LUCAS HILÁRIO DA COSTA

CLASSIFICAÇÃO DA QUALIDADE DO SINAL DE VOZ EM COMUNICAÇÃO VOIP UTILIZANDO DEEP LEARNING

LAVRAS - MG

2019

LUCAS HILÁRIO DA COSTA

CLASSIFICAÇÃO DA QUALIDADE DO SINAL DE VOZ EM COMUNICAÇÃO VOIP UTILIZANDO DEEP LEARNING

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Engenharia de Sistemas e Automação, área de concentração em Sistemas Inteligentes, para a obtenção do título de Mestre.

Prof. DSc. Demóstenes Zegarra Rodríguez
Orientador

Prof. DSc. Renata Lopes Rosa
Coorientadora

LAVRAS - MG

2019

Ficha catalográfica elaborada pelo Sitema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Costa, Lucas Hilário da.
Classificação da qualidade do sinal de voz em comunicação VoIP utilizando Deep Learning / Lucas Hilário da Costa. - 2019.
94 p. : il.

Orientador: Prof. DSc. Demóstenes Zegarra Rodríguez.
Coorientadora: Prof. DSc. Renata Lopes Rosa.
Dissertação (mestrado acadêmico)–Universidade Federal de Lavras, 2019.
Bibliografia.

1. VoIP. 2. Deep Learning. 3. Qualidade da voz. I. Rodríguez, Demóstenes Zegarra. II. Rosa, Renata Lopes. III. Título.

LUCAS HILÁRIO DA COSTA

CLASSIFICAÇÃO DA QUALIDADE DO SINAL DE VOZ EM COMUNICAÇÃO VOIP UTILIZANDO DEEP LEARNING

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Engenharia de Sistemas e Automação, área de concentração em Sistemas Inteligentes, para a obtenção do título de Mestre.

APROVADA em 26 de Junho de 2019.

Prof. DSc. Danton Diego Ferreira DEG / UFLA
Prof. DSc. Dante Coaquira Begazo LPS / USP

Prof. DSc. Demóstenes Zegarra Rodríguez
Orientador

Prof. DSc. Renata Lopes Rosa
Co-Orientadora

LAVRAS - MG
2019

Dedico este trabalho primeiramente a Deus, por ser essencial em minha vida, a minha noiva Daniele, meu pai Geraldo, minha mãe Adalgisa e aos meus irmãos. Dedico também ao meu orientador Prof. Dr. Demóstenes Zegarra Rodríguez, pela sua confiança, paciência, incentivo, amizade e excelente orientação. Sem o apoio de ambos, este trabalho não teria sido realizado. A eles, meu muito obrigado.

AGRADECIMENTOS

Para a realização da presente dissertação, contei com o apoio de diversas pessoas e quero deixar expresso os meus agradecimentos a todas elas:

Inicio meus agradecimentos a DEUS, já que Ele colocou pessoas tão especiais a meu lado que me deram apoio e que sem as quais certamente não teria dado conta.

Ao orientador desta dissertação o Doutor Professor Demóstenes Zegarra Rodríguez, pela orientação prestada, pelo seu incentivo, disponibilidade, apoio e paciência que sempre demonstrou. Aqui lhe exprimo a minha gratidão.

A co-orientadora Doutora Professora Renata Lopes Rosa, pela sua disponibilidade nos trabalhos de laboratório, pela sua disponibilidade e pelo seu apoio na elaboração deste trabalho.

A todos os amigos e colegas que de uma forma direta ou indiretamente, contribuíram e/ou auxiliaram na elaboração do presente estudo, pela paciência, atenção e força que prestaram em momentos difíceis, desde já os meus agradecimentos.

Não poderia deixar de agradecer à minha família por todo o apoio, pela força e pelo carinho que sempre me prestaram ao longo de toda a minha vida acadêmica, bem como, à elaboração da presente dissertação a qual sem o seu apoio teria sido impossível.

A minha noiva por ter caminhado ao meu lado, pela sua paciência, compreensão e ajuda prestada durante a elaboração da presente dissertação, especialmente por apresentar sempre um sorriso, quando sacrificava os dias, as noites, os fins-de-semana e os feriados em prol da realização deste estudo, eu agradeço.

Enfim, quero demonstrar o meu agradecimento, a todos aqueles que, de um modo ou de outro, tornaram possível a realização da presente dissertação.

A todos o meu sincero e profundo Muito Obrigado!

“A persistência é o menor caminho do êxito.”
(Charles Chaplin)

RESUMO

Atualmente a Voz sobre IP (VoIP - *Voice over IP*) é um dos serviços de comunicação mais utilizados, entretanto, sua qualidade está relacionada a diversos fatores externos que ocasionam diversos tipos de degradação do sinal de voz. Nos canais de comunicação, a perda de pacotes afeta significativamente o sinal de voz, fazendo com que a qualidade da comunicação seja menor, afetando diretamente a qualidade de experiência (QoE - *Quality of Experience*) do usuário. O objetivo deste trabalho foi a implementação e desenvolvimento de dois modelos de rede *Deep Learning* (DL) que são capazes de classificar a qualidade do sinal de voz transmitido em uma comunicação VoIP, afetado principalmente pela perda de pacotes. Os modelos propostos foram desenvolvidos utilizando um modelo de rede neural profunda (DNN - *Deep Neural Network*), onde através da análise do sinal da voz afetada pela taxa de perda de pacotes (PLR - *Packet Loss Rate*), dos sinais degradados, foi possível classificá-los em quatro classes distintas de acordo com a experiência do usuário. Para a realização dos testes foram preparadas duas bases de dados, contendo cada uma, quatro classes distintas, onde uma foi preparada com os arquivos da base de dados da recomendação ITU-T P.862, com diferentes taxas de perda de pacotes, e a outra base foi preparada com os arquivos da recomendação ITU-T P.501 de acordo com o índice MOS (*Mean Opinion Score*) de cada arquivo degradado. Para obter as bases de dados foi implementado um programa no MATLAB que degrada arquivos de voz original mudando os valores da taxa de perda de pacotes, após o processamento, os arquivos foram agrupados em quatro classes de acordo com a taxa de perda de pacotes aplicada a cada sinal de voz original. Para a base de dados preparada pelo índice MOS os arquivos degradados foram processados pelo algoritmo da recomendação ITU-T P.862 com o objetivo de determinar o MOS através da comparação do sinal de voz degradado com o sinal original de cada arquivo de áudio e depois agrupados em quatro classes de acordo com o MOS obtido. Para validar os modelos duas bases de dados adicionais foram preparadas contendo arquivos de áudio da base de dados VoxCeleb divididos em quatro classes com 250 arquivos cada, sendo agrupadas pela taxa de PLR e pelo MOS. Os resultados obtidos do modelo utilizando a base de dados preparada pela taxa de perda de pacotes foi de 94% de acurácia na validação e os resultados do modelo para a base de dados preparada pelo MOS foi de 91% de acurácia. O modelo alcançou uma acurácia de 86,96% para a base de dados adicional preparada de acordo com a taxa de perda de pacotes e de 83,29% de acurácia para a base adicional preparada de acordo com o MOS. Para determinar a eficiência do modelo desenvolvido foram comparados os seus resultados com os resultados obtidos pelos algoritmos da recomendações ITU-T P.563 e P.862, onde obteve-se uma média de 53,21% de acerto quando comparamos os resultados da definição do MOS do algoritmo da recomendação ITU-T P.563 com o definido pelo algoritmo da recomendação ITU-T P.862. Através dos resultados obtidos pode-se concluir que os modelos gerados foram capazes de classificar a taxa de perda de pacotes e o índice MOS de forma não intrusiva e com uma ótima taxa de acurácia. Pode-se destacar que quando comparamos os métodos não intrusivos, os resultados obtidos do modelo proposto para o indice MOS que foi de 91% de acurácia foi melhor em comparação dos com os resultados do algoritmo da recomendação ITU-T P.563 que obteve uma taxa 53,21% de acurácia em relação com os resultados do algortimo intrusivo da recomendação ITU-T P.862. Concluindo assim que o modelo gerado é capaz de determinar o MOS dos arquivos de voz degradados de forma mais eficiente que o algoritmo da recomendação ITU-T P.563. Consequentemente, uma contribuição importante deste trabalho é a apresentação de um modelo de avaliação não intrusivo capaz de identificar a qualidade do sinal de voz em tempo real.

