em sua maioria, entre três e seis segundos. Uma vez que não é usual que em um ambiente não controlado possa-se obter trechos significativos com tão pequena duração, optou-se por limitar a duração do áudio em 30 segundos. Dessa forma, os áudios que ultrapassam essa marca foram automaticamente desconsiderados como uma estratégia de manter a maior quantidade de similaridades construtivas entre o presente trabalho e os *datasets* públicos disponíveis.

5.1.1 Canais de Comunicação

O AUTOVM abrange uma grande quantidade de canais de comunicação que não necessariamente fazem parte do escopo deste trabalho. Através de análise junto à equipe de Engenharia de Operação, foram definidos os canais cujo conteúdo seria mais representativo, fornecendo um vocabulário mais próximo daquele observado no cotidiano. O Quadro 4 contempla os canais que foram utilizados.

Quadro 4 – Canais de comunicação analisados

Número do canal	Nome do canal
21	Posto 1 do CCO
24	Posto 2 do CCO
23	Posto 3 do CCO
55	Posto 4 do CCO
74	CCP de Laboreaux
40	CCE
41	CCM
91	Oficina de vagões
31	Torre A
32	Torre B
26	Torre C
33	Torre D
34	Torre E
46	Torre L

Fonte: Produção do próprio autor.

Foram selecionados quatro postos de trabalho do CCO. Por ser responsável pelas ações que ocorrem na linha tronco da ferrovia, o que implica em cobrir maiores distâncias de linha férrea, o CCO se torna um dos principais objetos de análise. O CCP de *Laboreaux*, bem como as torres

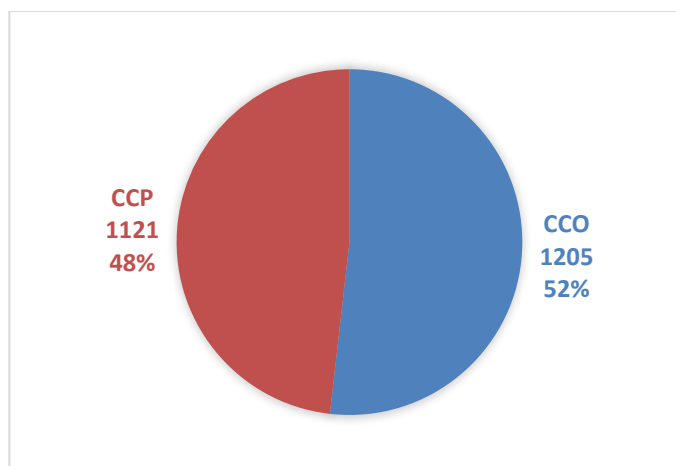
listadas (pertencentes ao CCP de Tubarão) e a Oficina de Vagões tem por objetivo ilustrar a rotina de manutenção dos pátios da ferrovia. Embora apareçam em maior quantidade que o CCO, estas contribuem individualmente com uma menor expressividade no volume de dados total. O CCM traz o vocabulário das rotinas de manutenção da ferrovia para o presente trabalho. Embora não haja muita recorrência é importante ter mapeado o vocabulário utilizado em tarefas afins. Por fim, o CCE traz a representatividade para as emergências. Embora raros acontecimentos tenham sido registrados durante o levantamento de dados, foram selecionados eventos pontuais para esclarecer a dinâmica da comunicação durante o atendimento às ocorrências. Uma vez que um dos maiores focos das auditorias são as emergências, é essencial que o CCE seja analisado.

Tabela 2 – Áudios e percentual de longos por canal

Canal	Áudios	Longos	Áudios aptos	% Total
Posto 1 do CCO	265	63 (23,77%)	202	8,68%
Posto 2 do CCO	502	108 (21,51%)	394	16,94%
Posto 3 do CCO	481	116 (24,12%)	365	15,69%
Posto 4 do CCO	350	106 (30,29%)	244	10,49%
CCP de Laboreaux	42	3 (7,14%)	39	1,68%
CCE	86	5 (5,81%)	81	3,48%
CCM	99	13 (13,13%)	86	3,70%
Oficina de vagões	36	1 (2,78%)	35	1,50%
Torre A	338	48 (14,2%)	290	12,47%
Torre B	139	11 (7,91%)	128	5,50%
Torre C	72	9 (12,5%)	63	2,71%
Torre D	156	10 (6,41%)	146	6,28%
Torre E	160	22 (13,75%)	138	5,93%
Torre L	118	3 (2,54%)	115	4,94%
Total	2844	518 (17,02%)	2326	100,00%

Fonte: Produção do próprio autor.

Gráfico 1 – Distribuição de áudios aptos



Fonte: Produção do próprio autor.

Para fins de comparação, os canais não referentes aos postos de trabalho do CCO serão englobados pelo CCP. Dessa forma, tem-se dados de dois postos distintos: CCO e CCP. Como o Gráfico 1 exhibe, a distribuição dos áudios aptos entre tais categorias é bem equilibrada. Tal distribuição é satisfatória pois denota o mesmo nível de importância para ambos tipos de canais de comunicação.

5.2 Escolha do Método de Rotulação

Avaliando os prós e contras das técnicas de rotulação de dados apresentadas na Seção 2.3, conclui-se que, por este trabalho ser parte de um projeto de parceria entre a UFES e uma empresa privada, não é interessante o compartilhamento de informações com empresas externas por questões de confidencialidade. Por isso, as opções envolvendo *outsourcing* e *crowdsourcing* são desconsideradas.

A rotulação sintética é interessante para a aplicação atual e se encaixaria no contexto. Entretanto, os dados analisados não são plenamente compatíveis com esse tipo de rotulação. Seria necessário conhecer profundamente os dados para que fosse possível trabalhar com esse método de forma a obter resultados consistentes. Embora seja descartado para o início das rotulações, a rotulação sintética permanece presente no horizonte de estudo. Uma vez tendo uma base de dados confiável e vasta, pode-se fazer uso desse conceito para expandir o volume de dados disponíveis através de técnicas de *data augmentation* (GANDHI, 2018).

Por se tratar de um *dataset* cuja aplicação é no reconhecimento de fala, é estritamente necessária uma alta acurácia dos dados rotulados. Partindo deste pressuposto, elimina-se também a possibilidade do uso de programação de dados como método de rotulação.

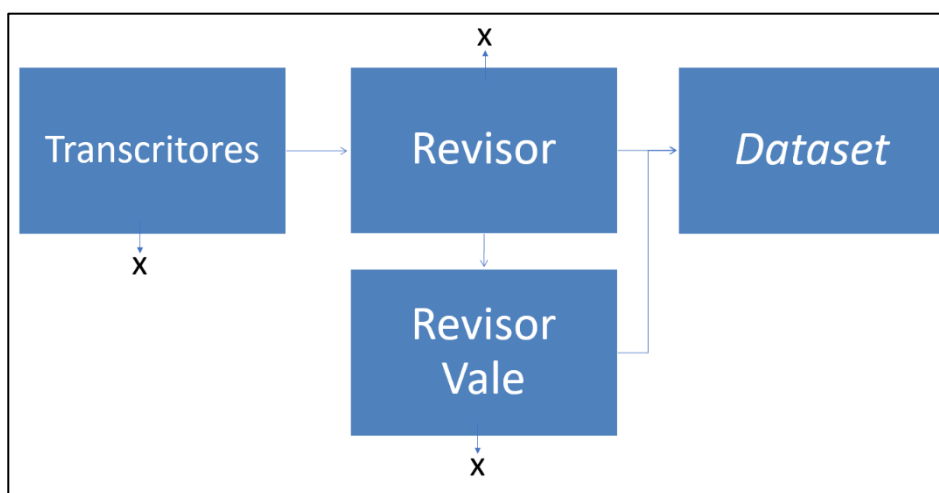
Por fim, têm-se a possibilidade do uso de rotulação interna. Embora não seja a forma de obtenção de resultados mais rápida, é o método mais interessante para aplicação neste trabalho. Observando os prós elencados no Quadro 1, pode-se concluir que tal técnica é a mais adequada. Resultados previsíveis, alta acurácia e possibilidade de acompanhamento de progresso são aspectos de grande valor que trazem maior confiabilidade ao *dataset*.

5.3 Rotulação dos Dados

Uma vez que o projeto tem um prazo determinado para ser concluído, pelo fato de a rotulação interna ser o método mais demorado, foram alocados dois alunos de iniciação científica (João Vitor Nunes e Gabriel Xavier) ligados ao projeto para darem suporte ao autor no processo de transcrição de áudios. Para que os dados mantivessem a qualidade, foi ministrado um treinamento a toda equipe do projeto para que estes fossem inteirados acerca das particularidades e protocolos de comunicação vigentes na ferrovia. Dessa forma, garante-se uma melhor compreensão do contexto dos enunciados, trazendo mais segurança aos transcritores, principalmente aos que não possuem experiência no contexto ferroviário.

A atividade desempenhada pelos transcritores, embora pareça simples, demanda muita atenção e esforço. Cada faixa disponibilizada deve ser escutada ao menos duas vezes. Caso haja qualquer dúvida em alguma palavra, deve-se reproduzir o arquivo novamente em velocidade reduzida. Persistindo a incerteza, o rótulo deve ser destacado como duvidoso e enviado ao autor que, na condição de revisor, por possuir experiência na operação ferroviária deve buscar compreender o áudio. Ainda assim, não havendo segurança quanto ao conteúdo, o revisor envia o arquivo para o instrutor do curso de comunicação ferroviária da empresa, aqui denominado Revisor Vale. O Revisor Vale por sua vez pode retornar o rótulo solucionado ou recomendar o seu descarte. Os transcritores e o revisor podem descartar um áudio em caso de ruído excessivo ou grande quantidade de palavras duvidosas. Na Figura 2 o **x** mostra a possibilidade de descarte.

Figura 2 – Fluxo dos rótulos de áudio



Fonte: Produção do próprio autor.

Para assegurar a qualidade e padronização da rotulação entre os três transcritores, algumas premissas foram definidas pelo autor no tocante à produção dos rótulos. A saber:

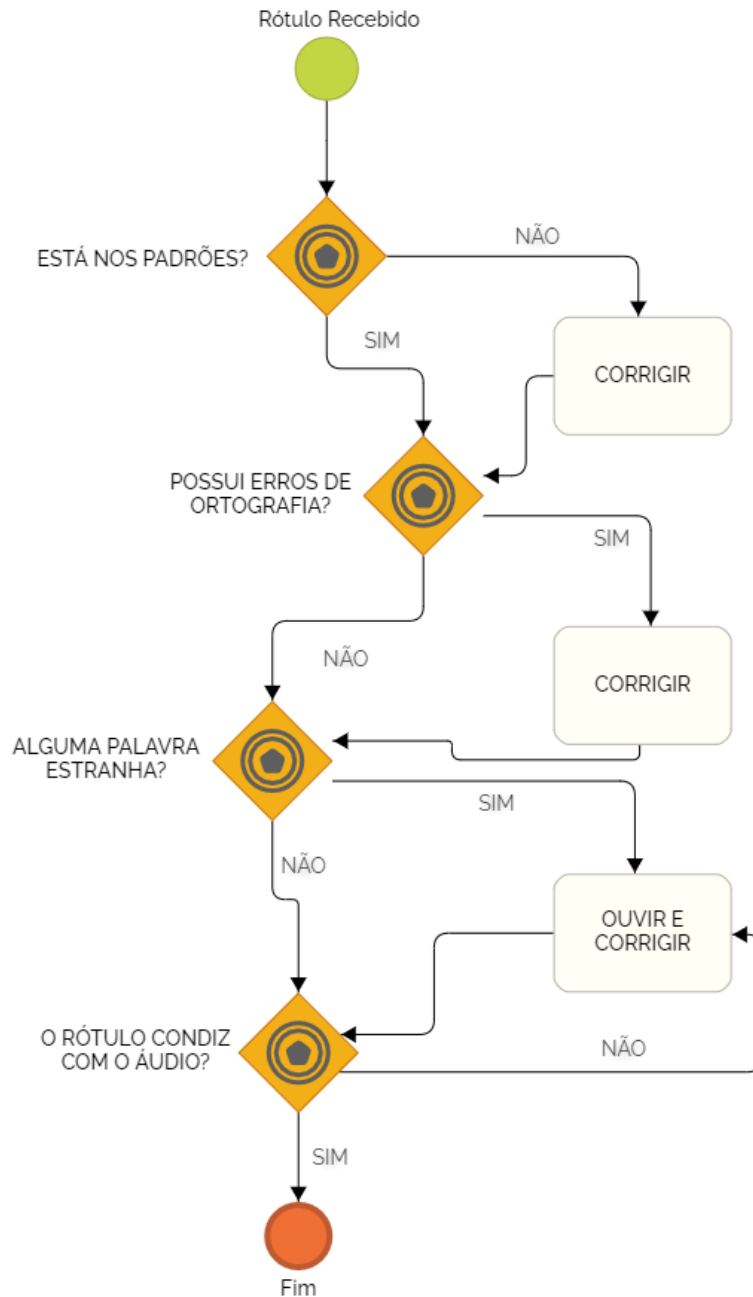
- A transcrição deve ser registrada em um arquivo de texto (extensão txt) homônimo à respectiva faixa de áudio;
- A transcrição deve ser exata;
- As palavras devem ser escritas mantendo-se a acentuação;
- Deve-se registrar a repetição de sílabas (gaguejos);
- Não se deve registrar símbolos de pontuação (interrogação, exclamação, vírgulas, pontos finais e afins);
- Deve-se registrar as siglas normalmente (CCO, CCP) e não silabicamente (ceceo, cecepe, por exemplo);
- Devem ser registradas as gírias utilizadas (passar vento ou shuntar, por exemplo);
- Deve-se registrar a forma como as palavras são ditas sem correção (brigado ou mineirin, por exemplo);
- Os números devem ser transcritos por extenso e nunca com algarismos;
- Os rótulos devem ser disponibilizados para o revisor através de um arquivo compactado no qual se possa diferenciar as faixas duvidosas, descartadas e precisas.

5.4 Rotina de Revisão de Rótulos em Lotes

Para garantir a conformidade e padronização dos rótulos, é necessária a atuação de um revisor com responsabilidade de avaliar se cada um dos rótulos está de acordo com os padrões listados na Seção 5.3. Dessa forma, certifica-se que todos os arquivos de texto tenham as mesmas formatações, características de conteúdo e particularidades. Além disso, cabe a este observar os rótulos em busca de palavras escritas em desacordo com a ortografia, erros de digitação, palavras estranhas ao contexto ferroviário e realizar as correções necessárias. Tal procedimento é ilustrado na Figura 3.

De acordo com a Tabela 2, tem-se 2326 faixas de áudio que cujas características se adequam ao contexto aqui apresentado. Revisar manualmente cada uma das faixas de áudio seria um retrabalho muito custoso ao revisor. Optou-se por criar uma ferramenta que permitisse a revisão em lotes dos rótulos fornecidos pelos transcritores. A Seção 5.5 trata deste tema.

Figura 3 – Fluxograma da revisão de rótulos



Fonte: Produção do próprio autor.

5.5 Ferramenta de Gerenciamento de Rótulos

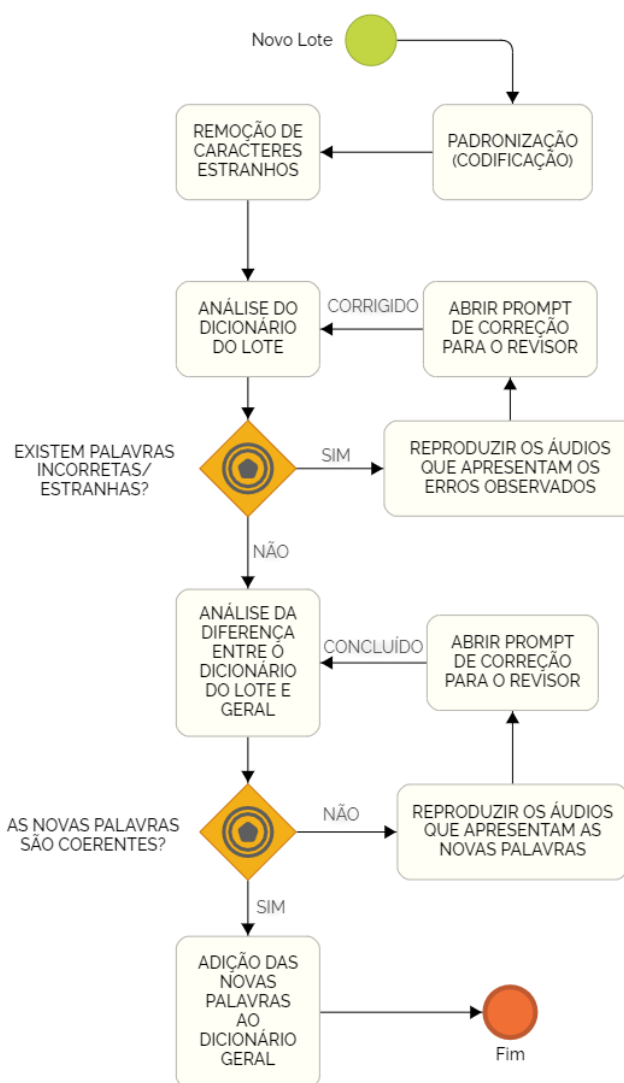
A Ferramenta de Gerenciamento de Rótulos (FGR) é uma aplicação desenvolvida em *Python* 3.8 que tem por objetivo auxiliar o revisor a analisar, corrigir e padronizar os arquivos de áudio e os rótulos recebidos pelos transcritores.