Palavras-chave: VoIP, Deep learning, Qualidade da voz, ITU-T P.862, ITU-T P.563, ITU-T P.501

ABSTRACT

Voice over IP (VoIP) is currently one of the most widely used communication services; however, its quality is related to several external factors that cause various types of voice signal degradation. In communication channels, packet loss significantly affects the voice signal, causing lower communication quality, directly affecting the user's quality of experience (QoE). The objective of this work was the implementation and development of two Deep Learning (DL) network models that are able to classify the quality of the voice signal transmitted in a VoIP communication, mainly affected by packet loss. The proposed models were developed using a Deep Neural Network (DNN) model, where through the analysis of the voice signal affected by Packet Loss Rate (PLR) of the degraded signals, it was possible to classify them into four distinct classes according to the user experience. To perform the tests two databases were prepared, each containing four distinct classes, one of which was prepared with the ITU-T P.862 recommendation database files, with different packet loss rates, and the another base was prepared with the ITU-T P.501 recommendation files according to the mean opinion score (MOS) index of each degraded file. To obtain the databases, a program was implemented in MATLAB that degrades original voice files by changing the packet loss rate values. After processing, the files were grouped into four classes according to the packet loss rate applied to each original voice signal. For the database prepared by the MOS index the degraded files were processed by the ITU-T P.862 recommendation algorithm in order to determine the MOS by comparing the degraded voice signal with the original signal of each audio file and then grouped into four classes according to the MOS obtained. To validate the models two additional databases were prepared containing VoxCeleb database audio files divided into four classes with 250 files each, being grouped by PLR rate and MOS. The results obtained from the model using the database prepared by the packet loss rate was 94% accuracy in the validation and the model results for the database prepared by the MOS was 91% accuracy. The model achieved an accuracy of 86.96% for the additional database prepared according to packet loss rate and 83.29% accuracy for the additional database prepared according to MOS. To determine the efficiency of the developed model, its results were compared with the results obtained by the ITU-T recommendations P.563 and P.862 algorithms, where an average of 53.21% accuracy was obtained when comparing the results. MOS definition of the ITU-T P.563 recommendation algorithm with that defined by the ITU-T P.862 recommendation algorithm. From the obtained results it can be concluded that the generated models were able to classify the packet loss rate and the MOS index in a non intrusive way and with an excellent accuracy rate. It can be highlighted that when comparing the non-intrusive methods, the results obtained from the proposed model for the MOS index which was 91% accuracy was better compared to the results from the ITU-T P.563 recommendation algorithm that obtained an accuracy rate of 53.21% compared to the intrusive algorithm results from the ITU-T P.862 recommendation. Thus, the generated model is able to determine the MOS of the degraded voice files more efficiently than the ITU-T P.563 recommendation algorithm. Consequently, an important contribution of this work is the presentation of a non-intrusive evaluation model capable of identifying the real-time voice signal quality.

Keywords: VoIP, Deep Learning, Voice Quality, ITU-T P.862, ITU-T P.563, ITU-T P.501.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

De acordo com o *Cisco Visual Networking Index Forecast*, é previsto que até 2022, o percentual da população mundial que utiliza a Internet alcance 60%, e consequentemente o número de dispositivos conectados à rede aumente de 2,4 para 3,6 dispositivos por pessoa (INDEX, 2019). Com a evolução das infraestruturas de rede existente, houve uma melhoria considerável na qualidade de serviço (QoS - *Quality of Service*) disponibilizada para o usuário final e isto aumentará ainda mais com o desenvolvimento de novas tecnologias, como, por exemplo, a quinta geração de redes de comunicações (5G). Com isso, a demanda de serviços multimídia e as expectativas dos usuários também estão aumentando. Isso levou a uma maior demanda por maior largura de banda resultando em mais desafios para os operadores de rede com recursos já limitados.

Com a crescente popularização de dispositivos móveis, o uso do serviço de VoIP está aumentando cada vez mais na rede, porém, uma vez que a condição da rede se relaciona diretamente com a QoS de VoIP, é essencial monitorar continuamente a qualidade do sinal da voz nas chamadas telefônicas em diversas condições da rede. Os métodos convencionais usam os parâmetros que contêm a informação de condição de rede ou sinal de voz transmitido para determinar a qualidade da comunicação de voz.

Apesar do desenvolvimento da infraestrutura atual da rede de comunicação e de seus recursos, ainda não é possível acompanhar a crescente demanda de novos serviços, fazendo com que as comunicações em serviços de VoIP possam ter uma baixa qualidade de experiência para seus usuários.

De acordo com (SINAM et al., 2014), as aplicações de VoIP estão se tornando populares nos dias de hoje gerando muito tráfego da Internet. Normalmente, o tráfego VoIP é transportado pelo *User Datagram Protocol* (UDP), exceto quando os firewalls bloqueiem o UDP, onde nesse caso, o sinal de voz e o tráfego de sinalização são transportados pelo *Transmission Control Protocol* (TCP).

Existem diferentes protocolos para transporte de dados em uma rede IP, sendo que o mais utilizado é o protocolo TCP, no entanto redes TCP/IP não foram projetadas para aplicações em tempo real, como foram as redes comutadas de telefonia, uma vez que as caraterísticas de retransmissão de pacotes TCP não trariam benefícios a estas aplicações. Porém o protocolo UDP envia os pacotes a uma taxa constante e sem a necessidade de uma confirmação de recibo do pacote no destinatário da rede. As características do protocolo UDP são adequadas para

as aplicações em tempo real, porém este protocolo não consegue oferecer um método de QoS confiável (SHANMUGAN, 2009), (SANCHEZ-IBORRA; CANO; GARCIA-HARO, 2013).

Os atrasos, perda de pacotes e *jitter* (variação do atraso), afetam diretamente a qualidade do sinal transmitido. Essas perdas de pacote podem estar ligadas ao descarte dos pacotes por roteadores congestionados e/ou por problemas no meio físico de transporte.

Para avaliar a qualidade do sinal da voz originada pelos fatores de degradação da rede exitem métodos bem definidos, que são especificados através das recomendações técnicas da *International Telecommunication Union - Telecommunication Standardization Sector* (ITU-T) que disponibiliza recomendações para diversas áreas e finalidades dentro do contexto de telecomunicações. Essas recomendações definem como será avaliada a qualidade da voz, pontuando a qualidade das comunicações, além disso, essas recomendações possuem uma grande importância para o monitoramento da qualidade de experiência (QoE - *Quality of Experience*) dos usuários do serviço.

Os métodos de avaliação da qualidade da voz podem ser classificados em métodos subjetivos ou objetivos. No método subjetivo, um número de pessoas são convidadas a avaliar a qualidade das amostras de voz, essa avaliação resulta no índice MOS (REC, 1996) que é a nota dada à qualidade da voz pela experiência do usuário. No entanto, é um método demorado e de alto custo para ser usado para monitorar continuamente a QoE do sistema em cenário de comunicação de voz em tempo real.

Por outro lado, já os métodos objetivos utilizam algoritmos e tentam predizer aproximadamente a pontuação referente a QoE do usuário que seria dada em testes subjetivos por indivíduos. Os métodos objetivos são subdivididos em dois métodos os intrusivos e não intrusivos, onde o não intrusivo necessita apenas do sinal degradado para determinar a qualidade da voz, enquanto o intrusivo necessita do sinal original como referencia para avaliar a qualidade do sinal degradado.

Segundo a recomendação (MALFAIT; BERGER; KASTNER, 2006), o método objetivo e não intrusivo é o mais recomendado para avaliar comunicações em tempo real como o VoIP. Buscando analisar e melhorar os problemas na comunicação VoIP com os sinais da voz originados de falhas da rede como a perda de pacotes, foi desenvolvido neste trabalho uma implementação de um algoritmo de DL (SCHMIDHUBER, 2015) capaz de avaliar e identificar a qualidade da voz em comunicações VoIP através da taxa de perda de pacote ou pelo seu índice MOS.

Por outro lado, o Machine Learning (ML) e Deep Learning (DL) são técnicas de análise de dados que automatizam o desenvolvimento de modelos analíticos, onde algumas das vantagens de se utilizar DL é a possibilidade de analisar dados complexos em grande escala com um alto grau de aprendizagem. Além de ter um aumento de desempenho, as soluções de ML e DL são capazes de substituir e automatizar diversas tarefas que poderiam ser feitas apenas por humanos, com uma velocidade e taxa de acerto superiores. Permitindo assim uma ágil identificação de problemas e as suas soluções, reduzindo portanto os custos na solução dos problemas.

Nos últimos anos várias pesquisas foram realizadas sobre o uso da voz utilizando a Inteligência Artificial (IA). Porém, percebe-se que grande parte delas não envolvem a qualidade da voz como objetivo principal e sim de usar seus parâmetros e suas características, onde geralmente essas pesquisas utilizam parâmetros como, a segmentação, energia, pitch, cruzamentos por zero, convolução e filtragem utilizando espectrograma para desenvolver novas soluções.

Em (LECUN; RANZATO, 2013) foi desenvolvido um algoritmo capaz de reconhecer a pessoa através de suas características próprias como o tom da voz, pausas e outras podendo identificar a pessoa que está falando, ou capaz de realizar o reconhecimento e tradução de idiomas em tempo real (AMODEI et al., 2016) devido a sua alta capacidade de aprendizado do modelo de rede Deep Learning.

Um outro exemplo de aplicação de DL é a capacidade de detectar doenças como por exemplo, tumores cerebrais (CHETLUR et al., 2014) através da classificação de imagens, também é capaz de identificar vários outros diagnósticos auxiliando no trabalho dos profissionais da área da medicina. Em (WAND; SCHULTZ, 2014) os autores mostram a extração e classificação da fonética da voz de pacientes através das características de dados dos eletromiográficos faciais (facial electromyographic - EMG), utilizando uma Deep Neural Network (DNN) para executar a tarefa de classificação.

De acordo com (EBERLE; PENDERS; YAZICIOGLU, 2011), a doença de Parkinson é uma doença incapacitante neurodegenerativa relativamente comum, que afeta o sistema nervoso com profundo efeito no sistema motor. Onde os sintomas mais comuns incluem lentidão, rigidez e tremor durante o movimento. Porém as cordas vocais estão entre os primeiros sintomas a serem afetados, fazendo com que a voz seja afetada em um estágio inicial da doença que continua a se deteriorar à medida que a doença avança. Diante disso, (EBERLE; PENDERS; YAZICIOGLU, 2011), apresentaram um modelo de rede de Deep Learning capaz de automa-

tizar o processo de diagnóstico da doença através da analise automática da voz, eliminando a necessidade de um pessoal treinado no processo de diagnóstico.

Neste trabalho, duas bases de dados foram construídas com arquivos de áudio degradado gerados com base nos arquivos de áudio originais da base de dados da recomendação ITU-T P.501 C, que contém 28 arquivos de áudio de conversação com uma frequência de amostragem de 16 kHz e da recomendação ITU-T P.862, que contém 20 arquivos de áudio de conversação com uma frequência de amostragem de 8 kHz. Onde ambas foram utilizadas no treinamento dos modelos de classificação desenvolvidos neste trabalho.

No primeiro modelo, para treinamento e validação foi utilizada a base de dados onde os arquivos foram separados em 4 classes com diferentes taxas de PLR. Já para treinar e validar o segundo modelo foi utilizada a base de dados onde os arquivos passaram por uma avaliação do índice MOS através do algoritmo da recomendação ITU-T P.862 e logo em seguida os arquivos de áudio degradado foram separados em 4 classes de acordo com a Tabela Fator-R de índice MOS do E-Model (BERGSTRA; MIDDELBURG, 2003).

Com isso os modelos foram capazes de identificar e classificar uma classe pela sua taxa de PLR e também pelo seu índice MOS para cada arquivo do banco de dados sendo possível predizer a QoE de um usuário final do sinal de voz na rede em tempo real, tendo uma acurácia de 94% e 91%, respectivamente.

A avaliação da solução proposta foi realizada com base na comparação entre os resultados obtidos pelo modelo desenvolvido e pelos os resultados dos algoritmos das recomendações ITU-T P.862 (intrusivo) e ITU-T P.563 (não intrusivo).

1.1 Objetivos

Nessa seção, apresentam-se com maiores detalhes os objetivos gerais e os específicos propostos para o projeto.

1.1.1 Objetivo Geral

O objetivo principal deste projeto foi o desenvolvimento de um sistema não intrusivo de identificação e classificação da qualidade da voz em comunicação VoIP utilizando Deep Learning. Onde através da especificação e treinamento de algoritmos de redes neurais, foram implementados dois modelos capazes de classificar o grau de degradação da voz através do índice MOS e através da análise da taxa de perda de pacotes do sinal da voz.

1.1.2 Objetivos Específicos

- Implementar um cenário de rede VoIP afetado pelo fator de degradação PLR.
- Desenvolver um algoritmo de simulação de taxa de perda de pacotes em arquivos de áudio, criando as bases de dados de arquivos de áudio degradados.
- Utilizar diferentes técnicas de DL para criar os modelos propostos para o trabalho.
- Avaliar o desempenho do modelo proposto em relação aos métodos de avaliação da qualidade da voz, como os resultados do algoritmo não intrusivo da recomendação ITU-T P.563 e dos resultados do algoritmo intrusivo da recomendação ITU-T P.862.

1.2 Estrutura da Dissertação

A presente dissertação está dividida em 5 capítulos organizados da seguinte forma.

No capítulo 2 apresenta-se um estudo do estado da arte em VoIP.

No capítulo 3 são apresentados os métodos e fundamentos teóricos da avaliação da qualidade da voz e alguns dos principais algoritmos utilizados atualmente, um estudo do estado da arte em Redes Neurais Artificiais e suas características principais para uso na dissertação e um estudo do estado da arte em Deep Learning e redes neurais convolutivas.

No capítulo 4 apresenta-se a metodologia utilizada na dissertação como as ferramentas utilizadas, definição da base de dados, definição do modelo utilizado e os testes realizados.

No capítulo 5 são apresentados os resultados para cada uma das bases de dados definidas no projeto e a comparação com os resultados dos algoritmos das recomendações ITU-T P.563 e ITU-T P.862.

Para finalizar, no capítulo 6 apresentam-se as conclusões extraídas no desenvolvimento desta pesquisa e as propostas de trabalhos futuros.