5.5.1 FGR Aplicada aos Rótulos

Na transcrição foram utilizados diferentes sistemas operacionais por parte dos transcritores. No momento de importação dos novos rótulos, a FGR padroniza automaticamente a forma de codificação de caracteres e remove *flags* características de alguns dos sistemas operacionais. Apesar disso, uma das características mais importantes da FGR é o seu papel na manutenção do dicionário do *dataset*. O dicionário, como o próprio nome sugere, é um objeto que reúne e contabiliza cada palavra que fora transcrita pelo menos uma vez. Através do uso do dicionário, assim que um novo lote de dados é recebido, pode-se analisar quais são as palavras estranhas ao vocabulário já indexado, permitindo uma análise mais focada em *outliers*, uma vez que as demais palavras são consideradas existentes pela aplicação. A Figura 4 detalha o funcionamento da aplicação.

Assim que é recebido um novo lote, é gerado um dicionário contendo todo o vocabulário apresentado nos rótulos. Tal dicionário é exibido ao revisor para que este possa elencar quais entradas devem ser revisadas. É válido ressaltar que a rotina do revisor, antes orientada ao rótulo, com uso da FGR passa a ser orientada à palavra. Após observar todas as palavras do dicionário, corrigir os erros de escrita e digitação e avaliar as palavras estranhas ao contexto, é exibido ao revisor uma lista contendo as palavras que não possuem cadastro no dicionário geral. Dessa forma, pode-se focar nas palavras que são novas e apresentam maior possibilidade de inconsistência. Assim que estas são avaliadas, o dicionário do lote é integrado ao dicionário geral e este tem sua contagem de palavras atualizada.

Figura 4 – Fluxograma de funcionamento da FGR



Fonte: Produção do próprio autor.

5.5.2 FGR Aplicado aos Áudios

Uma vez que o dicionário foi atualizado, é renovada também a listagem de arquivos transcritos. Feito isso, é oferecido ao revisor a opção de gerar uma versão do *dataset*. Dando início a esse processo, os arquivos de áudio, juntamente com seus rótulos, são copiados para um novo ficheiro reservado para o armazenamento dos *datasets* de teste. Um problema apresentado, entretanto, é o *sample encoding* do arquivo de áudio coletado. Como mencionado na Seção 3.2.2, o *codec* utilizado é o Sistema Global para Comunicação Móvel (GSM, do inglês *Global System for Mobile Communication*), cujo processamento não é suportado por grande parte das ferramentas de manipulação de áudio. Buscando solucionar essa incompatibilidade, foi utilizada a ferramenta SoX para realizar a descompressão do arquivo. Sua escolha é justificada

pela sua ampla utilização (BROWN; GARG; MONTGOMERY, 2018) e pela sua integração com o *prompt* de comando (terminal) de diversos sistemas operacionais. O Quadro 5 evidencia as alterações nas características dos áudios após o processamento pela FGR.

Quadro 5 – Áudios manipulados na FGR

Característica	Pré-processamento	Pós-processamento
<i>Bit rate</i>	13,0 kbps	128 kbps
Canais	1	1
<i>Sample rate</i>	8000 Hz	8000 Hz
<i>Sample encoding</i>	GSM	16-bit Signed Integer PCM

Fonte: Produção do próprio autor.

De acordo com Faria (2016), os *codecs* modulados por codificação de pulso (PCM, do inglês *pulse-code modulation*) foram criados em 1937 e são os *codec* que mais se aproximam do áudio analógico, uma vez que não há a introdução de nenhum tipo de compressão do sinal. Daí a sua classificação como *lossless* (sem perda). É válido observar que o *bit rate* é aproximadamente 10 vezes superior ao *bit rate* apresentado antes do processamento. Dessa forma, pode-se esperar um aumento do espaço ocupado em disco na mesma proporção. Apesar disso, o uso desse *codec* é vantajoso por ser compatível com qualquer aplicação de processamento de áudio, inclusive aquelas utilizadas no treinamento de redes neurais. Dessa forma, o *dataset* gerado pela FGR é compatível com as aplicações onde será utilizado.

5.6 Descarte de Áudios

Durante o processamento dos lotes de rótulos recebidos, pode-se perceber uma grande quantidade de arquivos sendo descartados. As razões para descarte são basicamente duas: ruído e má dicção. Analisando os arquivos, ficou evidente que a contribuição da primeira é mais volumosa. A Tabela 3 apresenta o panorama geral dos arquivos de áudio.

Alguns dos números são dignos de observação. De todo o conjunto de áudios aptos coletados, pouco mais de 4% foram considerados extremamente ruidosos e foram automaticamente eliminados. Cerca de 19% do conjunto apresentou alguma dificuldade para os transcritores e foram devidamente assinalados com essa condição. Dos 437 áudios marcados como duvidosos, com a ação minuciosa do revisor foi possível recuperar 309 faixas, o que corresponde a aproximadamente 70% do conjunto.

Tabela 3 – Relação de áudios duvidosos, ruidosos, descartados e recuperados

Canais	Aptos	Duvidosos	Ruidosos	Descartados	Recuperados*	Aceitos*
Posto 1 do CCO	202	28 (13,86%)	6 (2,97%)	24 (11,88%)	10 (35,71%)	6 (21,43%)
Posto 2 do CCO	394	57 (14,47%)	17 (4,31%)	50 (12,69%)	24 (42,11%)	13 (22,81%)
Posto 3 do CCO	365	53 (14,52%)	39 (10,68%)	91 (24,93%)	1 (1,89%)	10 (18,87%)
Posto 4 do CCO	244	40 (16,39%)	15 (6,15%)	16 (6,56%)	39 (97,5%)	11 (27,5%)
CCP de Laboreaux	39	7 (17,95%)	3 (7,69%)	3 (7,69%)	7 (100%)	7 (100%)
CCE	81	29 (35,8%)	3 (3,7%)	9 (11,11%)	23 (79,31%)	10 (34,48%)
CCM	86	12 (13,95%)	1 (1,16%)	1 (1,16%)	12 (100%)	4 (33,33%)
Oficina de vagões	35	7 (20%)	0 (0%)	1 (2,86%)	6 (85,71%)	2 (28,57%)
Torre A	290	57 (19,66%)	3 (1,03%)	14 (4,83%)	46 (80,7%)	14 (24,56%)
Torre B	128	27 (21,09%)	0 (0%)	5 (3,91%)	22 (81,48%)	7 (25,93%)
Torre C	63	20 (31,75%)	1 (1,59%)	2 (3,17%)	19 (95%)	5 (25%)
Torre D	146	41 (28,08%)	10 (6,85%)	10 (6,85%)	41 (100%)	11 (26,83%)
Torre E	138	41 (29,71%)	1 (0,72%)	5 (3,62%)	37 (90,24%)	13 (31,71%)
Torre L	115	24 (20,87%)	0 (0%)	2 (1,74%)	22 (91,67%)	8 (33,33%)
Total	2326	443 (19,05%)	99 (4,26%)	233 (10,02%)	309 (69,75%)	121 (27,31%)

Fonte: Produção do próprio autor.

Nota: * Percentual relativo aos duvidosos.

Dentre os áudios recuperados existem duas categorias distintas: áudios resolvidos e áudios aceitos. Um áudio é considerado resolvido quando as dúvidas elencadas pelos transcritores são plenamente sanadas. Os áudios categorizados como aceitos são áudios que apresentam um total de incertezas inferior a 5% do total de palavras contidas no rótulo. O total de áudios aceitos corresponde a aproximadamente 40% dos áudios recuperados e, por consequência, 5% do total de arquivos aptos. Do montante total de 2326 arquivos de áudio, ao todo 233 (10%) foram descartados devido a ruído e/ou incompreensibilidade do conteúdo.

É importante ressaltar que uma parcela considerável dos dados assinalados como duvidosos foram descartados devido ao alto nível de ruído. O número de áudios ruidosos levantado na Tabela 3 se refere às faixas que foram descartadas pelos transcritores pelo ruído, sem que houvesse sequer associação de rótulo.

6 FERRAMENTA DE SUPRESSÃO DE RUÍDO

Observando que ao menos 10% do volume de dados elegíveis à transcrição tiveram de ser descartados principalmente pelo ruído, buscou-se desenvolver uma ferramenta de supressão de ruído com objetivo de melhorar não apenas a transcrição dos áudios, como também montar o *dataset* com faixas de áudio mais limpas. Tal recurso tem por finalidade melhorar a compreensão de dados tanto para os ouvidos humanos como para os sistemas de processamento em uso.

6.1 Método de Supressão de Ruído

Como explicitado na Seção 5.5.2, a FGR possui uma rotina de descompactação de arquivos baseada em SoX. Como descrito na Seção 2.5, esta aplicação foi uma das responsáveis pela popularização da técnica de supressão de ruído de subtração espectral. Partindo dessa análise, decidiu-se desenvolver um supressor de ruído usando a ferramenta SoX e, por conseguinte, fazendo uso do método de supressão de ruído SS.

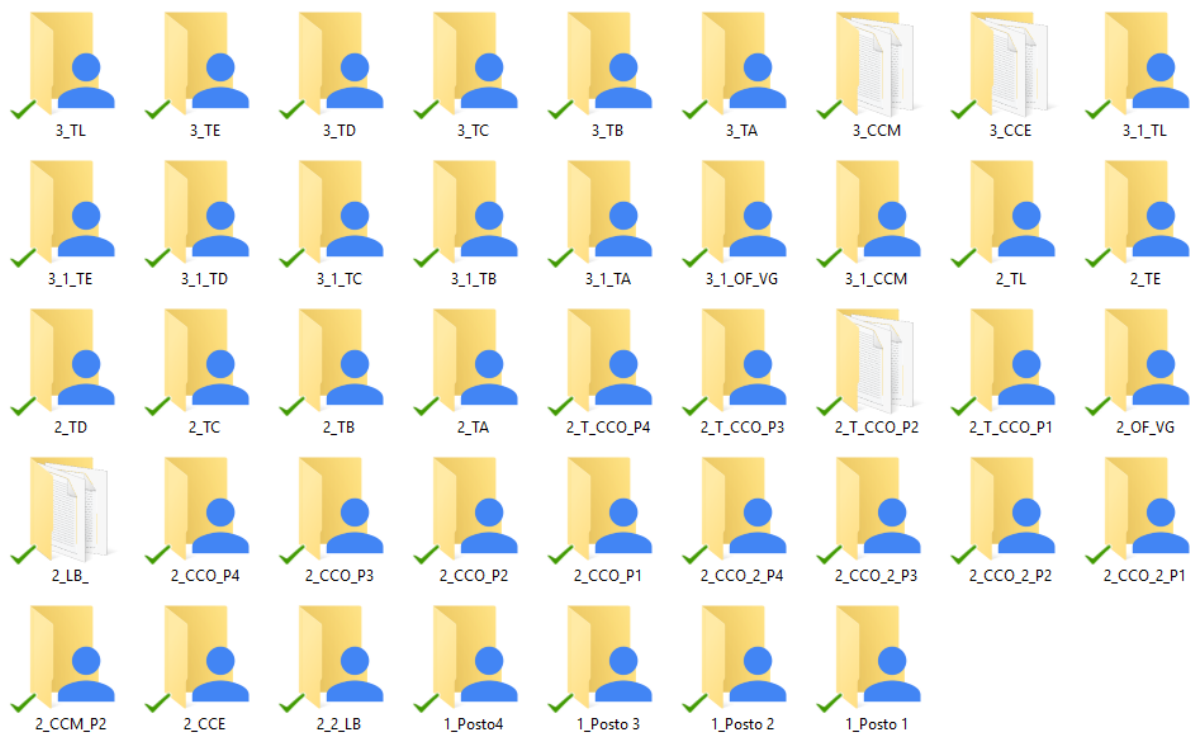
Para uso desse método, vale salientar que é necessária a criação de um perfil de ruído. Um novo problema que surge, portanto, é obter um arquivo de áudio que contenha exclusivamente ruído para criar tal perfil. É característico da comunicação via rádio um intervalo entre as instruções quando da troca de interlocutores (geralmente existe uma pausa de quase um segundo após o câmbio, até que o destinatário do comando se pronuncie). Nesse intervalo, o sinal é composto unicamente de ruído, sendo ideal para geração do PR.

Tomando por base essas informações, foi desenvolvida uma aplicação que avalia a amplitude do sinal. Quando este é inferior ao limiar da fala (valor obtido através de extensa observação), garantindo que se mantenha por tempo superior a um segundo, este trecho é tomado como PR a ser utilizado na supressão de ruído deste arquivo de áudio. Fazendo isso, é garantido que o ruído da faixa é completamente removido. Em arquivos com duração superior a 20 segundos é comum haver mais de um trecho elegível a PR. Nestas situações são utilizados aqueles de maior duração.

6.2 Algoritmo de Supressão de Ruído Aplicado a Ficheiros

Para compreensão do funcionamento do supressor de ruído aplicado aos ficheiros do depósito de rótulos e áudios, se faz necessário compreender a estrutura sob a qual estes estão organizados. A Figura 5 mostra a separação dos dados e é imprescindível para compreensão dos processos realizados em lotes.

Figura 5 – Distribuição dos rótulos e arquivos de áudio



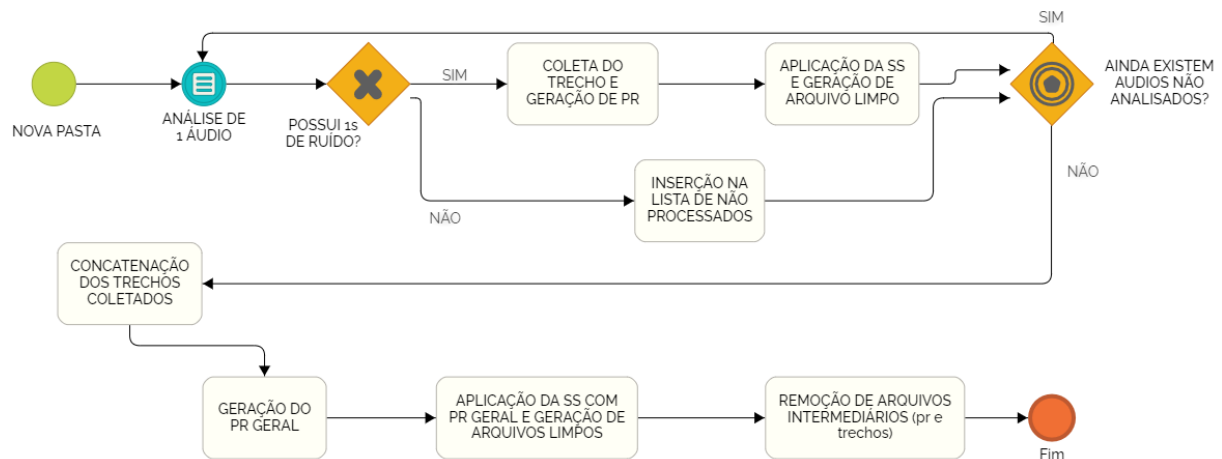
Fonte: Produção do próprio autor.

Os arquivos foram classificados através de dois critérios primordiais: data de gravação do áudio e canal. Dessa forma, foi possível agrupar os áudios que possuem basicamente as mesmas características de ruído próprio do canal, partindo do pressuposto que as condições do ambiente e do canal não se alterem no período no qual os dados foram coletados. Tal pensamento, entretanto, não pode ser propagado para faixas de mesmo canal oriundas de diferentes dias.