2 REFERENCIAL TEÓRICO

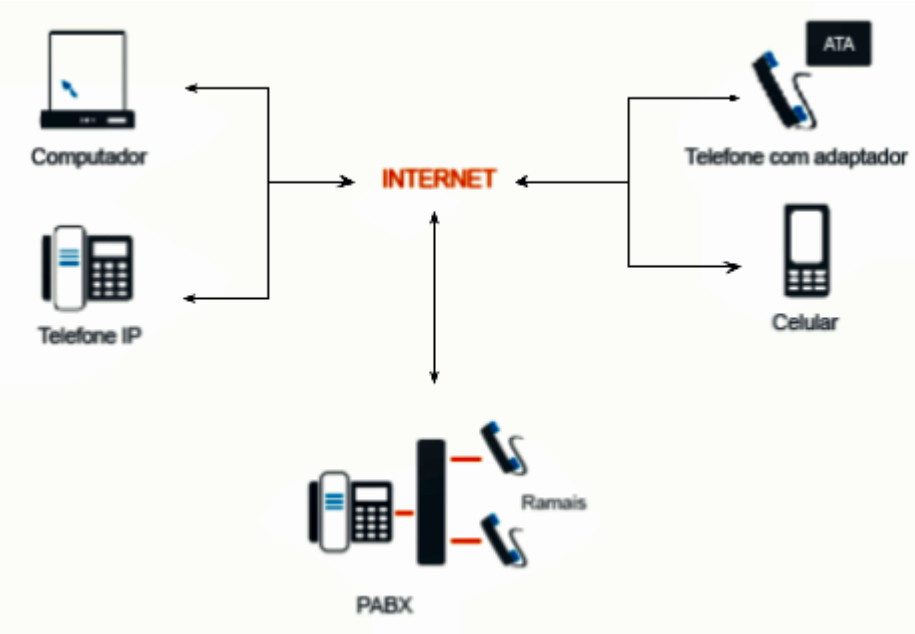
2.1 Voz sobre IP

De acordo com (CUNY; LAKANIEMI, 2003) durante os últimos anos, os serviços de voz sobre redes de dados ganharam popularidade crescente. Com o rápido crescimento das redes baseadas no Protocolo de Internet (IP), a transmissão do VoIP despertou muito interesse substituindo alguns casos a tradicional tecnologia de telefonia de longa distância devido aos custos reduzidos para o usuário final.

Embora a tecnologia VoIP envolva a transmissão de voz digitalizada em pacotes, o próprio telefone pode ser analógico ou digital. Pois a voz pode ser digitalizada e codificada antes ou ao mesmo momento que o empacotamento, onde por muitos anos, o PSTN operou estritamente com o padrão ITU-T G.711 (GOODE, 2002).

VoIP é uma implementação de telefonia sobre redes IP. Porém, este tipo de aplicação é sensível a vários parâmetros de qualidade em nível de rede, que influenciam na qualidade da voz percebida pelos seus usuários.

Figura 2.1 – VoIP estrutura básica.



Fonte: Autor.

A internet não foi otimizada para o tráfego em tempo real, uma vez que foi originalmente desenvolvida para transportar grandes dados em rajadas. No entanto, a comunicação VoIP é vulnerável ao mal desempenho da rede devido a algumas deficiências, onde acordo com

as expectativas dos utilizadores e as infraestruturas disponíveis, diferentes parâmetros como latência, jitter, perda de pacotes, largura de banda, tráfego de rede, clima, entre outros, afetam a QoS e deterioram a voz (ORTEGA; ALTAMIRANO; ABAD, 2018), além da escolha do Codec que será utilizado (GOODE, 2002).

Atualmente, os serviços de VoIP estão enfrentando diferentes mudanças para melhorar a qualidade das comunicações. A análise de QoS da rede VoIP é um método comum que avalia a qualidade do serviço VoIP (BEHDADFAR; FAGHIHI; SADEGHI, 2015). Grandes desafios surgem quando se tenta otimizar as deficiências da rede, em um esforço para melhorar a qualidade do serviço VoIP de maneira eficiente durante a comunicação em tempo real (TRAORE et al., 2018).

De acordo com (MANOUSOS et al., 2005) o atraso no VoIP ocorre devido ao processamento de voz nos estágios de codificação, decodificação e *look-ahead*, que é o buffer de amostras de voz que certos codecs empregam para melhorar a taxa de compactação e a qualidade de voz compactada, os atrasos ocorrem também na etapa de empacotamento de dados de voz, transmissão de pacotes e atrasos de carregamentos de quadros de voz que chegam ao receptor para eliminar o jitter da rede, pois as aplicações VoIP não toleram o jitter, que pode ser eliminado com buffers de reprodução estáticos ou adaptativos.

Os pontos finais podem controlar todos esses fatores, exceto o atraso de transmissão, para minimizar seus efeitos na qualidade da conversação. Consequentemente, a QoS de ponta a ponta depende fortemente das estratégias e mecanismos de controle implementados nos pontos finais (MANOUSOS et al., 2005).

A QoS garante o desempenho na transmissão das informações durante a comunicação, desde que existam as seguintes condições: infraestrutura de rede adequada, otimização dos recursos tecnológicos, aplicação de técnicas de tráfego inteligente, controle de banda, uso de codecs apropriados, entre outras. A QoS é afetada por parâmetros como latência, jitter e perda de pacotes enumerados abaixo com seus respectivos valores limites de acordo com o trabalho desenvolvido por (ORTEGA; ALTAMIRANO; ABAD, 2018).

- Latência: É o atraso que os pacotes experimentam de ponta a ponta. A recomendação ITU-T G.114 define um valor limite de 150 ms.

- Jitter: É o tempo de atraso entre os pacotes de ponta a ponta. É dado devido ao tráfego de rede, roteamento e sincronização. O valor recomendável é menor que 100 ms de acordo com as recomendações ITU-T G.114 e ITU-T Y.1541.

- Perda de pacotes: É produzida por causa do tipo de rede em que as informações são transportadas. Para evitar a degradação da voz, é recomendável um valor de perda inferior a 1%.

Segundo (ORTEGA; ALTAMIRANO; ABAD, 2018) uma maneira de medir a qualidade da voz é através do índice MOS, que é definido pela recomendação ITU-T P.800. Onde este sistema quantifica a qualidade geral da voz por meio de uma escala numérica entre 1 e 5 conforme a Tabela 2.1.

Tabela 2.1 – Escala de índice MOS da recomendação ITU-T P.800

MOS	**Qualidade**	**Níveis de degradação**
5	Excelente	Imperceptível
4	Boa	Perceptível, mas não é incômoda
3	Razoável	Levemente incômoda
2	Ruim	Incômoda
1	Péssima	Muito incômoda

Fonte: Adaptado da Recomendação ITU-T P.800

2.2 Codificação de voz

De acordo com (SANTOS et al., 2014) e (MATIAS, 2010), a voz é um conjunto de vibrações acústicas representada naturalmente na forma analógica, onde o sinal analógico produzido por essas vibrações é transformado em um sinal digital antes de ser transmitido na rede. Quando chega no receptor, o sinal digital de voz é convertido de volta para a sua forma analógica e transformado pelo ouvido humano em percepções ao cérebro, que identifica um padrão e monta uma mensagem.