Como pode ser observado na Figura 6, o algoritmo de supressão coleta um trecho sem fala de cada arquivo de áudio, o utiliza na criação de um PR e por fim, elimina o ruído da faixa original salvando a faixa limpa no diretório de acomodação de *datasets* de teste. É recorrente, entretanto, que algumas faixas não possuam o trecho sem fala de duração superior a um segundo. Nesses

casos, estes arquivos não possuem um PR extraído de si próprio que possa servir de base para o processo de SS. Utilizando a estrutura das pastas listadas na Figura 5, optou-se por manter salvo o arquivo da amostra de ruído de cada faixa até o fim do processamento da pasta. Uma vez que todos os áudios da pasta foram analisados, cria-se uma lista daqueles que não foram tratados. As amostras de ruído são concatenadas objetivando criar uma amostra de ruído geral, que, embora não apresente as características de um único arquivo de áudio especificamente, contempla o padrão de ruído presente naquele canal durante o tempo de gravação dos dados coletados. Passa-se então a tratar as faixas restantes com o PR generalista obtido. Dessa forma, garante-se que as características específicas daquele canal (durante o tempo de gravação dos arquivos) sejam observadas, trazendo maior assertividade na remoção de ruído.

Figura 6 – Fluxograma do algoritmo de supressão de ruído



Fonte: Produção do próprio autor.

6.3 Análise Qualidade de Remoção de Ruído

Buscando avaliar a eficácia da ferramenta em atender os objetivos propostos, os arquivos de áudio processados foram avaliados em quatro frentes principais: amplitude do sinal, Transformada Rápida de Fourier (FFT, do inglês *Fast Fourier Transform*) e Coeficientes Cepstrais de Frequência Mel (MFCC, do inglês *Mel-Frequency Cepstral Coefficients*), além de uma audição junto à equipe do projeto e representantes da Vale S.A. Para apresentação neste documento foram selecionados dois arquivos de áudio para serem analisados sob a ótica das quatro seções seguintes. O Quadro 6 apresenta as informações pertinentes.

Quadro 6 – Arquivos de áudio usados na análise de ruído

Característica	Áudio 1	Áudio 2
Canal	21 – Posto 1 CCO	21 – Posto 1 CCO
Data	20 jan. 2020	20 jan. 2020
Hora inicial	07:02:13	07:22:28
Duração	00:00:28	00:00:13
Interlocutores	CCO e Campo	Apenas CCO
Rótulo	atenção a três cinco um na quatro um câmbio três cinco um na quatro um atende cco câmbio o senhor gasta quanto tempo na manobra aí câmbio a manobra dá pra fazer com quinze minutos câmbio tá entendido hein cargueiro livrando aí o senhor me chama por favor câmbio cargueiro acabou de livrar aqui câmbio	mas ele vai livre livrando a quatro um ele vai falando pra gente depois ele trabalha na três oito linha um tá bom três oito linha um circulação até gv nada

Fonte: Produção do próprio autor.

Foram escolhidas estas duas faixas de áudio para elucidar algumas peculiaridades encontradas. Quando se tem apenas um interlocutor no registro, não existe a presença do trecho de ruído que é utilizado na supressão de ruído. Por outro lado, havendo mais de um interlocutor, em geral, após o uso da palavra câmbio, existe o trecho de ruído necessário para uso do algoritmo de supressão de ruído.

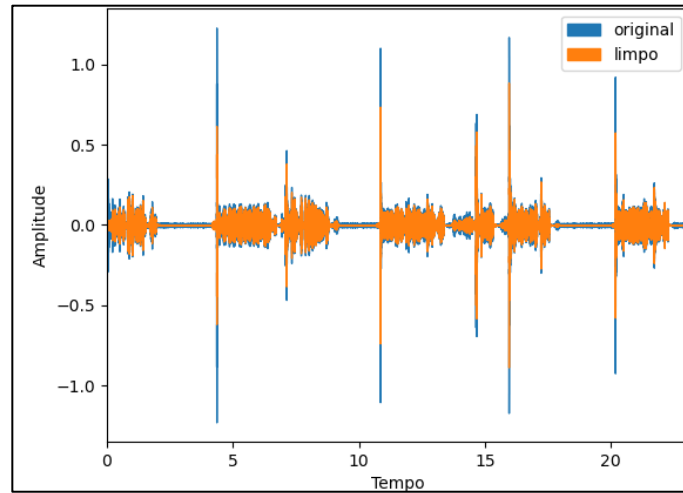
Para as análises subsequentes foram desenvolvidas rotinas em *Python* 3.8 com o intuito de gerar gráficos que permitissem a análise das características que não necessariamente são perceptíveis ao ouvido humano. Devido à sua grande versatilidade, foi utilizada a biblioteca *Librosa* para manipulação dos sinais de áudio e geração dos gráficos (MCFEE et al., 2015).

6.3.1 Amplitude

O primeiro aspecto a ser observado foi o gráfico de amplitude dos arquivos estudados. Através desse gráfico, pode-se perceber o sinal variando em amplitude ao longo do tempo. Aspecto interessante que pode exibir mais claramente a influência do ruído no sinal. Para uma melhor

visualização das informações contidas nos gráficos, optou-se pela plotagem do áudio limpo sobreposto ao áudio original.

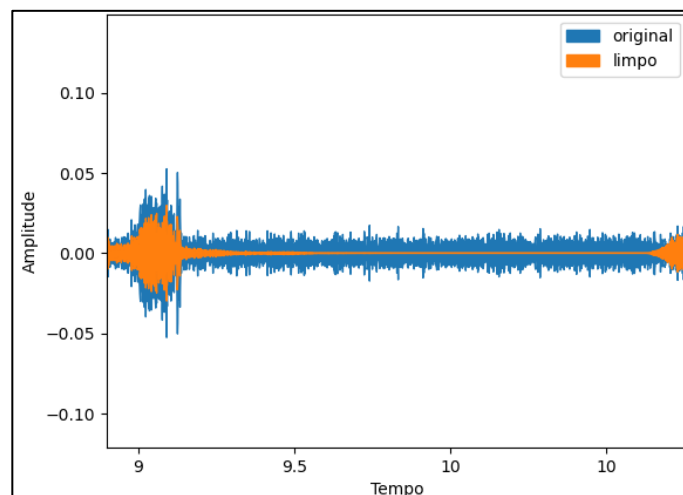
Gráfico 2 – Gráfico de amplitude do áudio 1



Fonte: Produção do próprio autor.

Percebe-se que, no Gráfico 2, o sinal limpo se encontra completamente envolvido pelo sinal original. Isso significa que todos os pontos de amostragem tiveram sua amplitude reduzida. O Gráfico 3 foca no trecho compreendido entre 9 e 11 segundos, aproximadamente. Pode-se afirmar que o trecho de menor amplitude é composto unicamente de ruído, até pela sua evidente distinção do restante do gráfico. Pelo Gráfico 2, sabe-se que o trecho evidenciado, devido à sua duração, não gera o PR. Apesar disso, nota-se uma alta redução do ruído no trecho destacado.

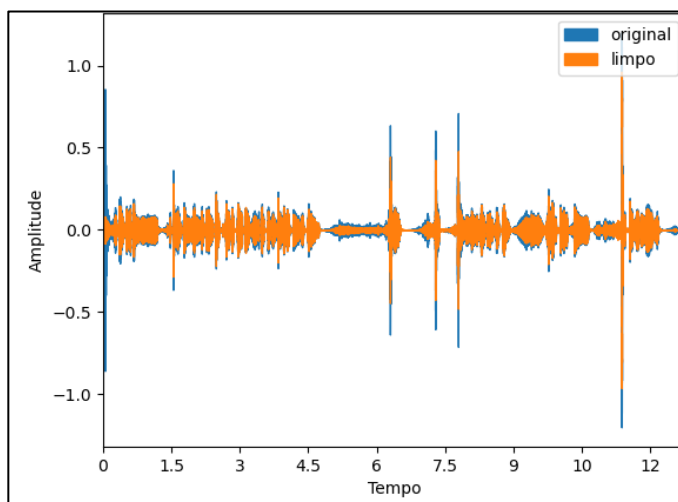
Gráfico 3 – Foco do gráfico de amplitude do áudio 1



Fonte: Produção do próprio autor.

De acordo com os requisitos para criação do PR (vide Seção 6.1), o Gráfico 4 mostra que o áudio 2 foi limpo com uso do PR geral, uma vez que os trechos de baixa amplitude não possuem o mínimo de um segundo ininterrupto. De acordo com o Quadro 6, é sabido que, por ser pertencente ao mesmo canal e oriundo do mesmo dia, o áudio 1 foi indiretamente utilizado para remover o ruído do áudio 2.

Gráfico 4 – Gráfico de amplitude do áudio 2



Fonte: Produção do próprio autor.

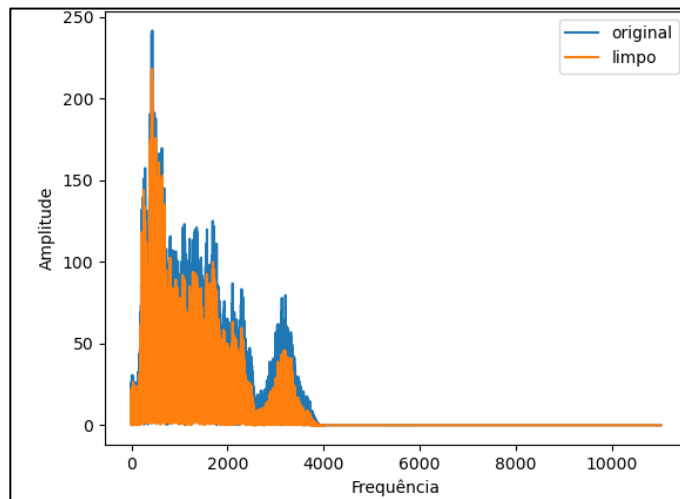
Uma particularidade do áudio 2 é a presença de apenas um interlocutor. Isso justifica a ausência de um trecho sem fala, uma vez que não há espera para efetivar a comunicação. Percebe-se ainda que o pico inicial (ocasionado por impactos mecânicos ao se pressionar o botão para dar início à comunicação, por exemplo) não existe no áudio limpo. A diminuição dos picos de amplitude (observada em ambos os arquivos) auxilia na diminuição de *outliers*. Dessa forma, o volume de dados fica menos disperso, auxiliando, em tese, a assimilação dos dados pelas redes neurais.

6.3.2 FFT

“A Transformada de Fourier é, em essência, uma ferramenta matemática que realiza a transição entre as variáveis tempo e frequência de sinais” (GONÇALVES, 2004, p. 3). Enquanto na plotagem da amplitude observa-se o comportamento do sinal no tempo, através da plotagem da FFT avalia-se o sinal no domínio da frequência. A mesma estratégia na plotagem dos sinais foi adotada para esta análise: sobreposição de gráficos.

Uma vez que a voz humana possui energia significativa para composição do sinal de voz apenas até o limiar de 4 kHz (SOTERO FILHO, 2017) e sabendo que o *codec* utilizado foi desenvolvido para sistemas de comunicação de voz (HUERTA; STERN, 1997), os arquivos coletados têm sua frequência limitada a este valor. Tanto o Gráfico 5 como o Gráfico 6 mostram claramente esse efeito. Uma das implicações esperadas com a remoção de ruído é a diminuição da energia de frequências onde há ruído periódico. Como os dados coletados são limitados a frequências de até 4 kHz, os ruídos periódicos presentes são reduzidos, mas não se observa graficamente de forma clara, uma vez que estas frequências não são zeradas por portarem informações oriundas da fala dos interlocutores.

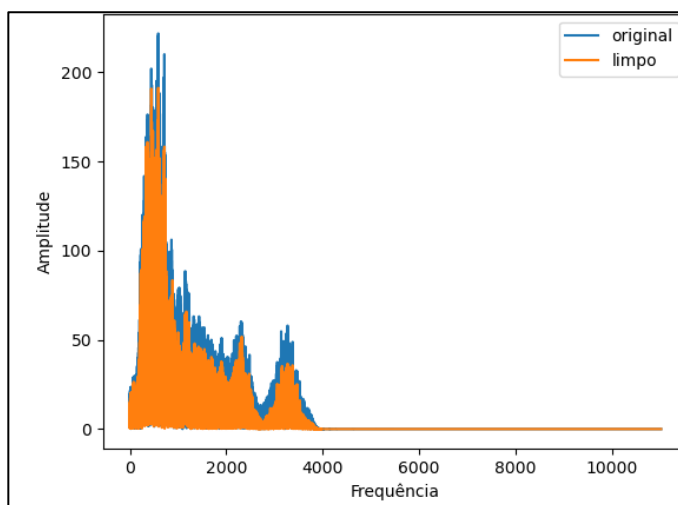
Gráfico 5 – FFT do áudio 1



Fonte: Produção do próprio autor.

Analisando ambos os gráficos, pode-se perceber uma maior concentração da energia do sinal compreendida entre 0 e 2,75 kHz, seguida de uma concentração de menor intensidade em torno de 3,3 kHz. Percebe-se que a energia do sinal como um todo sofreu uma redução após o procedimento de SS, de forma que, assim como observado no domínio do tempo, o sinal limpo está envolvido pelo sinal original. É possível, portanto, afirmar que há uma melhora obtida pela ferramenta de supressão de ruído.

Gráfico 6 – FFT do áudio 2



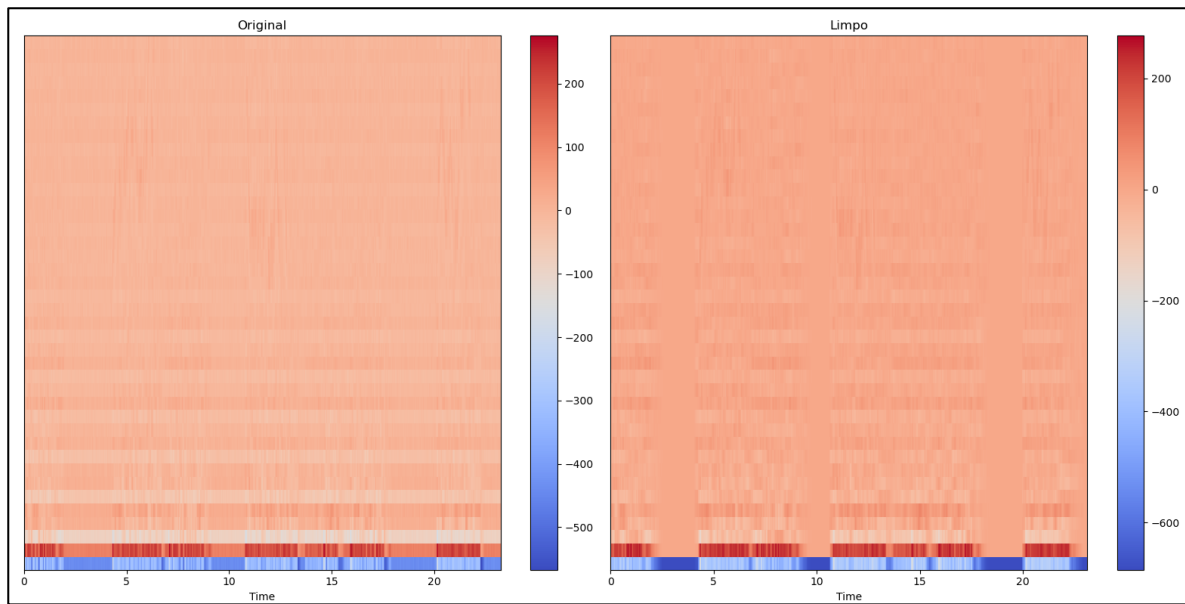
Fonte: Produção do próprio autor.

6.3.3 MFCC

Segundo Nassif e outros (2019), os MFCC são os componentes principais de sistemas modernos voltados para reconhecimento de fala. “O objetivo dos coeficientes cepstrais é extrair características a respeito do trato vocal da fonte de excitação, pois as características do trato vocal contêm informações sobre a dicção dos fonemas” (SCART, 2019, p. 2019). Fazendo uso de gráficos que mostram os MFCC, busca-se avaliar se existe alguma mudança notável trazida pelo processo de supressão de ruído. O modelo de exibição desse tipo de gráfico impede a sobreposição, sendo necessário recorrer à justaposição dos estados pré e pós-processamento.

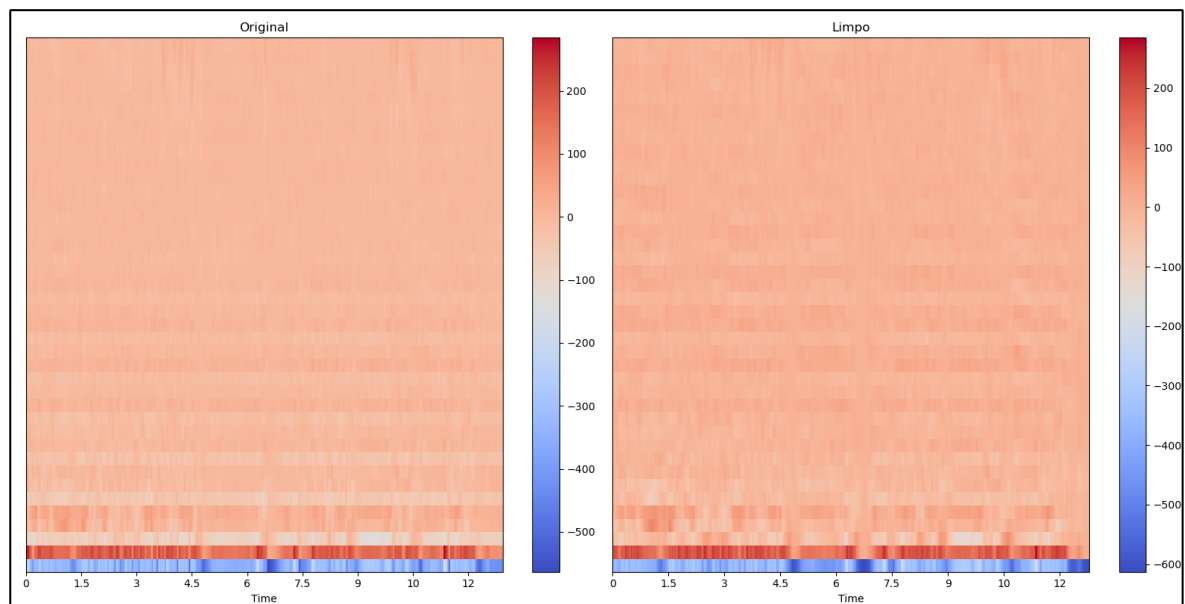
O gráfico gerado a partir da obtenção dos MFCC tem sua leitura visual dificultada pela alta densidade de informação que este agrupa. Em uma análise simples direta, é perceptível um aumento do contraste nas representações do sinal limpo: os trechos em azul se tornam mais vívidos, bem como os trechos em vermelho na linha imediatamente superior. Além disso, pode-se notar o surgimento de colunas onde a intensidade dos coeficientes sofre menos alteração. No sinal limpo do Gráfico 7 isso é mais evidente. Acima dos espaços cuja primeira linha possui intensidade da ordem de -600 (trechos em azul mais escuro), praticamente não há variação na intensidade dos coeficientes. Tal efeito pode ser observado no Gráfico 8 com menos recorrência, porém ainda está presente.

Gráfico 7 – MFCC do áudio 1



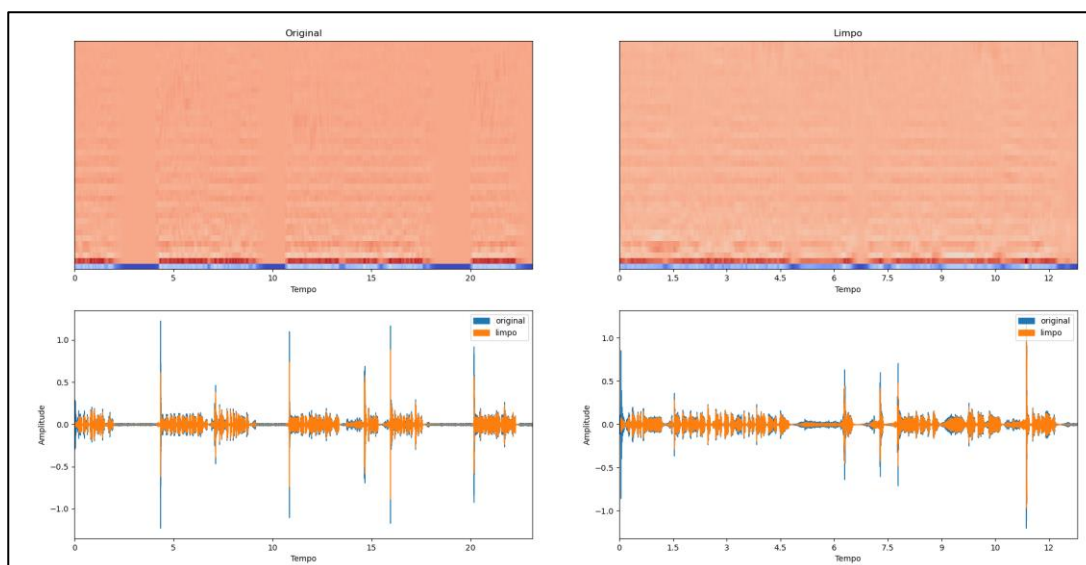
Fonte: Produção do próprio autor.

Gráfico 8 – MFCC do áudio 2



Fonte: Produção do próprio autor.

Gráfico 9 – MFCC limpos e gráficos e amplitude dos sinais



Fonte: Produção do próprio autor.

Em ambos os gráficos nota-se que o sinal limpo excursa o valor dos coeficientes até a marca de -600, enquanto as faixas originais não ultrapassam o limiar de -500. Percebe-se ainda que existem alterações que são feitas até mesmo nos coeficientes de mais alta ordem, que se posicionam no topo do gráfico.

O Gráfico 9 foi gerado apenas com a intenção de mostrar simultaneamente a amplitude do sinal e os MFCC, pelo fato de, como ambos são apresentados no domínio do tempo, ser possível observar a distribuição dos coeficientes e relacionar características de amplitude a estes. Uma das observações feitas, por exemplo, é de que quando existe silêncio no sinal, os coeficientes inferiores atingem amplitudes ainda menores (da ordem de -500).

6.3.4 Audição

Uma outra análise realizada, essa em conjunto com a equipe do projeto ao qual este trabalho reporta conhecimento, foi a reprodução e avaliação qualitativa de alguns arquivos de áudio. Percebeu-se que o ruído característico do canal foi removido de forma satisfatória. Ou seja, a captação do áudio apresenta ruídos que estão sendo filtrados de forma a atender as expectativas. Quando se avalia, no entanto, áudios com mais de um interlocutor nota-se que estes apresentam qualidade variável de acordo com o falante.

Quando da apresentação dos resultados da supressão de ruído à equipe da Engenharia de Operação da Vale S.A. houve uma aceitação considerável da qualidade alcançada.

6.4 Aplicabilidade

Na Seção anterior foram apresentados dois arquivos de áudio com uma diferença crítica: a quantidade de interlocutores presentes na faixa de áudio. Havendo mais de uma voz na gravação, a supressão de ruído tende a ser mais eficaz no locutor que está em ambiente administrativo, devido a menor quantidade de ruído externo. Havendo apenas uma voz, existe uma leve queda na qualidade da supressão de ruído por fazer uso do PR geral.

Uma das conclusões que se pode chegar analisando as FFT obtidas é que, pelo fato de o sinal excursar frequências inferiores ao limiar da voz humana, métodos de supressão de ruído baseados em restrição de banda não são aplicáveis. Dessa forma, apenas métodos mais robustos serão capazes de obter resultados superiores aos aqui demonstrados. Uma outra dificuldade encontrada é a obtenção do perfil de ruído. Como tanto a SS quanto a NG se baseiam na existência desse perfil, é possível que resultados usando técnicas de *Noise Gating* sejam semelhantes.

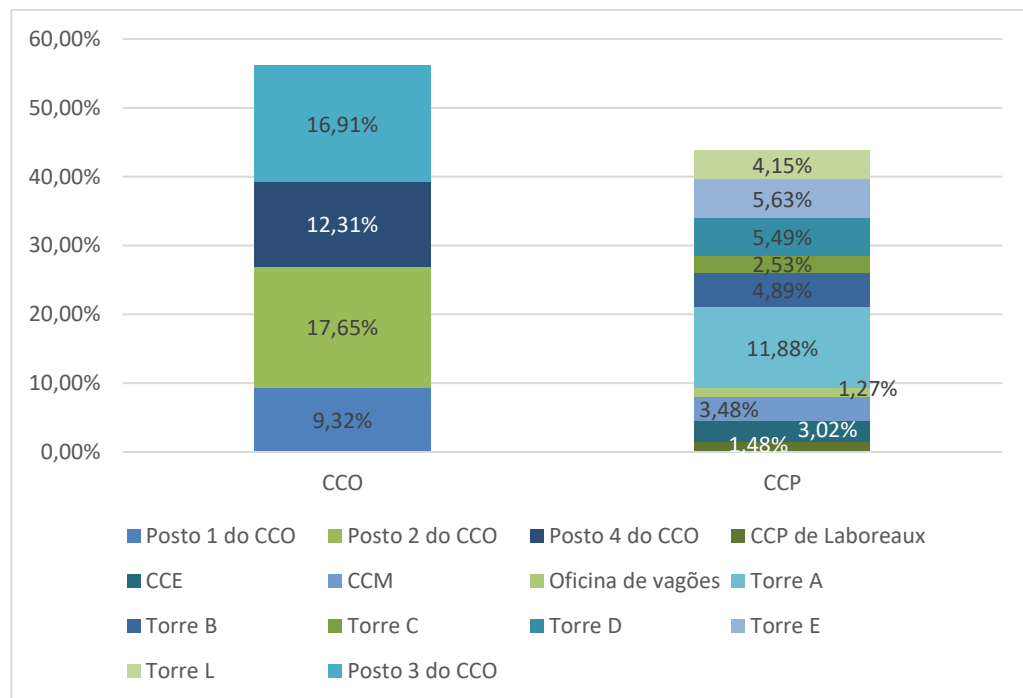
Para compreensão humana, portanto, pode-se considerar a remoção de ruído aceitável. No entanto, é clara a diferença de ruído dos ambientes de trabalho dos controladores e maquinistas. A locomotiva insere um nível considerável de ruído nas comunicações pelo fato de trabalhar com motores à combustão interna com potência da ordem de 4000 hp. Tais ruídos são periódicos, uma vez que o motor diesel tem suas rotações bem definidas.

Uma possibilidade de solução do ruído captado dentro da locomotiva é utilizar os equipamentos de comunicação da ferrovia para obter um áudio contendo apenas o barulho dos motores diesel para cada velocidade de rotação predefinida. Utilizando esses áudios na geração de PRs, é provável obter melhores resultados. Apesar disso, ainda seria necessário realizar a segmentação dos sinais do *dataset* através de alguma técnica de ASD para então realizar a remoção de ruído.

7 ANÁLISE DOS DADOS

Durante o período de aquisição de dados, foram coletados o total de 2844 arquivos de áudio registrados pelo console de gravação de rádio da ferrovia. Aproximadamente 56% dos áudios são do CCO e 44% são originários dos canais que compõem o grupo CCP. A distribuição de cada uma dessas categorias é evidenciada no Gráfico 10.

Gráfico 10 – Distribuição dos áudios coletados



Fonte: Produção do próprio autor.

Após a coleta dos arquivos e o devido tratamento dos dados (remoção dos arquivos longos e ruidosos, análise e solução dos áudios duvidosos e descarte dos não compreensíveis), tem-se uma distribuição ligeiramente mais homogênea na relação entre CCO e CCP (51% e 49%, respectivamente). Tal equilíbrio é importante pelo caráter duplo das comunicações da ferrovia: linha tronco e pátios.

Os áudios recebidos foram trabalhados observando-se vários critérios ao longo do processo de análise. A priori, foram descartados os áudios com duração superior a trinta segundos (Seção 5.1). Por fim, foram removidas as amostras que não puderam ser compreendidas e as amostras que apresentavam pequenas incertezas foram sinalizadas (Seção 5.6). Dessa forma, o grupo de áudios transcritos é uma categoria que representa o *dataset* de interesse e é composto de 2093

sinais de áudio e texto divididos em áudios precisos e aceitos (incerteza inferior ou igual a 5%, vide Seção 5.6). A Tabela 4 apresenta detalhadamente a composição do conjunto de dados.

Tabela 4 – Classificação dos áudios coletados.

Canais	Col.	Longos	Descartados	Precisos	Aceitos	Transcritos
Posto 1 do CCO	265	63 (23,77%)	24 (9,06%)	172 (64,91%)	6 (2,26%)	178 (67,17%)
Posto 2 do CCO	502	108 (21,51%)	50 (9,96%)	331 (65,94%)	13 (2,59%)	344 (68,53%)
Posto 3 do CCO	481	116 (24,12%)	91 (18,92%)	264 (54,89%)	10 (2,08%)	274 (56,96%)
Posto 4 do CCO	350	106 (30,29%)	16 (4,57%)	217 (62%)	11 (3,14%)	228 (65,14%)
CCP de Laboreaux	42	3 (7,14%)	3 (7,14%)	29 (69,05%)	7 (16,67%)	36 (85,71%)
CCE	86	5 (5,81%)	9 (10,47%)	62 (72,09%)	10 (11,63%)	72 (83,72%)
CCM	99	13 (13,13%)	1 (1,01%)	81 (81,82%)	4 (4,04%)	85 (85,86%)
Oficina de vagões	36	1 (2,78%)	1 (2,78%)	32 (88,89%)	2 (5,56%)	34 (94,44%)
Torre A	338	48 (14,2%)	14 (4,14%)	262 (77,51%)	14 (4,14%)	276 (81,66%)
Torre B	139	11 (7,91%)	5 (3,6%)	116 (83,45%)	7 (5,04%)	123 (88,49%)
Torre C	72	9 (12,5%)	2 (2,78%)	56 (77,78%)	5 (6,94%)	61 (84,72%)
Torre D	156	10 (6,41%)	10 (6,41%)	125 (80,13%)	11 (7,05%)	136 (87,18%)
Torre E	160	22 (13,75%)	5 (3,13%)	120 (75%)	13 (8,13%)	133 (83,13%)
Torre L	118	3 (2,54%)	2 (1,69%)	105 (88,98%)	8 (6,78%)	113 (95,76%)
Total	2844	518 (18,21%)	233 (8,19%)	1972 (69,34%)	121 (4,25%)	2093 (73,59%)

Fonte: Produção do próprio autor.

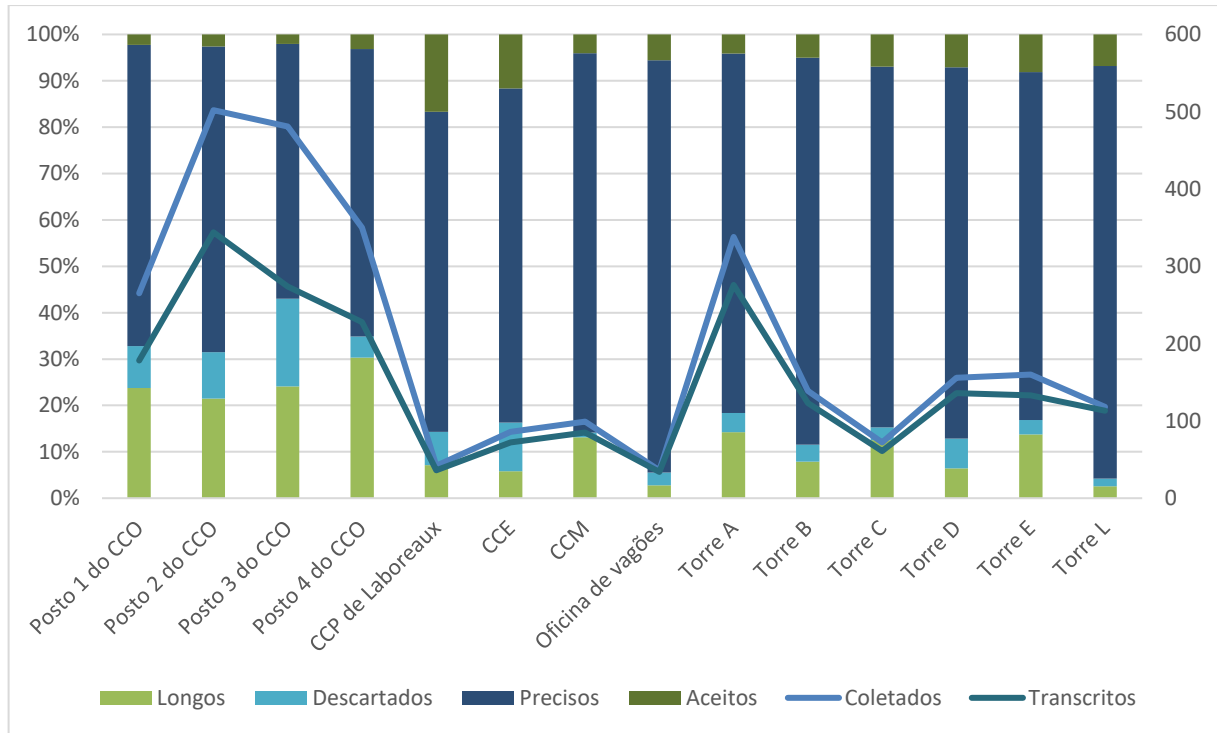
Nota: Os valores percentuais são em relação à coluna de áudios coletados.