2.2.1 Algoritmos de codificação de voz - Codecs

As características do tráfego de voz gerado são dependentes principalmente do codificador utilizado. Normalmente o arquivo de áudio é comprimido para a eficiência da utilização da largura de banda. Toda chamada VoIP exige a utilização de dois algoritmos onde um é utilizado para comprimir os dados da origem (codificação) e outro para descomprimir no destino (decodificação). Na literatura, esse conjunto de algoritmos são conhecidos como codecs (CAVALCANTE, 2018)

Segundo (BEHDADFAR; FAGHIHI; SADEGHI, 2015), os codecs são responsáveis pela tarefa de codificação e decodificação do fluxo de dados digitais, onde dependendo da

qualidade de voz desejada e das possíveis taxas de bits, diferentes tipos de codec podem ser aplicados, como codec de forma de onda, codec paramétrico e codec híbrido. Em que quando utilizado codecs de forma de onda resulta em alta qualidade de fala enquanto codecs paramétrico torna a voz sintética. Entretanto, sistemas que usam codecs híbridos, a qualidade da fala é aceitável e a taxa de bits é moderada.

Segundo (SANTOS et al., 2014), um algoritmo de codificação de voz é normalmente avaliado em quatro requisitos:

- Eficiência: normalmente representada pela taxa de bits necessária para a transmissão da voz;
- Complexidade: medida em milhões de instruções por segundo (MIPS) necessárias na codificação do sinal;
- Atraso: corresponde ao tempo em milisegundos que é necessário armazenar a voz para que o algoritmo de compressão seja aplicado;
- Qualidade: possui o objetivo de alcançar a melhor qualidade de voz com a menor taxa de transmissão de bits possível.

De acordo com (MANOUSOS et al., 2005) a codificação G.729A é a mais utilizada em aplicativos VoIP e em todos os fluxos de voz durante a experimentação.

2.2.1.1 Codecs de Forma de onda

(CARVALHO; DANILO, 2000) define que codificadores de forma de onda, ou de linha, são esquemas que tentam aproximar o sinal gerado ao sinal de voz original, onde a forma básica de codificação aplicada no sinal de voz é a digitalização, uma vez que o sinal obtido é analógico ou contínuo no tempo.

Segundo (SANTOS et al., 2014), os codificadores de onda fornecem um sinal codificado o mais próximo possível do sinal analógico original, com base nas suas características estatísticas, temporais ou espectrais. Onde normalmente são codificadores de baixa complexidade e que introduzem um pequeno retardo na voz.

Algumas técnicas adicionais podem ser utilizadas para diminuir a taxa de bits nos codificadores de forma de onda. Entre elas convêm destacar o PCM Adaptativo (APCM), o PCM

Diferencial (DPCM) e o PCM Adaptativo e Diferencial (ADPCM) (CARVALHO; DANILO, 2000) e (SANTOS et al., 2014).

- PCM (Pulse Code Modulation) – É um método usado para representar digitalmente amostras de sinais voz analógicos baseado na forma de onda. Possui uma taxa de amostragem de 8000 amostras/segundo e cada amostra é codificada por uma sequência de 8 bits resultando em uma frequência de amostragem de 64 kbit/s (CARVALHO; DANILO, 2000);

- APCM (Adaptive Pulse Code Modulation) – Nesta modulação utiliza-se um passo de quantização que varia com o tempo para acompanhar as variações de amplitude do sinal de voz, baseando-se nas amostras passadas do mesmo. Desta forma, reduz-se a faixa dinâmica do sinal e consequentemente a taxa final de transmissão (CARVALHO; DANILO, 2000);

- DPCM (Differential Pulse Code Modulation) - Nesta modulação quantiza-se a diferença de amplitude entre amostras adjacentes. Como essa diferença é relativamente pequena, pode ser representada com menos bits. O sinal de entrada no quantizador é a diferença entre o sinal original e uma predição do mesmo, baseada nas amostras passadas, resultando em um sinal chamado de erro de predição, que por sua vez é codificado a uma taxa de 32 kbit/s (CARVALHO; DANILO, 2000).

- ADPCM (Adaptive Differential Pulse Code Modulation) - Nesta modulação empregam-se a quantização e/ou a predição adaptativa. A predição adaptativa consiste no ajuste dinâmico do preditor de acordo com variações no sinal de voz. O quantizador adaptativo realiza um ajuste do valor da diferença observada entre amostras consecutivas. Codificadores ADPCM apresentam boa qualidade de voz para taxas entre 24 e 48 kbit/s (CARVALHO; DANILO, 2000).

2.2.1.2 Codecs paramétricos

Os codificadores paramétricos, ou vocoders, utilizam um modelo de como o sinal foi gerado, e extraem os parâmetros que representam este modelo (CARVALHO; DANILO, 2000). São esses parâmetros que são enviados ao decodificador. Os codificadores paramétricos são baseados nas estatísticas do sinal de voz, não funcionam para outro tipo de sinais (CARVALHO; DANILO, 2000). Além disso, a qualidade é inferior à qualidade dos padrões telefônicos, e a voz

reproduzida tem um aspecto sintético ou não. Os vocoders operam com baixas taxas de transmissão, normalmente inferiores a 4 kbit/s e possuem valores de atrasos grandes e complexidade elevada (SANTOS et al., 2014).

A qualidade da voz para os codificadores paramétricos é baixa, soando de forma sintética. Seu principal uso são aplicações militares, na qual a fidelidade da voz não é tão importante quanto à obtenção de uma baixa taxa de transmissão, permitindo uma criptografia forte com pouca necessidade de largura de banda (SANTOS et al., 2014).

Um tipo de vocoder muito utilizado é a Codificação por Predição Linear (LPC - Linear Predictive Coding) que extrai os parâmetros para o modelo do trato vocal diretamente da forma de onda no tempo, obtendo um resultado melhor que outros tipos de vocoders que obtêm seus parâmetros a partir do espectro de frequência (CARVALHO; DANILO, 2000).

2.2.1.3 Codecs de híbridos

Com taxas de bits entre 4 e 16 kbit/s e mais complexos, os codificadores híbridos exploram técnicas tanto dos codificadores paramétricos como dos de forma de onda, e com isso conseguem uma qualidade muito próxima à dos codificadores de linha, que requerem taxas acima de 16 kbit/s (SANTOS et al., 2014) e (CARVALHO; DANILO, 2000).

Os codificadores híbridos também são baseados nos modelos de produção da voz e utilizam uma excitação mais apurada para o sintetizador, que propicia uma melhora na qualidade da voz sintetizada, tornando-a mais inteligível que nos vocoders convencionais (CARVALHO; DANILO, 2000).

Os codificadores híbridos também utilizam LPC, porém foram desenvolvidas técnicas de codificação do resíduo para que mais informações sobre este também fossem transmitidas, aumentando a qualidade do sinal. Desta forma, tornam-se mais inteligíveis que os vocoders convencionais e com qualidade próxima à obtida com os codificadores de forma de onda (SANTOS et al., 2014).

Alguns exemplos de codificadores híbridos:

- CELP (*Code Excited Linear Prediction*) é uma técnica que foi desenvolvida para minimizar o aumento da taxa de transmissão de bits para que possa ser enviado mais informações sobre o resíduo. No CELP, uma tabela com os valores mais comuns de resíduos é utilizada pelo codificador para comparar o valor do resíduo com os valores da tabela e, ao encontrar o melhor valor, o codificador envia apenas o código que representa o valor da

tabela. O receptor busca na sua tabela de resíduos os valores correspondentes ao código recebido, e então usa esse valor para excitar o filtro das frequências formadoras. Para conter todos os valores de resíduos é necessária uma tabela grande o suficiente para conter todos os valores, aumentando muito o tempo de procura pelo valor correto. Na prática são utilizadas duas tabelas. A tabela fixa possui valores fixos de resíduos que são determinados durante a construção do sistema. A tabela adaptativa é preenchida durante a operação do sistema com cópias atrasadas do resíduo usado anteriormente, onde o atraso representa a mudança de frequência (SANTOS et al., 2014) (HAN et al., 2015).

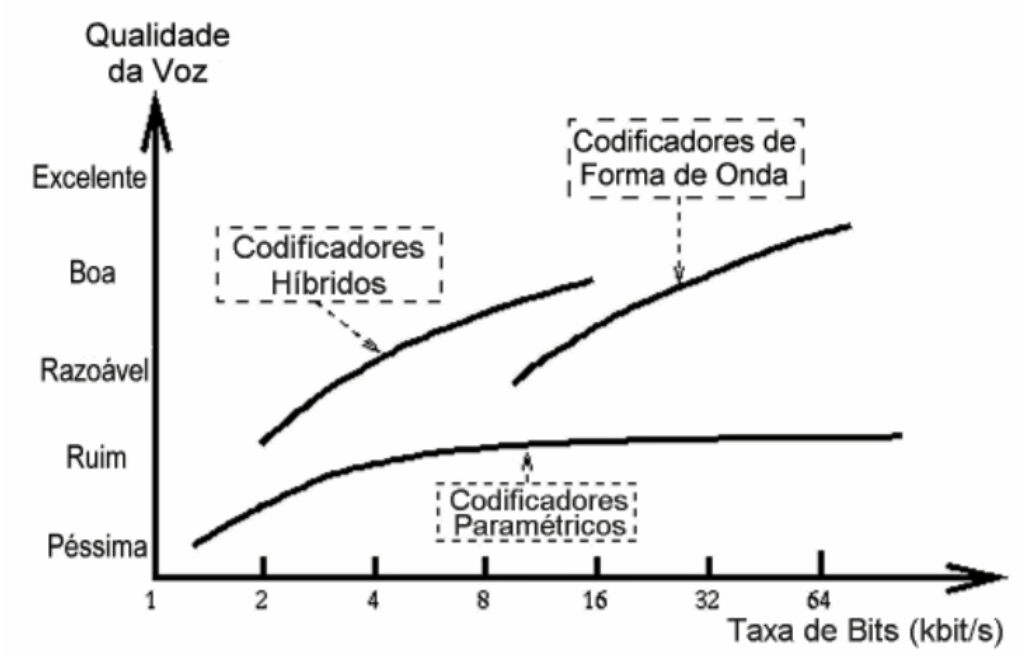
- ACELP (*Algebraic Code Excited Linear Prediction*) - A diferença básica entre o ACELP e o CELP está na tabela fixa de valores de resíduos. No ACELP, são utilizados códigos algébricos promovendo um aumento na qualidade das componentes harmônicas (SANTOS et al., 2014).

- CS-ACELP (*Conjugate Structure Algebraic Code Excited Linear Prediction*) utiliza um algoritmo com uma forma diferente de armazenamento que faz uma pré-seleção dos futuros candidatos a serem escolhidos para representar a próxima janela a ser analisada, otimizando ainda mais o processo de procura nas tabelas de resíduos (SANTOS et al., 2014).

- LD-CELP (*Low Delay Code Excited Linear Prediction*) é um codificador que possui uma taxa de codificação de 16 kbit/s e que trabalha com blocos de cinco amostras PCM, onde para cada amostra ocorre um atraso de 0,125 ms, fazendo com que o atraso desse algoritmo seja de 0,625 ms (RODRÍGUEZ, 2009).

A qualidade da voz, em função da taxa de bits e do tipo de codificador, de acordo com (CARVALHO; DANILO, 2000), é apresentada no gráfico da Figura 2.2.

Figura 2.2 – Comparação entre classes de codificadores de voz.



Fonte: Adaptado de (CARVALHO; DANILO, 2000).

2.2.1.4 Codec G.711

O codec de áudio G.711 é um dos padrões utilizados pelo codificador H.323 e utiliza PCM (Pulse Code Modulation) para representar amostras de frequências de voz, amostrada a 8 kHz.

Aprovado em 1972, possuía uma taxa de codificação de 64 kbit/s contendo 8 bits por amostra e utilizava a codificação por PCM que obedecia ao critério de Nyquist, em que a frequência de amostragem é igual ou superior ao dobro da maior frequência presente no espectro (SANTOS et al., 2014). Assim, o sinal de voz com a codificação PCM possui uma taxa padrão 64 kbit/s resultante de 8000 amostras/seg x 8 bits/amostra. Segundo (SANTOS et al., 2014) o atraso do algoritmo é de apenas 0,125 ms.

De acordo com (BUENO, 2008), essa recomendação prevê o uso de dois tratamentos de erros de quantização, denominadas por -law, usado nos Estados Unidos e Japão, e o A-law, usado em outros países fora dos EUA. Ambos são logarítmicos, mas o A-law é mais simples para ser processado.

2.2.1.5 Codec G.726

De acordo com (BUENO, 2008) o G.726 é um codec ADPCM (*Adaptive Pulse Code Modulation*) aprovado em 1990 com transmissão de 16, 24, 32 e 40 kbit/s, que possui uma qualidade praticamente idêntica à do codec G.711, porém com a metade do consumo de largura de banda.

De acordo com (SANTOS et al., 2014), a Recomendação G.726 possui um quantizador adaptativo, podendo fazer um ajuste no preditor linear com base nas variações do sinal a ser codificado, operando com taxa de amostragem de 8 kHz (SANTOS et al., 2014).

2.2.1.6 Codec G.728

O G.728 é um padrão da ITU-T para compressão e descompressão de áudio. O G.728 é um padrão de codificação de voz, que funciona em 16 Kbps na taxa de amostragem de 8.000 amostras/segundo. Ele é usado em sistemas de transmissão digital onde possui o objetivo de codificar os sinais analógicos em sinais digitais. Segundo (BUENO, 2008), este codec usa LD-CELP (Low Delay Code Excited Linear Predicition). O codec possui um atraso de apenas 5 amostras (0.625ms).

2.2.1.7 Codec G.729

O codec de áudio G.729 é um algoritmo que comprime voz em pacotes de 10 ms de duração e foi aprovado em 1996. Tem uma baixa largura de banda, de 8 kbit/s, usando CS-ACELP (BUENO, 2008).

Segundo (SANTOS et al., 2014), a recomendação G.729 foi concebida para codificar um sinal de voz com qualidade total em 8 kbit/s, e ser usado por aplicações com comunicação sem fio e por redes com fio que necessitem de compressão da banda usada pelo sinal codificado, como por exemplo circuitos transoceânicos. O áudio é codificado em quadros de 10 ms, correspondentes a 80 amostras das 8000 do PCM. Adicionalmente, possui um tempo de look-ahead de 5 ms, resultando em 15 ms o tempo de atraso do algoritmo.

G.729 - Anexo A apresentado em maio de 1996, tinha como objetivo reduzir a complexidade e manter a interoperabilidade com o codec G.729 original. O funcionamento básico do algoritmo no codec G.729 Anexo A é o mesmo do codec G.729, porém suas principais simplificações foram feitas com relação à operação dos filtros e forma de busca nos dicionários de vetores (SANTOS et al., 2014).

G.729 - Anexo B foi aprovado em outubro de 1996 e descreve o detector de voz ativa e gerador de ruído de conforto, ambos usados na compressão de silêncio, tanto na G.729 como na G.729 - Anexo A (SANTOS et al., 2014).

2.2.1.8 Codec G.723.1

Segundo (BUENO, 2008), o codec é baseado em ADPCM (*Adaptative Differential Pulse Code*) com 24 e 40 kbit/s, foi aprovado em 1996 e possui as seguintes taxas de codificação 5,3 e 6,3 kbit/s com os seguintes tipos de codificação ACELP (Algebraic-Code-Excited Linear-Prediction) para 5,3 kbit/s e MP-MLQ (*Multipulse Maximum Likelihood Quantization*) para 6,3 kbit/s.