Analisando a Tabela 4 é possível chegar a algumas conclusões interessantes: (i) o CCO possui uma dinâmica de comunicação que leva mais tempo para ser concluída, uma vez que mais de 75% dos áudios longos são desta categoria; (ii) alguma falha ou condição estranha fez com que o posto 3 tivesse quase 19% dos dados descartados; (iii) a Torre L e a Oficina de Vagões se destacam pelo alto índice de áudios precisos: 89%, o que é um indicador preliminar de melhores condições do canal ou melhor dicção dos interlocutores; (iv) os áudios aceitos são majoritariamente do CCP (67%), indicando que, embora as comunicações sejam mais curtas, são mais difíceis de alcançar a compreensão plena; (v) a quantidade de arquivos transcritos tem um comportamento similar em ambas categorias, havendo uma variação no percentual de transcritos em no máximo 12% dentro de cada categoria (56% a 68% para CCO e 83% a 95% para o CCP).

O Gráfico 11 tem por finalidade fazer uma demonstração visual dos dados reunidos na Tabela 4. Uma qualidade da utilização desse gráfico é a percepção de como cada canal de áudio é composto. Totalizando os áudios coletados, as colunas são divididas em longos, descartados, precisos e aceitos, dando uma ideia mais clara do percentual de cada categoria para a composição de cada canal. As linhas, por sua vez, trazem uma noção da magnitude de cada uma

das colunas. Pelo fato de as colunas serem expressas em valores percentuais, o gráfico de linhas (lido no eixo a direita) nos permite compreender melhor a participação de cada canal no total obtido.

Gráfico 11 – Distribuição dos áudios coletados por canal



Fonte: Produção do próprio autor.

Nota: O eixo à esquerda corresponde às barras, enquanto o da direita, às linhas.

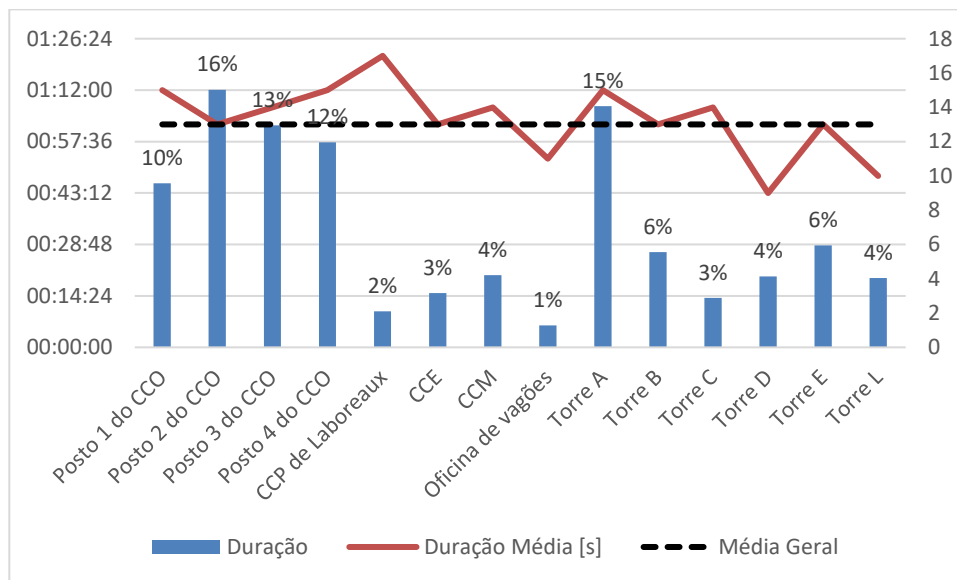
7.1 Métricas

Para uma melhor compreensão dos dados obtidos, é importante selecionar as características que serão avaliadas de forma quantitativa e qualitativa. A contagem de dados pertencentes a cada canal pode fornecer uma ideia não muito clara da sua real contribuição na composição do *dataset*, considerando que cada áudio contribui com diferentes quantidades de informações. Partindo desse pressuposto, foram realizadas duas análises de cunho estatístico: duração e quantidade de palavras em cada item do *dataset* (áudio e rótulo).

7.1.1 Duração

É interessante avaliar a duração total que cada canal contribui pois, através disso, pode-se notar quanto tempo é necessário para se efetivar a comunicação característica de um determinado posto de trabalho. Embora exista a restrição de trinta segundos, tal apreciação permanece pertinente para construção de um conjunto de dados que consiga abranger as particularidades do problema que busca solucionar. Para a melhor compreensão dessas informações, foi elaborado o Gráfico 12. Neste, as linhas correspondem ao eixo a direita, enquanto as barras podem ser lidas no eixo a esquerda. Os valores percentuais representam a contribuição do canal para o total do dataset.

Gráfico 12 – Duração total e média por canal



Fonte: Produção do próprio autor.

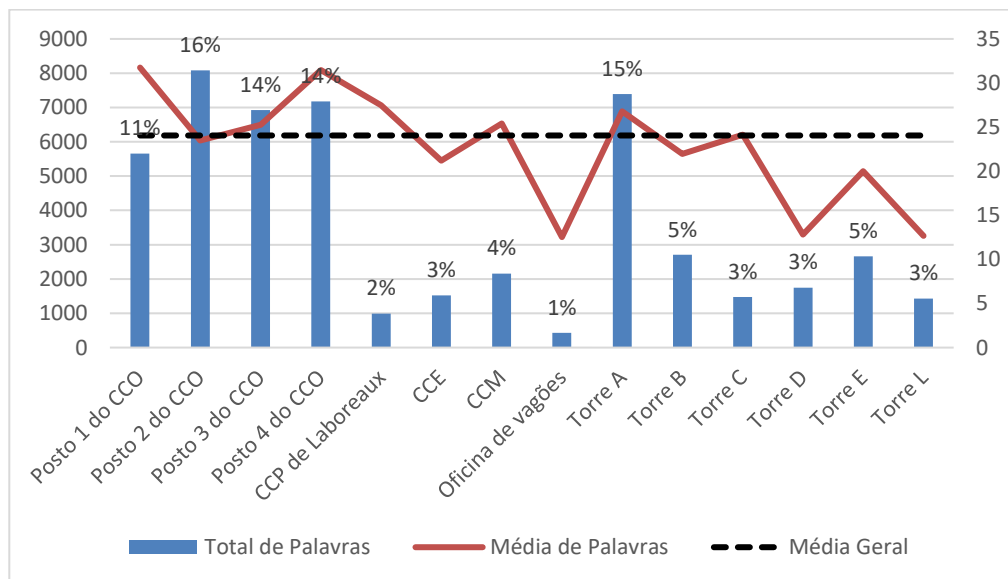
Observando o Gráfico 12, pode-se perceber que a variabilidade da média de duração fica limitada entre 9 e 17 s, sendo que as médias mais altas são oriundas de canais com pequena colaboração. Os canais que possuem mais expressividade (valores próximos a 1 hora) têm seus valores entre 13 e 15 s, trazendo a ideia de que quanto maior a amostragem de dados, mais converge a média para o valor de 13 s. Tal suposição é confirmada pela média geral, que é aproximadamente 13 s. Uma observação que pode ser feita é que a duração do canal CCM é superior à Torre D e à Torre L, embora a quantidade de áudios transcritos destes seja cerca de 40% superior. A diferença de duração total entre canais CCO e CCP gira em torno de 10

minutos (2%). Num universo de 7 horas e 45 minutos, é uma diferença irrisória, podendo-se afirmar que o *dataset* é equilibrado no tocante à duração de gravações entre CCO e CCP.

7.1.2 Contagem de palavras

Outra métrica interessante que permite checar o equilíbrio de contribuição entre as categorias é a contabilização de palavras contidas em rótulos do canal. Posto que o *dataset* é voltado para o reconhecimento de fala, essa seria a maneira mais precisa de dimensionar a influência de cada canal para a formação do *dataset*. Mesmo sendo uma distribuição homogênea no que se refere à duração, a quantidade de palavras pode variar de acordo com algumas características específicas da comunicação de dado canal.

Gráfico 13 – Contagem de palavras e média por canal



Fonte: Produção do próprio autor.

Ao confrontar o Gráfico 12 e o Gráfico 13 é possível notar um comportamento semelhante dos canais, levando à conclusão de que duração e quantidade de palavras são diretamente relacionadas. Algumas diferenças, entretanto, são interessantes: (i) a exceção do canal Posto 2 do CCO, a categoria como um todo apresentou um aumento nos seus índices; (ii) além da Oficina de Vagões e das Torres D e L, que já figuram abaixo da média de duração geral, para contagem de palavras, o CCE e as Torres B e E são adicionadas a esse grupo; (iii) as torres que estão na média ou acima (A e C) são opostos no tocante à quantidade de dados disponíveis, não

sendo possível chegar a conclusões de convergência no contexto de TC; (iv) no panorama geral, estão acima da média canais com mais de 170 arquivos transcritos e também canais com 36 a 85 arquivos transcritos, não havendo correlação entre quantidade de palavras e quantidade de rótulos gerados para o canal.

Avaliando o total das categorias CCO (55%) e CCP (45%), percebe-se que, em um panorama homogêneo, o CCO tende a incorrer numa comunicação mais veloz. Tal conclusão se apresenta como um incentivo à avaliação mais profunda da velocidade de fala empregada em cada canal, o que leva à seguinte Seção.

7.1.3 *Speech Pace*

Pace nada mais é do que o compasso, ou velocidade com que se realiza algo. No contexto aqui apresentado, *speech pace* ou simplesmente *pace* se refere à velocidade de fala de um interlocutor, ou à velocidade de fala medida em um arquivo de áudio. Tal informação é obtida através da combinação das duas análises anteriores e pode oferecer informações importantes sobre a qualidade da comunicação efetivada em cada canal.

De acordo com *SpeakerHub* (2017), um *pace* mais rápido traz a ideia de urgência, emoção, paixão e entusiasmo, enquanto um *pace* mais devagar é mais ligado a importância, tristeza, confusão e introdução a novas ideias. De certa forma, o ideal é que haja uma variação do *pace* para que a comunicação seja efetivada com maior significado quando aplicada a discursos ou situações correlatas. No contexto de troca de informações via rádio, onde geralmente há uma necessidade de clareza e compreensibilidade, não é interessante a recorrência de um *pace* rápido, chegando a ser até um dos critérios da avaliação realizada pelos inspetores dos centros de controle.

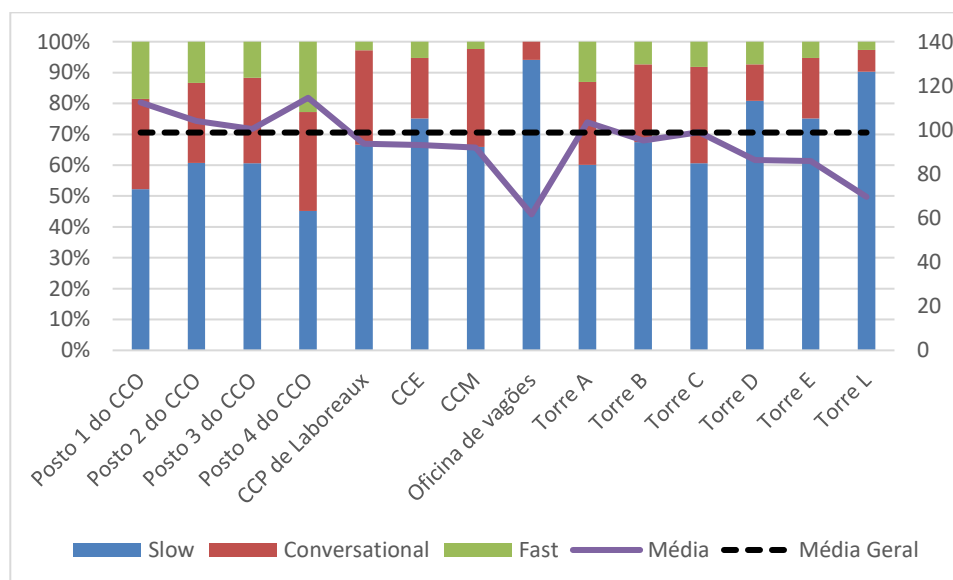
SpeakerHub (2017) introduz a unidade palavras por minuto (WPM, do inglês *words per minute*) para avaliar a velocidade de fala de um locutor qualquer e um conjunto de informações de referência para categorização do discurso. Tais informações são condensadas no Quadro 7.

Quadro 7 – Limites das categorias de *pace*

Limite inferior	Limite superior	Classificação
0	110	<i>Slow</i>
110	150	<i>Conversational</i>
150	-	<i>Fast</i>

Fonte: *SpeakerHub* (2017).

Em primeira instância obteve-se o *pace* de cada arquivo de áudio cruzando informações de duração do áudio e quantidade de palavras no rótulo. A partir daí foram feitas duas análises: (i) cálculo do *pace* médio do canal e (ii) contagem de arquivos em cada categoria de *pace*. Tais análises são apresentadas no Gráfico 14.

Gráfico 14 – *Pace* médio e distribuição dos canais nas categorias de *pace*

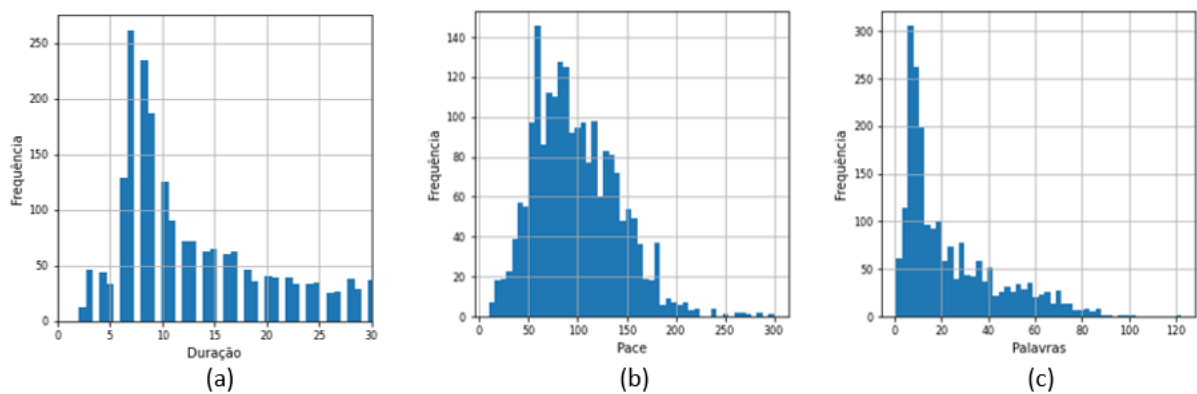
Fonte: Produção do próprio autor.

Naturalmente, o comportamento da linha de média por canal é diretamente influenciado pela distribuição das porções no gráfico de barras. Algumas informações importantes são extraídas dessa análise: (i) *slow* é a categoria predominante, representando no mínimo, 45% de cada canal; (ii) *slow* é predominante na Oficina de Vagões e na Torre L, que apresentam ainda irrisória porção *fast*; (iii) Posto 4 do CCO é o único canal que não é majoritariamente *slow*, seu *pace* médio figura na categoria *conversational* juntamente com o Posto 1 do CCO; (iv) os canais que possuem mais de 10% de sua composição na classe *fast* são os únicos com *pace* acima de 100 WPM e ainda são os canais com maior número de transcritos (todos acima de 170 rótulos criados).

7.1.4 Matriz de Correlação

A Figura 7 traz uma visualização através de histogramas para a duração, a quantidade de palavras e o *speech pace* do *dataset*. Por meio de sua análise é possível perceber detalhes estatísticos referentes à distribuição dos dados e repetibilidade de características, auxiliando a perceber o comportamento do conjunto de dados e identificar pontos destoantes, comumente chamados de *outliers*. É típico o comportamento dos histogramas obtidos e estes apresentam a característica de alcançar o mais alto patamar antes do primeiro terço do eixo das ordenadas. À exceção do histograma de duração, por este ser um critério de exclusão dos dados, os demais apresentam um comportamento de queda lenta com densidade de dados baixa quanto mais à direita do gráfico.

Figura 7 – Histogramas de (a) duração, (b) *pace* e (c) quantidade de palavras



Fonte: Produção do próprio autor.