Esta recomendação segundo (SANTOS et al., 2014) especifica uma codificação usada para compressão de voz de um serviço multimídia para meios de muito baixa velocidade de transmissão, onde são necessários 30 ms para a formação de um quadro, independente da velocidade em uso, além de 7,5 ms de look-ahead, resultando em 37,5 ms o tempo de atraso do algoritmo.

A recomendação ITU-T G.273.1 define que as duas velocidades de transmissão devem estar disponíveis, podendo trocar de velocidade de um quadro para outro ou nos períodos de descontinuidade de transmissão nos intervalos sem voz. A diferença entre essas taxas resulta do tipo de excitação a ser utilizada e transmitida para o decodificador, ACELP para 5,3 kbit/s ou MP-MLQ para 6,3 kbit/s (SANTOS et al., 2014).

2.2.1.9 Codec iLBC

Aprovado em 2004 o codificador iLBC utiliza a codificação preditiva linear, suporta quadros de 20 ms em uma taxa de 15,2 kbit/s e quadros de 30 ms em uma taxa de 13,33 kbit/s (RODRÍGUEZ, 2009).

Segundo (WISNEVSKI; FAGUNDES; COSSIO, 2010), o iLBC (*Internet Low Bitrate Codec*) é um codificador de voz, de banda estreita, amostrado a uma frequência de 8 kHz, ele reduz a taxa de bits devido as suas técnicas de codificação, alcançando um bom percentual de qualidade, explorando as técnicas de mascaramento das distorções, considerando as características do ouvido humano. Por ser um codec aplicado a uma rede de pacotes, os impactos de qualidade da voz inerente ao meio de transmissão, estão diretamente ligados à perda de pacotes,

ao delay e ao jitter. Estes parâmetros são importantes e devem ser considerados para manter a qualidade da ligação entre os interlocutores.

A essência do codec iLBC é a Codificação Linear Preditiva, onde o objetivo do codificador é enviar ao decodificador os coeficientes do filtro LPC, e o resíduo resultante da aplicação deste ao sinal original de voz.

Segundo (WISNEVSKI; FAGUNDES; COSSIO, 2010), a codificação entre os blocos é realizada de forma independente, resultando na eliminação de propagação de degradações perceptual devido à perda de pacotes. O método facilita a ocultação de perda de pacotes de alta qualidade (PLC - *Packet Loss Concealment*) .

2.2.2 Fatores de degradação

Os fatores de degradação são responsáveis pela má experiência na comunicação de voz entre um emissor e o receptor. Geralmente nas redes cabeadas a degradação da qualidade de uma transmissão está sempre ligada a fatores físicos, como por exemplo meio de transporte (tipo de cabos), conectores com defeito e equipamentos sobrecarregados ou danificados, esses problemas podem gerar perda de pacotes e atrasos.

2.2.2.1 A perda de pacotes

A perda de pacotes (PLR - Packet Loss Rate) está ligada ao descarte dos pacotes por roteadores e/ou switches congestionados e também por problemas no meio físico de transporte. Portanto, em transmissões em tempo real como VoIP, a perda de pacotes pode afetar o sinal recebido fazendo com que não seja igual ao enviado, porém nem sempre pode ocorrer uma retransmissão dos dados perdidos, especialmente em transmissões em tempo real.

Nas redes wireless, diferente das redes cabeadas, são encontrados outros fatores degradantes, que prejudicam a qualidade do sinal de voz transmitida, como o desvanecimento ou fadding, (LEE et al., 2013), (GARZÓN, 2014), (CHOUDHURY; GIBSON, 2007), (YUHE; JIE, 2009), que pode ocorrer por diversos fatores, como a dificuldade de contornar objetos no caminho do sinal, os múltiplos percursos formados por refletores das ondas, o afastamento, dentre outros.

2.2.2.2 Atrasos

O tempo de transmissão de um dado inclui o atraso devido ao processamento do codec, bem como o atraso da propagação. A Recomendação ITU-T G.114 (ABBAS; MOSBAH; ZEMMARI, 1996) recomenda os seguintes limites de tempo de transmissão unidirecional para conexões com eco adequadamente controlado da Recomendação G.131 (REC, 2003):

- 0 a 150 ms: aceitável para a maioria das aplicações de usuários;
- 150 a 400 ms: aceitável para conexões internacionais;
- 400 ms: inaceitável para fins gerais de planejamento de rede.

O Anexo B da Recomendação ITU-T G.114 descreve os resultados de testes subjetivos para avaliar os efeitos do atraso puro na qualidade da voz. Um teste concluído em 1989 mostrou que a porcentagem de usuários classificando a chamada como ruim ou inferior (POW - *Poor or Worse*) para a qualidade geral começou a aumentar acima de 10% apenas para atrasos superiores a 500 ms, mas o POW para interrupção foi superior a 10% para atrasos de 400 ms. Um dos testes, concluído em 1990, foi projetado para obter reações subjetivas, em contexto de interrupção e qualidade, para circuitos telefônicos eco-livres em que foram introduzidas várias quantidades de atraso. Os resultados indicaram que longos atrasos não reduziram significativamente a qualidade de experiência média no intervalo de atraso testado. No entanto, as observações durante o teste e as entrevistas de sujeito após o teste mostraram que os sujeitos experimentaram algumas dificuldades reais de comunicação nos atrasos mais longos, embora os sujeitos nem sempre associam-se à dificuldade ao atraso (ABBAS; MOSBAH; ZEMMARI, 1996).

Em 1991 foi realizado a medição do efeito do atraso usando seis tarefas diferentes envolvendo mais ou menos interrupções no diálogo. O limite de detecção de atraso foi definido como o atraso detectado em 50% dos assuntos de uma tarefa. A medida que a interatividade exigida pelas tarefas diminuiu, o limite de detecção de atraso aumentou de 45 para 370 ms de atraso unidirecional. A medida que o atraso unidirecional aumentou de 100 a 350 ms, a qualidade do índice MOS diminuiu de 3.74 para 3.48, e a aceitabilidade da conexão diminuiu de 80% para 73% (ABBAS; MOSBAH; ZEMMARI, 1996).

A variação de atraso, as vezes chamada de jitter, também é importante. O gateway ou o telefone de recepção deve compensar a variação de atraso com um buffer de jitter, o que impõe um atraso nos pacotes iniciais e passa os pacotes atrasados com menos atraso, de modo que a

voz decodificada se transmite para fora do receptor a uma taxa constante. Todos os pacotes que chegam depois do comprimento do buffer de jitter são descartados, uma vez que queremos uma baixa perda de pacotes, o atraso do buffer de jitter é a variação de atraso máximo que esperamos.

Este atraso do buffer de jitter deve ser incluído no atraso de ponta a ponta total que o ouvinte experimenta durante uma conversa usando a telefonia de pacotes.

A voz em bloco tem grandes atrasos de ponta a ponta do que um sistema TDM (*Time-Division Multiplexing*), tornando desafiantes os objetivos de atraso acima. Este orçamento não é preciso. O atraso do buffer de jitter alocado de 60 ms é apenas uma estimativa; O atraso real pode ser maior ou menor. Uma vez que o orçamento da amostra não inclui atrasos específicos para a compactação e descompressão do cabeçalho, podemos considerar que, se essas funções são empregadas, o atraso de processamento associado é agrupado no atraso do link de acesso.

Os atrasos da rede na região Ásia-Pacífico, bem como entre a América do Norte e a Ásia, podem ser superiores a 100 ms. De acordo com a recomendação ITU-T G.114, esses atrasos são aceitáveis para os links internacionais. No entanto, os atrasos de ponta a ponta para chamadas VoIP são consideravelmente maiores do que para chamadas PSTN (*Public Switched Telephone Network*).