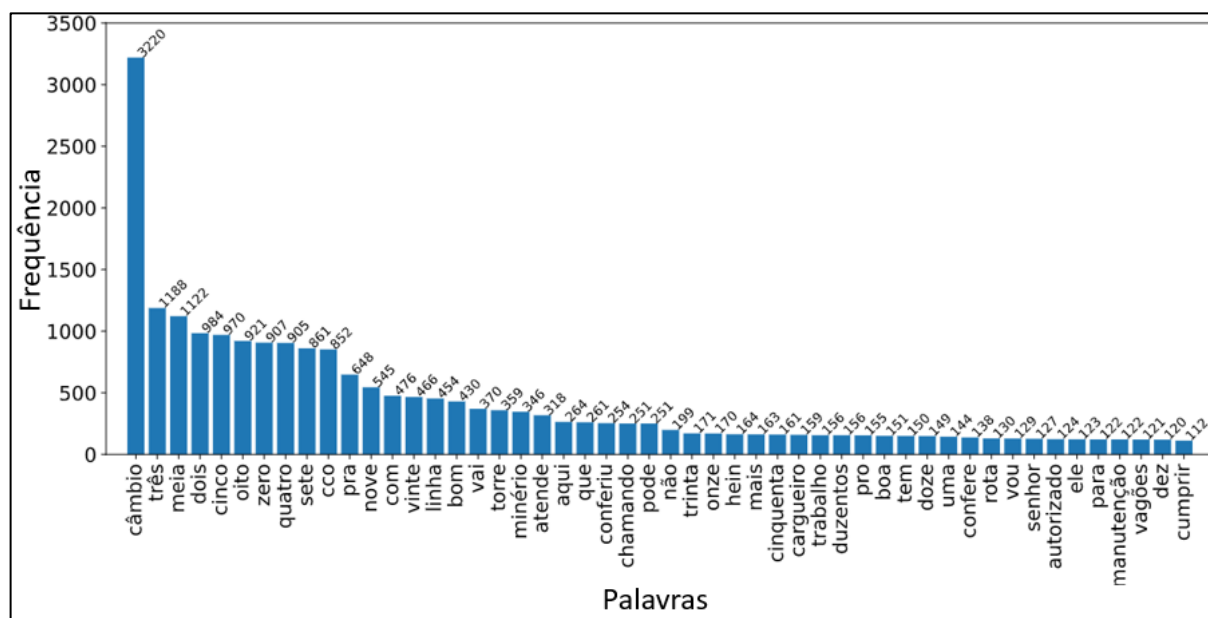
Buscando uma maneira de compreender a influência de uma variável sobre as outras, motivado pela Figura 7, buscou-se uma maneira de realizar essa análise. Através da matriz de correlação é possível traçar de modo analítico e visual tais influências. O Quadro 8 traz a matriz de correlação criada a partir do valor médio de cada característica já apresentada.

No Quadro 8 é possível notar que a quantidade de áudios transcritos não tem influência expressiva sobre nenhuma das demais características. A duração de um áudio impacta fortemente a quantidade de palavras e influencia ainda o *pace* do canal. A quantidade de palavras, por sua vez, tem forte ligação com o *pace* e com a duração, o que é plenamente justificável, uma vez que a obtenção do *pace* se dá justamente através de operações envolvendo exclusivamente duração e quantidade de palavras.

Um item que chama atenção quando da observação atenta do diagrama é a ocorrência do termo **brigado**. Como proposto no item 5.3, as palavras devem ser transcritas sem nenhum tipo de correção de pronúncia, respeitando omissões silábicas e semelhantes. Tal requisito foi obedecido. Outro requisito é a escrita por extenso de números, cujo resultado é percebido pela presença de diversos números no diagrama.

Para uma análise mais direta e objetiva, foi elaborado um gráfico que apresenta as cinquenta palavras mais comuns, desconsiderando aquelas que possuem duas letras ou menos. Essa análise mais quantitativa é válida por perceber a importância de cada um dos verbetes.

Gráfico 15 –Maiores recorrências no *dataset* para palavras com mais de duas letras



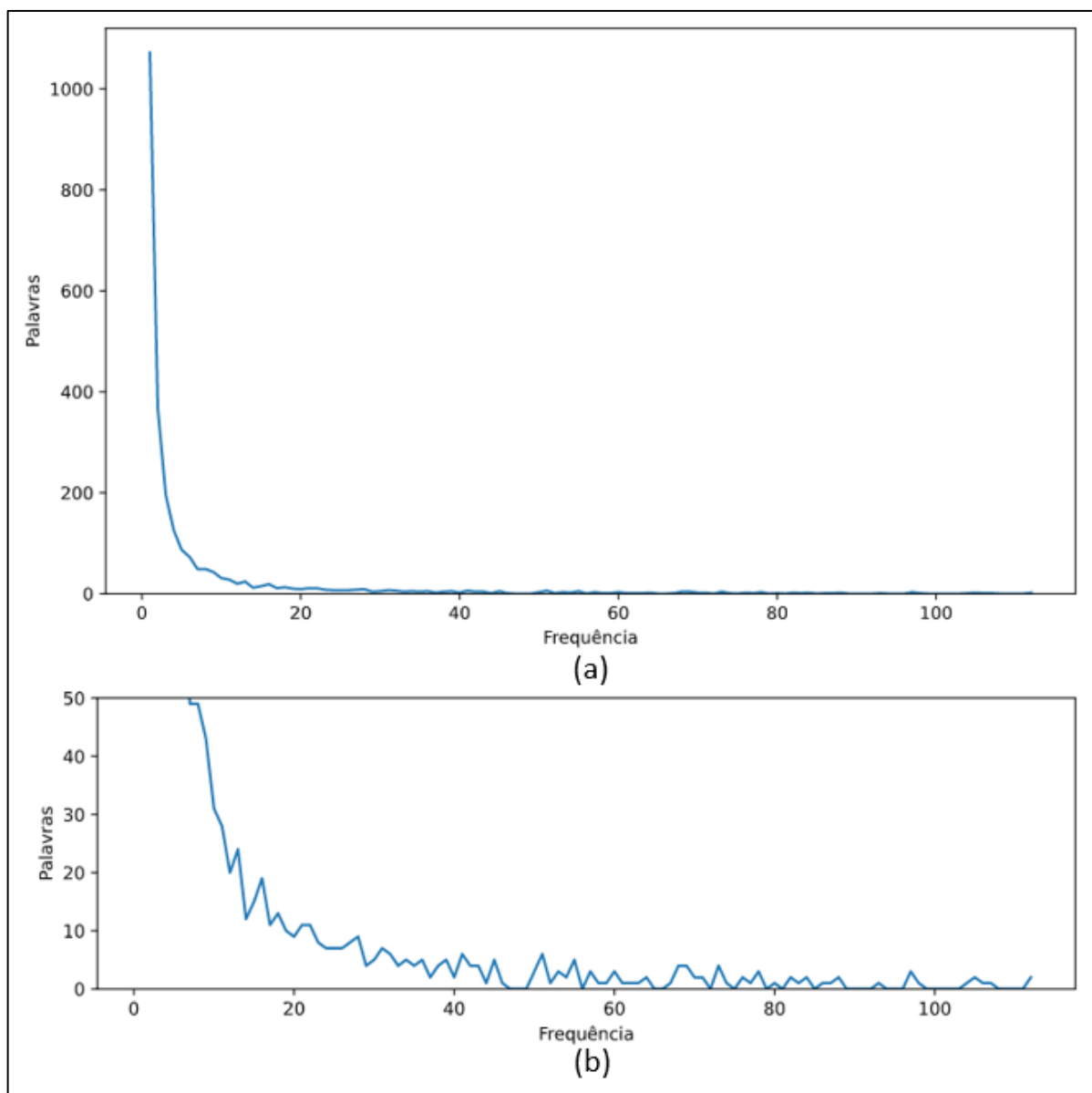
Fonte: Produção do próprio autor.

A palavra *câmbio* de fato é um ponto fora da curva. Se somados dentre os vinte verbetes mais recorrentes, aqueles que não representam números, têm-se o montante de 3129 aparições, valor ainda assim inferior à marca da palavra *câmbio*, que representa 6,4% das palavras do *dataset*. Observando o Gráfico 13, é possível ainda notar que a palavra *câmbio* tem uma representação superior à contribuição individual de todos os canais diferentes da Torre A e postos do CCO.

Para uma compreensão mais abrangente da distribuição de frequência das palavras no *dataset*, o Gráfico 16 faz uma abordagem contrária e complementar ao Gráfico 15. Enquanto o anterior traz a quantidade de ocorrências por palavra até a 50ª palavra, que possui 112 ocorrências, o

atual exibe a quantidade de palavras que se repetem (eixo) x vezes, excursando até a frequência de 112 ocorrências. Assim sendo, é clara a percepção de que o conjunto de palavras é bem esparso, uma vez que aproximadamente um terço das palavras aparecem apenas uma vez, como se percebe pelo queda brusca da curva que contabiliza a quantidade de palavras do Gráfico 16. Outro fato de relevância é que 29 das palavras mais recorrentes representam 50% das entradas do dicionário obtido. O Gráfico 16 (b) exibe com maior detalhe as alterações na quantidade de palavras de maior frequência.

Gráfico 16 – Quantidade de palavras por frequência



Fonte: Produção do próprio autor.

Apesar da baixa frequência, a diversidade de palavras contempladas pelo *dataset* é essencial para um treinamento mais abrangente e que possa ser capaz de identificar mais verbetes.

7.3 Comparativo Outros *Datasets*

Algumas das métricas apresentadas na Seção 7.1 são pontos comparativos válidos na tentativa de observar a expressividade e abrangência dos dados utilizados na montagem do *dataset*. Quintanilha, Biscainho e Netto (2020) elencam as características apresentadas no Quadro 9 referentes aos *datasets* analisados na Seção 2.4. A última linha do referido quadro traz os dados do *dataset* próprio, denominado ACV, de forma a remeter ao projeto ao qual está ligado este trabalho.

Quadro 9 – Métricas dos *datasets* abertos

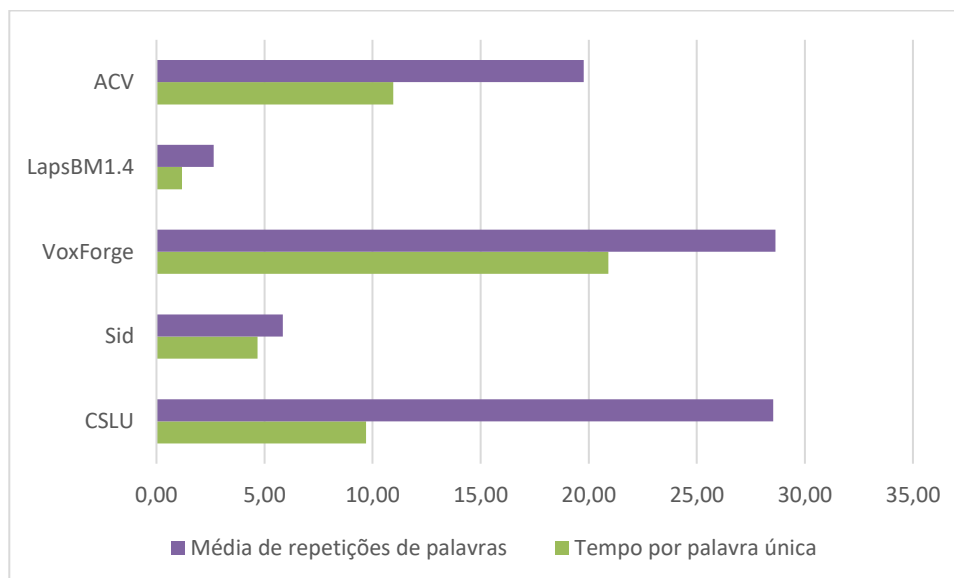
<i>Dataset</i>	Duração (HH:MM)	Total de palavras	Palavras únicas	Falantes
CSLU	1:35	16776	588	477
Sid	7:23	33189	5676	72
VoxForge	4:14	20879	729	111
LapsBM1.4	0:54	7228	2731	35
ACV	7:45	50338	2547	-

Fonte: Quintanilha, Biscainho e Netto (2020) e produção do próprio autor.

A primeira observação que é feita quanto aos dados do *dataset* de produção autoral (ACV) é a impossibilidade de efetuar a contagem de falantes, uma vez que cada áudio possui dois falantes que podem ou não se repetir nos demais áudios. Em relação às demais características, vale ressaltar que o *dataset* apresentado neste trabalho possui a maior duração e quantidade total de palavras pronunciadas, sendo ainda o terceiro vocabulário mais extenso.

Para melhor ilustrar as características de cada *dataset* foi elaborado um gráfico que contrasta a relação entre o tempo e as palavras únicas e também entre quantidade de palavras totais e palavras únicas (repetição média de verbetes). O Gráfico 17 permite perceber que a média de repetições do ACV, sendo inferior apenas ao CSLU e ao VoxForge, apresenta um valor satisfatório. Quanto ao tempo por palavra única, estando ligeiramente acima do valor do CSLU, o ACV alcança a segunda posição. De uma forma geral, o *dataset* apresentado neste documento assume características dentro da variabilidade dos conjuntos existentes, podendo ser considerado um *dataset* robusto, tal qual os demais.

Gráfico 17 – Relação da duração e palavras como vocabulário



Fonte: Produção do próprio autor.

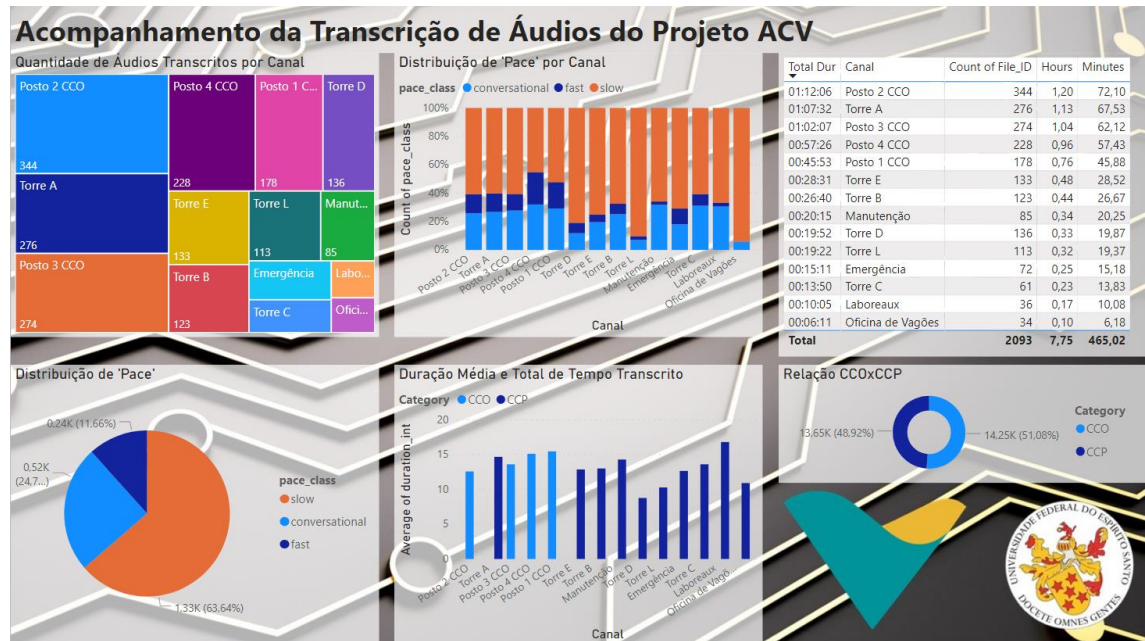
7.4 PowerBI

Para acompanhamento do andamento das transcrições e dados coletados do banco de dados da empresa, foi desenvolvido e compartilhado periodicamente com a equipe de Engenharia Ferroviária um painel desenvolvido na aplicação *PowerBI*. Este sintetiza os principais dados como duração de áudios coletados, distribuição entre canais, duração média e relação entre CCO e CCP.

A Figura 9 ilustra o painel e os gráficos elaborados. Os únicos itens de cunho não meramente estatístico, dentre os gráficos apresentados, são aqueles referentes ao *pace* característico do canal. Por ser algo de simples resolução, estes eram tratados como indicadores apenas para controle interno no contexto do projeto (sem qualquer divulgação), justamente para evitar qualquer interferência no processo de comunicação ou na maneira como este é conduzido.

A apresentação dos dados foi feita de forma mensal, por meio de reuniões entre a equipe do projeto, liderada pelo Prof. Dr. Jorge Leonid Aching Samatelo, com os responsáveis da empresa, bem como reuniões do autor com a engenharia da ferrovia na condição de estagiário da empresa.

Figura 9 – Painel de acompanhamento no *PowerBI*



Fonte: Produção do próprio autor.

8 APLICAÇÃO EM UMA REDE NEURAL

Inicialmente, para que o conjunto de dados possa ser utilizado em uma rede neural, foi necessário fazer a divisão deste entre *train* e *valid*. A proporção adotada foi de 80-20, sendo 80% do conjunto reservado para treino e 20%, para validação. Para que se mantivesse a distribuição dos dados, tal proporção foi seguida para cada pasta apresentada na Figura 5, uma vez que estas reúnem os áudios de mesmo canal e mesmo momento de coleta.

Para o treinamento da rede neural é interessante fazer uso de diversos *datasets* voltados para língua portuguesa com o objetivo de aumentar a massa de dados utilizada, expandindo o vocabulário. Dessa forma, para a aplicação do *dataset* foram obtidos outros dois *datasets* descritos e utilizados por Quintanilha, Biscainho e Netto (2020), a saber BRSDv1 e BRSDv2. O primeiro é composto pelos *datasets* públicos listados no Quadro 9, enquanto o segundo é composto por um *dataset* único denominado CETUC, possuidor de 144 horas de gravação.

Para a aplicação do conjunto de dados, foram disponibilizadas duas versões do *dataset*: *raw* e *clean*. A primeira reúne os dados sem qualquer tipo de supressão de ruído, enquanto a segunda versão detém os mesmos áudios após serem processados pela ferramenta de supressão de ruído apresentada na Seção 6. A rede neural será utilizada também para avaliar a eficácia da supressão de ruído aplicada.

Para aferição dos ganhos obtidos com o uso do conjunto de dados apresentado neste documento, foi feita uma série de testes em uma rede neural conforme estruturada por Scart (2019). Os testes consistiram em avaliar a taxa de erro de caracteres (CER do inglês, *Character Error Rate*) para os cenários apresentados no Quadro 10.

Quadro 10 – Cenários de teste

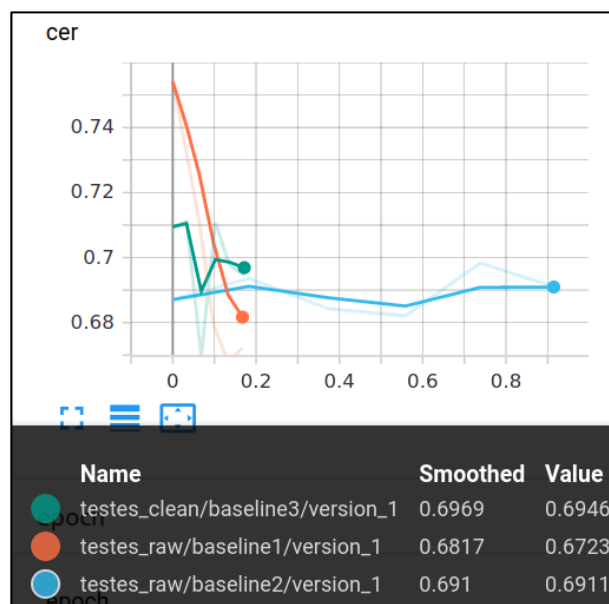
Experimento	Treino	Validação
<i>Baseline 1</i>	BRSDv1	ACV <i>raw</i>
Experimento 1	BRSDv1 + ACV <i>raw</i>	ACV <i>raw</i>
Experimento 2	ACV <i>raw</i>	ACV <i>raw</i>
<i>Baseline 2</i>	BRSDv2	ACV <i>raw</i>
Experimento 3	BRSDv2 + ACV <i>raw</i>	ACV <i>raw</i>
<i>Baseline 3</i>	BRSDv1	ACV <i>clean</i>
Experimento 4	BRSDv1 + ACV <i>clean</i>	ACV <i>clean</i>
Experimento 5	ACV <i>clean</i>	ACV <i>clean</i>

Fonte: Produção do próprio autor.

Os experimentos denominados *baseline* tem por objetivo estabelecer a resposta da rede neural para o problema da transcrição de fala aplicada em áudios da ferrovia sem que haja uso deste *dataset* no treinamento. Os experimentos 1, 3 e 4 avaliam o uso dos *datasets* públicos em conjunto com o ACV na etapa de treinamento. Os experimentos 2 e 5, por fim, buscam averiguar a resposta da rede quando treinada exclusivamente com os dados apresentados neste trabalho.

A primeira análise de interesse abordada consiste na comparação dos *baselines* estabelecidos, isto é, uma observação do estado da arte (no tocante a *datasets*) aplicado ao problema da transcrição da comunicação ferroviária. Os testes envolveram o BRSDv1 sendo validado com o ACV *raw e clean* (*baselines* 1 e 3) e o BRSDv2 com validação do ACV *raw*. É possível observar na Figura 10 que não existe diferença significativa nas variações apresentadas posto que há uma variação da ordem de 2% e, portanto, ambos conjuntos de dados públicos trariam o mesmo resultado, ou seja, uma taxa de erro próxima de 70%.

Figura 10 – Comparativo entre as *baselines*

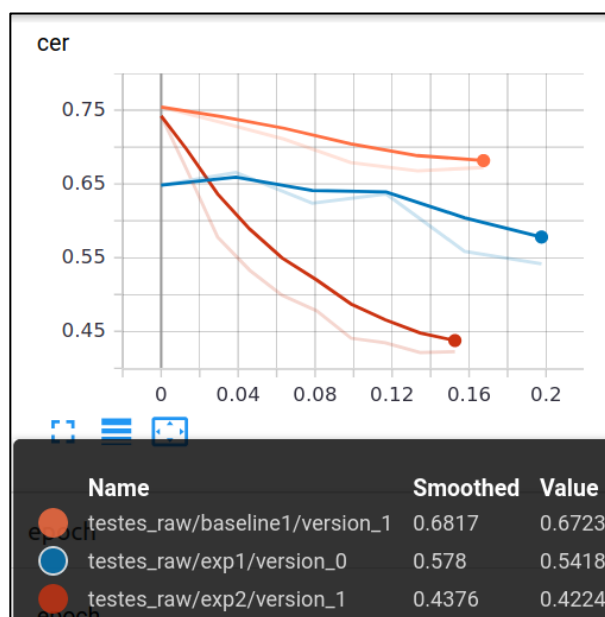


Fonte: Produção do próprio autor.

A segunda análise feita consiste na observação do comportamento do treinamento envolvendo o *dataset* BRSDv1. As curvas apresentadas na Figura 11 representam três condições, isto é, o *baseline* que realiza o treinamento apenas com *datasets* públicos, o experimento 1 no qual o treinamento é misto entre BRSDv1 e ACV *raw* e o experimento 2 sendo usado para treinamento e validação exclusivamente do ACV *raw*.

De acordo com os valores apresentados, há um ganho considerável quando se insere o *dataset* ACV no treino da rede neural, levando o CER a cair de aproximadamente 70% para 54%. Quando se faz o treinamento exclusivamente com o conjunto de testes próprio, o resultado salta para o valor de 42%. Dessa forma, utilizar o ACV *raw* como *dataset* de treino e validação melhora o CER em quase 40% em relação ao *baseline*.

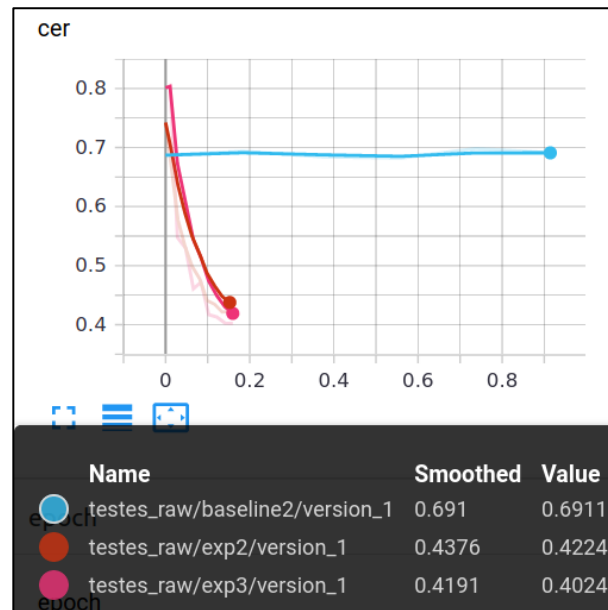
Figura 11 – Testes envolvendo BRSDv1 e ACV *raw*



Fonte: Produção do próprio autor.

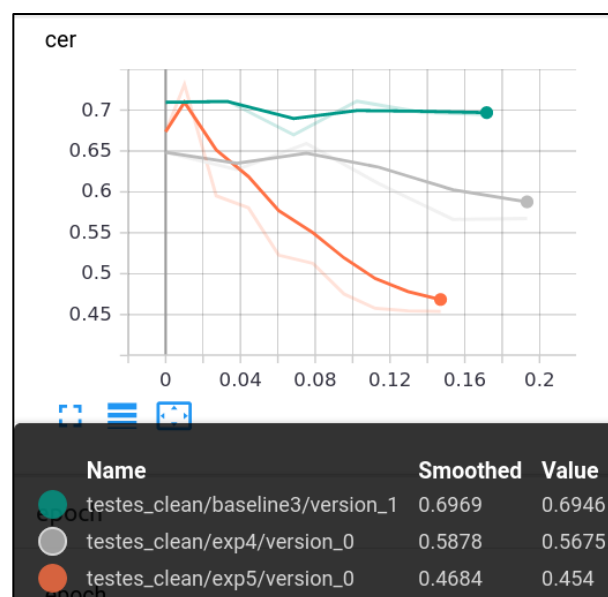
A metodologia do teste supracitado foi repetida alterando-se apenas o *dataset* público utilizado. Fazendo uso do BRSDv2 como *baseline*, foram obtidas as curvas apresentadas na Figura 12. Vale ressaltar que este possui 144 horas de conteúdo de voz (QUINTANILHA; BISCAINHO; NETTO, 2020), o que explica a discrepância entre a duração do treinamento da rede neural.

O BRSDv2, embora apresente o mesmo nível de assertividade quando utilizado sozinho na etapa de treinamento, percebe-se um ganho considerável no seu emprego em conjunto com o ACV *raw*, alcançando o CER de 40% (experimento 3). Este valor é ainda inferior ao obtido no experimento 2 (uso exclusivo de *dataset* próprio), aprimorando sua marca em aproximadamente 5% (de 42% para 40%). Em relação ao *baseline*, há uma evolução de pouco mais de 42%.

Figura 12 – Testes envolvendo BRSDv2 e ACV *raw*

Fonte: Produção do próprio autor.

Em seguida, na Figura 13, foi avaliado o desempenho do *dataset* montado a partir dos áudios limpos de ruído. As curvas apresentam um comportamento bem próximo do observado na Figura 11, sendo o melhor resultado também obtido com o uso exclusivo do *dataset* próprio, alcançando CER de 45%.

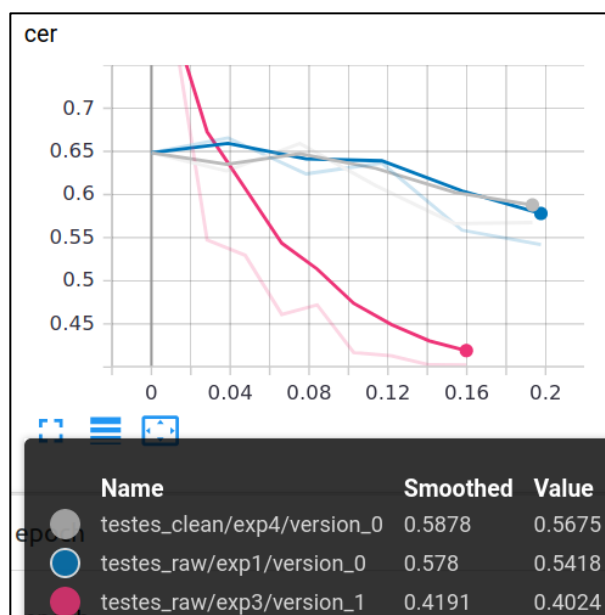
Figura 13 – Testes envolvendo BRSDv1 e ACV *clean*

Fonte: Produção do próprio autor.

Numa tentativa de observar alguma correlação entre a composição dos treinamentos e os resultados obtidos, foram traçadas as curvas referentes aos três experimentos cujo treinamento é realizado com *datasets* públicos e próprios, a saber, os experimentos de índice 1, 3 e 4.

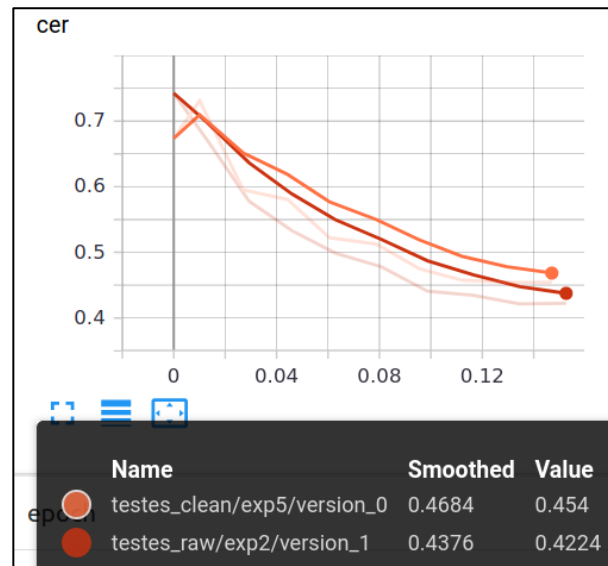
Na Figura 14 pode ser percebido um comportamento similar (variação inferior a 2,5%) entre os testes envolvendo o BRSDv1. O BRSDv2, por sua vez, alcançou o menor dos valores registrados nos experimentos desenvolvidos (CER de 40%). Em comparação com as demais curvas, há uma melhora média da ordem de 27%. Considerando que o método utilizado foi o mesmo, o BRSDv2 se apresenta como mais interessante para o contexto aqui apresentado.

Figura 14 – Comparativos entre testes mistos



Fonte: Produção do próprio autor.

Por fim, foi feita uma comparação dos experimentos que utilizam apenas o *dataset* próprio tanto para treinamento quanto para validação. Ao contrário do esperado, o conjunto que obteve o melhor desempenho não passou pelo processo de supressão de ruído. Enquanto o ACV *clean* alcançou o CER de 45%, o ACV *raw* obteve apenas 42%. Dessa forma, utilizar dados ruidosos parece aumentar a qualidade da transcrição em quase 2,5% em relação ao uso de áudios tratados.

Figura 15 – Comparativo entre *datasets* próprios

Fonte: Produção do próprio autor.

Durante os testes, o desempenho do conjunto de dados ACV *clean* se mostrou inferior ao ACV *raw*. A Figura 11, a Figura 13 e a Figura 15 mostram que, para todos os cenários, tanto de treino exclusivo quanto misto, o desempenho do ACV *raw* foi superior. Dessa forma, optou-se por não realizar os testes envolvendo o ACV *clean* e o BRSDv2.

9 CRIAÇÃO DE UM ROTEIRO

Após construir o *dataset* e fazer uma avaliação superficial do vocabulário e da recorrência de cada verbete, foi percebida a necessidade de que a rede neural compreenda algumas palavras críticas com uma assertividade maior. Para reforçar o aprendizado desses casos específicos, com a anuência da equipe de Engenharia de Operação, optou-se por desenvolver um roteiro a ser enunciado pelos controladores que trabalham em um local fixo.

O texto apresentado no Apêndice A – ROTEIRO DE ENUNCIÇÃO busca simular algumas características da comunicação ferroviária, embora as frases elaboradas não necessariamente façam sentido com operações reais. Foram elencadas as palavras mais recorrentes e, dentre estas, foram retiradas algumas para compor o vocabulário do enunciado. Por outro lado, em conjunto com a equipe de Engenharia Ferroviária foram feitas reuniões entre o autor e membros que já exerceram atividades de comunicação de rádio. Tais encontros tiveram por objetivo discutir e perceber quais as palavras de maior importância que não foram contempladas no texto lúdico preparado.

Com o emprego do roteiro, existe uma mudança no procedimento para coleta de dados rotulados. Este trabalho apresenta uma técnica de rotulação de sinais de áudio previamente produzidos, enquanto o roteiro de enunciação se comporta como um rótulo a ser convertido em sinal de áudio.

No tocante à aplicabilidade, a obtenção de diversos registros com o mesmo rótulo aumenta a quantidade de repetições das palavras importantes e, em tese, pode auxiliar a reduzir o erro observado nestas. Em termos de praticidade, uma vez que o rótulo está definido, será necessário um menor esforço na rotulação e organização dos dados, sendo sugerida a segmentação dos dados arquivos de diferentes durações. Por exemplo, os dados de um orador são divididos em 5 partes de igual tamanho, enquanto o do segundo orador, contendo o mesmo rótulo é particionado em 4 porções. Dessa forma, há uma sobreposição dos rótulos, e não uma repetição exata. Os intervalos de 2 a 3 segundos sugeridos no texto a ser enunciado podem facilitar essa tarefa. Tal prática pode evitar ou reduzir as chances de incorrer em *overfitting*, onde a máquina aprende um problema muito específico e perde sua capacidade de resolução de situações similares.

10 CONCLUSÕES E PROJETOS FUTUROS

Este trabalho apresentou uma proposta de confecção de um *dataset* de sinais de áudio e um modelo de rotulação orientado a aplicações de reconhecimento de fala no contexto específico da comunicação de rádio operada pela empresa Vale S.A.

O conjunto de dados final obtido é expressivo e abrangente ao contexto ferroviário, sendo balanceado entre as categorias que fazem uso do sistema de rádio da EFVM. Seu vocabulário contempla as mais diversas peculiaridades da rotina da ferrovia.

Foi utilizada como metodologia de construção do conjunto de dados a rotulação interna, aliada ao desenvolvimento de algumas ferramentas para gerenciamento e tratamento de rótulos. O *dataset* obtido apresentou-se robusto em contraste com as opções abertas disponíveis. Quando comparado ao estado da arte disponível no início do desenvolvimento (*datasets* que compõem o BRDSv1, que foi utilizado como *benchmark* quando da proposta deste trabalho), o conjunto de dados desenvolvido chega a possuir uma duração superior, vocabulário de tamanho compatível e média de repetição de palavras acima da média para *datasets* de sua duração.

É de suma importância o uso deste conjunto de dados nos trabalhos futuros relacionados ao projeto de parceria entre empresa e universidade, uma vez que seu emprego reduziu a taxa de erros da rede neural em até 42%.

Devido ao alto nível de ruído, principalmente nos trechos de áudio correspondente aos falantes que trabalham no interior de locomotivas, a qualidade do áudio coletado é baixa. Apesar disso, foi obtido um CER de 40%. Para efeitos de comparação, as referências estudadas e citadas nesse documento alcançam a marca de CER de até 25% trabalhando com sinais consideravelmente menos ruidosos.

A ferramenta de supressão de ruído desenvolvida e relatada neste documento se mostrou ineficaz na tentativa de melhorar os resultados da rede neural. Embora tenha trazido mais clareza para percepções humanas, tal efeito não se repetiu na extração de características dos sinais. Sugere-se o estudo e desenvolvimento de ferramentas de supressão de ruído baseadas

em outros modelos que não necessitam de um perfil de ruído. O uso de redes neurais na remoção de ruídos, é uma opção.

O roteiro de enunciação proposto está em trâmite junto à equipe da Engenharia de Operação da EFVM. Uma vez que seja realizada a narração pelos controladores, sugere-se a coleta dos dados para uso em duas frentes distintas: criação de um *dataset* para reconhecimento de fala (motivo de seu desenvolvimento) e outro para reconhecimento de oradores.

Aconselha-se o desenvolvimento de uma ferramenta capaz de isolar os oradores distintos dentro de um mesmo arquivo de áudio. Dessa forma, pode-se obter um melhor resultado na remoção de ruído, além de apresentar uma integração interessante com sugestão de identificação de oradores. Os dados brutos coletados neste trabalho podem ser utilizados para desenvolvimento de *datasets* voltados para estas aplicações.

Por fim, recomenda-se a ampliação do banco de dados com o intuito de alcançar resultados melhores.

REFERÊNCIAS BIBLIOGRÁFICAS

ALTEXSOFT. **How to Organize Data Labeling for Machine Learning: Approaches and Tools**. 2018. Disponível em: <https://www.kdnuggets.com/2018/05/data-labeling-machine-learning.html>. Acesso em: 28 out. 2019

ANDERL, R; STRANG, D. **Assembly process driven component data model in cyber-physical production systems**. 2014 apud NETO, A. A.; PEREIRA, G. B.; DROZDA, F. O.; SANTOS, A. P. L. A Busca de uma Identidade para a Indústria 4.0. **Brazilian Journal of Development**, Curitiba, v. 4, n. 4, p. 1379-1395, jul./set. 2018. Disponível em: <https://www.brazilianjournals.com/index.php/BRJD/article/view/183>. Acesso em: 19 nov. 2020.

ANTONELI, G. C.; NEITZEL, I. Aplicação de redes neurais artificiais na indústria de fios de algodão. **GEPROS. Gestão da Produção, Operações e Sistemas**, Bauru, ano 11, n. 2, p. 1-20, abr.-jun. 2016. Disponível em: <https://revista.feb.unesp.br/index.php/gepros/article/view/1355>. Acesso em: 19 nov. 2020.

BAGWELL, C. **SoX**. 2013. Disponível em: <http://sox.sourceforge.net/soxformat.html#:~:text=SoX%20can%20read%20and%20write,overriding%20the%20file%20type%2C%20e.g>. Acesso em: 12 out. 2020.

BOLL, S. Suppression of acoustic noise in speech using spectral subtraction. **IEEE Transactions on Acoustics, Speech, and Signal Processing**, v. 27, n. 12, p.113-120, abr. 1979. Disponível em: <https://ieeexplore.ieee.org/document/1163209>. Acesso em: 19 nov. 2020.

BRASIL. **EFVM – Estrada de Ferro Vitória a Minas Gerais**. 2018. Disponível em: <https://www.ppi.gov.br/efvm-estrada-de-ferro-vitoria-a-minas>. Acesso em: 18 out. 2020.

BRIAN, A.; SHI, Y.Q. **Mp3 bit rate quality detection through frequency spectrum analysis**. 2009. Disponível em: https://www.researchgate.net/publication/228941766_Mp3_bit_rate_quality_detection_through_frequency_spectrum_analysis. Acesso em: 12 out. 2020.

BROWN, A.; GARG, S.; MONTGOMERY, J. Automatic and efficient denoising of bioacoustics recordings using MMSE STSA. **IEEE Access**, v. 6, p. 5010-5022, fev. 2018. Disponível em: <https://ieeexplore.ieee.org/document/8194836>. Acesso em: 19 nov. 2020.

FARIA, J. **Entendendo o que é um codec de áudio**. 2016. Disponível em: <https://www.stereotool.com.br/entendendo-codec-de-audio>. Acesso em: 17 out. 2020.

FISCHLER, M.A; FIRSCHEIN, O. **Intelligence: The Eye, The Brain and The Computer**. Massachusetts: Addison – Wesley, 1987 apud HAYKIN, S. **Neural Networks and Learning Machines**. 3. ed. Hamilton: Prentice Hall, 2009.

GANDHI, A. **Data Augmentation: How to use Deep Learning when you have Limited Data – Part 2**. 2018. Disponível em: <https://nanonets.com/blog/data-augmentation-how-to-use-deep-learning-when-you-have-limited-data-part-2/>. Acesso em: 17 out. 2020.

GONÇALVES, L. A. **Um estudo sobre a Transformada de Fourier e seu uso em processamento de imagens**. 2004. Dissertação (Mestrado em Matemática Aplicada) – Programa de Pós-Graduação em Matemática Aplicada, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2004. Disponível em: <https://www.lume.ufrgs.br/bitstream/handle/10183/6748/000446124.pdf>. Acesso em: 19 nov. 2020.

HAYKIN, S. **Neural Networks and Learning Machines**. 3. ed. Hamilton: Prentice Hall, 2009.

HUERTA, J.; RICHARD, S. **Speech recognition from GSM codec parameters**. 1997. Disponível em: https://www.researchgate.net/publication/2599643_Speech_Recognition_From_Gsm_Codec_Parameters. Acesso em: 12 out. 2020.

LOPES, R. C. S. **A relação professor-aluno e o processo de ensino-aprendizagem**. 2017. Disponível em: <http://www.diaadiaeducacao.pr.gov.br/portals/pde/arquivos/1534-8.pdf>. Acesso em: 11 out. 2020.

MCFFEE, B.; RAFFEL, C.; LIANG, D.; ELLIS, D.P.W.; MCVICAR, M.; BATTENBERG, E.; NIETO, O. *librosa: Audio and music signals analysis in python*. In: SciPy. 14., 2015, Austin. **Proceedings** [...]. Austin, 2015. p. 18-25. Disponível em: http://conference.scipy.org/proceedings/scipy2015/pdfs/brian_mcftee.pdf. Acesso em: 19 nov. 2020.

MICROPYRAMID. **Understanding audio quality: bit rate, sample rate**. 2017. Disponível em: <https://medium.com/@MicroPyramid/understanding-audio-quality-bit-rate-sample-rate-14286953d71f>. Acesso em: 12 out. 2020.

MINISTÉRIO DA INDÚSTRIA, COMÉRCIO E SERVIÇOS. **Agenda Brasileira para a Indústria 4.0**. 2019. Disponível em: <http://www.industria40.gov.br>. Acesso em: 26 out. 2019.

MONTEIRO, L. V.; CARNEIRO, M. P.; MOREIRA, F. C. Aplicação de Redes Neurais Artificiais para Redução da Variabilidade no Processo Produtivo de uma Indústria Alimentícia. In: ENCONTRO NACIONAL DE ENGENHARIA DE PRODUÇÃO, 32., 2012, Bento Gonçalves. **Anais**[...]. Bento Gonçalves: 2012. v. 1, p. 1-13. Disponível em: http://www.abepro.org.br/biblioteca/enegep2012_TN_STP_158_921_19801.pdf. Acesso em: 19 nov. 2020.

NASSIF, A. B.; SHAHIN, I.; ATTILI, I.; AZZEH, M.; SHAALAN, K. Speech recognition using deep neural networks: A systematic review. **IEEE Access**, v. 7, p. 19143–19165, fev. 2019. Disponível em: <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8632885>. Acesso em: 19 nov. 2020.

NETO, A. A.; PEREIRA, G. B.; DROZDA, F. O.; SANTOS, A. P. L. A Busca de uma Identidade para a Indústria 4.0. **Brazilian Journal of Development**, Curitiba, v. 4, n. 4, p. 1379-1395, jul./set. 2018. Disponível em: <https://www.brazilianjournals.com/index.php/BRJD/article/view/183>. Acesso em: 19 nov. 2020.

NYQUIST, H. Certain topics in telegraph transmission theory. **Transactions of the American Institute of Electrical Engineers**, p. 617-644, abr. 1928. Reimpressão, Nova Iorque 2002. Disponível em: https://monoskop.org/images/2/2e/Nyquist_Harry_1928_Certain_Topics_in_Telegraph_Transmission_Theory.pdf. Acesso em: 19 nov. 2020.

OCHI, L. S.; DIAS, C. R.; SOARES, S. S. F. Clusterização em Mineração de Dados. *In*: ENCONTRO REGIONAL DE INFORMÁTICA RJ/ES, 2004, Vitória. **Anais [...]**. Vitória: 2004. 1 CD-ROM. Disponível em: https://www.researchgate.net/publication/251910507_Clusterizacao_em_Mineracao_de_Dados. Acesso em: 12 out. 2020.

QUINTANILHA, I. M.; BISCAINHO, L. W. P.; NETTO, S. L. An open-source end-to-end ASR system for Brazilian Portuguese using DNNs built from newly assembled corpora. **Journal of Communication and Information Systems**, v. 35, n. 1, p. 230-242, jan. 2020. Disponível em: <https://jcis.sbrt.org.br/jcis/article/view/721/498>. Acesso em: 19 nov. 2020.

QUINTANILHA, I. M. **End-to-End Speech Recognition Applied to Brazilian Portuguese Using Deep Learning**. 2017. Dissertação (Mestrado em Engenharia Elétrica) – Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2017. Disponível em: <http://www.pee.ufrj.br/index.php/pt/producao-academica/dissertacoes-de-mestrado/2017/2016033174-end-to-end-speech-recognition-applied-to-brazilian-portuguese-using-deep-learning/file>. Acesso em: 19 nov. 2020.

ROCHMAN, D. **High Quality AI and Machine Learning Data Labeling at Scale: A Brief Research Report**, 2019. Disponível em: <https://www.kdnuggets.com/2019/07/high-quality-ai-machine-learning-data-labeling-research-report.html>. Acesso em: 28 out. 2019.

SCART, L. G. **Reconhecimento automático de fala em português utilizando arquiteturas de redes neurais profundas**. 2019. Trabalho de Conclusão de Curso (Graduação em Engenharia Elétrica) – Centro Tecnológico, Universidade Federal do Espírito Santo, Vitória, 2019.

SEN, S.; DUTTA, A.; DEY, N. **Audio Processing and Speech Recognition**. Singapura: Springer, 2019.

SOTERO FILHO, R. F. B. **Novas abordagens para codificação de voz e reconhecimento automático de locutor projetadas via mascaramento pleno em frequência por oitava**. 2017. Dissertação (Mestrado em Engenharia Elétrica) – Programa de Pós-Graduação em Engenharia Elétrica, Centro de Tecnologia e Geociência, Universidade Federal de Pernambuco, Recife, 2009. Disponível em: <https://repositorio.ufpe.br/bitstream/123456789/26231/1/DISSERTA%C3%87%C3%83O%20Roberto%20Fernando%20Batista%20Sotero%20Filho.pdf>. Acesso em: 19 nov. 2020.

SPEAKERHUB. **Your speech pace: guide to speeding and slowing down**. 2017. Disponível em: <https://medium.com/@speakerhubHQ/your-speech-pace-guide-to-speeding-and-slowing-down-be150dcb9cd7>. Acesso em: 8 jun. 2020

TORRES JÚNIOR, R. G.; MACHADO, M. A. S.; BARRETO, J. M. Previsão de Falhas em Manutenção Industrial Usando Redes Neurais. *In*: SEGeT – SIMPÓSIO DE EXCELÊNCIA

EM GESTÃO E TECNOLOGIA, 3., 2006, Resende. **Anais** [...]. Resende. 2006. v.1, p. 1-9. Disponível em: https://www.aedb.br/seget/arquivos/artigos06/403_Rubiao-FORMATADO.pdf. Acesso em: 19 nov. 2020.

VOXFORGE. **VoxForge**. 2006. Disponível em: <https://www.voxforge.org>. Acesso em: 01 nov. 2019.

APÊNDICE A – ROTEIRO DE ENUNCIÇÃO

O roteiro de enunciação deverá ser lido fielmente, com entonação e velocidades usuais, respeitando os períodos de pausa determinados e sem encerrar a comunicação até o fim da narração.

Bom dia torre A, minério 1 na linha 3 em contato com o CCO pedindo rota pra entrar no pátio e manobrar na pera.

[pause 2-3s]

Passageiro cruzando com o cargueiro na intermediária RH 9, câmbio

[pause 2-3s]

Manutenção 2 chamando CCM pra abrir uma LDL abaixo da housing 9.

[pause 2-3s]

Boa tarde oficial cobrindo recuo na linha do hump yard, olhando agora aqui tá autorizado a recuar mais 10 vagões, entendido?

[pause 2-3s]

Manutenção 02 chamando operador, então o senhor confere pra mim se o shunt de verificação tá na linha um ou se tenho que trocar ele, brigado

[pause 2-3s]

Conferiu aqui, shunt de verificação na linha um, SB de cima.

[pause 2-3s]

Passageiro atendendo CCO, pode confirmar se é ou não pra gente cumprir o PRO ou a restrição até livrar a cauda no circuito de chave?

[pause 2-3s]

Laurindo na entrehouse da 9 com a 10, chamando CCO câmbio.

[pause 2-3s]

Controlador atende equipe do turno, câmbio. CCO, vai vir a equipe pra troca de turno, e assim que eles chegarem eu vou te avisar. Bom trabalho, bom descanso.

[pause 2-3s]

Laurindo chamando CCE, avistei a placa de precaução aqui na sinalizada e tô cumprindo a curva. Pode seguir no permissivo, ou tem alguém em emergência mesmo?

[pause 2-3s]

CCE atende Laurindo. Aciona a emergência, a linha tá interditada. Uma composição avançou acima da velocidade na chave contra. Estamos avaliando se ocorreu descarrilamento, abalroamento, atropelamento ou algum vazamento de combustível. É perigoso passar no travessão com velocidade alta se tiver pra contrária.

[pause 2-3s]

Inspetor chamando Torre A. Dá pra verificar se no *way side* tem alguma coisa de *hot box* ou *hot wheel* nessa composição, câmbio?