2.2.3 Empacotamento e Transmissão

Informações em chamadas VoIP são geralmente transmitidas usando UDP como protocolo de transporte e o protocolo IP é utilizado para encaminhamento de dados na rede. Além disso o UDP é um protocolo no qual os pacotes podem ser entregues fora de ordem ou até perdidos e, ao contrário do TCP, não possui um controle de fluxo com reenvio de pacotes perdidos e nem controles de congestionamento da rede (SANTOS et al., 2014).

Entretanto, de acordo com (SANTOS et al., 2014), o serviço de entrega de pacotes disponibilizado pelo UDP não é suficiente para garantir uma qualidade aceitável no serviço de chamada VoIP, porém é totalmente eficiente para aplicações em tempo real, onde utilizam o RTP (*Real-Time Transport Protocol*) e o RTCP (*Real Time Control Protocol*). O RTP possui o objetivo de transportar a mídia e oferecer serviços essenciais para aplicações de tempo real. Porém o RTP não oferece nenhuma reserva de recurso na banda, nem recuperação dos pacotes e não garante QoS. Enquanto o RTCP tem como objetivo o monitoramento da QoS obtendo informações da rede como quantidade de jitter, perda média de pacotes, atrasos e outras.

2.3 Processo de reconhecimento da qualidade do sinal de voz

O processo de reconhecimento da qualidade do sinal de voz é realizado através do processamento de arquivos de áudio, onde é identificado as características da voz contidos nos sinais de áudio e por fim realizando uma classificação de acordo com os dados obtidos.

Mais há um problema no reconhecimento que pode se dar de duas abordagens diferentes, no reconhecimento ou na verificação. Em ambas as abordagens o reconhecimento da qualidade da voz está sendo limitado pelo universo de características da voz para serem analisados e reconhecidos, ou seja, quanto maior o universo de padrões diferentes definidos, maior será a complexidade do sistema e com isso menor é a acurácia do sistema (IRIYA, 2014). Algumas características estão definidas na Tabela 2.2, onde representa as características extraídas pela biblioteca Librosa que utiliza a biblioteca scikit-learn em python (PEDREGOSA et al., 2011) para extração das características da voz.

Tabela 2.2 – Características extraídas do áudio pela biblioteca Librosa

Característica	**Descrição**
Taxa Zero Crossing	A taxa de ano faz sinal durante a duração de um quadro em específico.
Energia	A normalidade do tamanho dos desenhos é normalizada pelo tamanho do quadro.
Entropia da Energia	A entropia das energias normalizadas dos subquadros. Pode ser interpretado como uma medida de alterações rápidas.
Centróide Espectral	O centro de gravidade do espectro.
Espalhamento do Espectro	O segundo momento central do espectro.
Entropia Espectral	Entropia das energias espectrais normalizadas para um conjunto de sub-quadros.
Fluxo Espectral	A diferença quadrática entre as grandezas normalizadas do espectro de dois quadros sucessivos.
Deslocamento espectral	Uma frequência abaixo da qual se concentra 90% da distribuição de magnitude do espectro.
MFCC	Os coeficientes de páscoa de discussão podem formar uma representação cepstral, como as bandas de saída não são lineares, mas distribuídas de acordo com a escala Mel.
Vetor de Chroma	Uma apresentação da energia com 12 elementos em que os esquemas representam 12 classes de passos com constante de músicas ocidentais (espaçamento de semitom).
Desvio de Chroma	O desvio padrão dos 12 coeficientes de chroma.

Fonte: Autor.

Em um estudo conduzido por (IRIYA, 2014), foi selecionado para o estudo um conjunto de 96 características de curto prazo, incluindo pitch, energia, as 5 primeiras formantes, os 13 primeiros MFCC e os 12 primeiros LFPC, além das primeiras e segundas derivadas de cada característica, conforme a Tabela 2.3.

Tabela 2.3 – Índices das característica extraídos

Indice	**Parâmetro**	**Indice**	**Parâmetro**	**Indice**	**Parâmetro**
1	Pitch	33	3º LFPC - 2ª Derivada	65	8º MFCC
2	Pitch - 1ª Derivada	34	4º LFPC - 2ª Derivada	66	9º MFCC
3	Pitch - 2ª Derivada	35	5º LFPC - 2ª Derivada	67	10º MFCC
4	LogEnergy	36	6º LFPC - 2ª Derivada	68	11º MFCC
5	LogEnergy - 1ª Derivada	37	7º LFPC - 2ª Derivada	69	12º MFCC
6	LogEnergy - 2ª Derivada	38	8º LFPC - 2ª Derivada	70	13º MFCC
7	1º LFPC	39	9º LFPC - 2ª Derivada	71	1º MFCC - 1ª Derivada
8	1º LFPC	40	10º LFPC - 2ª Derivada	72	2º MFCC - 1ª Derivada
9	3º LFPC	41	11º LFPC - 2ª Derivada	73	3º MFCC - 1ª Derivada
10	4º LFPC	42	12º LFPC - 2ª Derivada	74	4º MFCC - 1ª Derivada
11	5º LFPC	43	1º Formante	75	5º MFCC - 1ª Derivada
12	6º LFPC	44	2º Formante	76	6º MFCC - 1ª Derivada
13	7º LFPC	45	3º Formante	77	7º MFCC - 1ª Derivada
14	8º LFPC	46	4º Formante	78	8º MFCC - 1ª Derivada
15	9º LFPC	47	5º Formante	79	9º MFCC - 1ª Derivada
16	10º LFPC	48	1º Formante - 1ª Derivada	80	10º MFCC - 1ª Derivada
17	11º LFPC	49	2º Formante - 1ª Derivada	81	11º MFCC - 1ª Derivada
18	12º LFPC	50	3º Formante - 1ª Derivada	82	12º MFCC - 1ª Derivada
19	1º LFPC - 1ª Derivada	51	4º Formante - 1ª Derivada	83	13º MFCC - 1ª Derivada
20	2º LFPC - 1ª Derivada	52	5º Formante - 1ª Derivada	84	1º MFCC - 2ª Derivada
21	3º LFPC - 1ª Derivada	53	1º Formante - 2ª Derivada	85	2º MFCC - 2ª Derivada
22	4º LFPC - 1ª Derivada	54	2º Formante - 2ª Derivada	86	3º MFCC - 2ª Derivada
23	5º LFPC - 1ª Derivada	55	3º Formante - 2ª Derivada	87	4º MFCC - 2ª Derivada
24	6º LFPC - 1ª Derivada	56	4º Formante - 2ª Derivada	88	5º MFCC - 2ª Derivada
25	7º LFPC - 1ª Derivada	57	5º Formante - 2ª Derivada	89	6º MFCC - 2ª Derivada
26	8º LFPC - 1ª Derivada	58	1º MFCC	90	7º MFCC - 2ª Derivada
27	9º LFPC - 1ª Derivada	59	2º MFCC	91	8º MFCC - 2ª Derivada
28	10º LFPC - 1ª Derivada	60	3º MFCC	92	9º MFCC - 2ª Derivada
29	11º LFPC - 1ª Derivada	61	4º MFCC	93	10º MFCC - 2ª Derivada
30	12º LFPC - 1ª Derivada	62	5º MFCC	94	11º MFCC - 2ª Derivada
31	1º LFPC - 2ª Derivada	63	6º MFCC	95	12º MFCC - 2ª Derivada
32	2º LFPC - 2ª Derivada	64	7º MFCC	96	13º MFCC - 2ª Derivada

Fonte: Adaptado de (IRIYA, 2014).

O processo de reconhecimento da qualidade do sinal de voz se compreende nos seguintes passos: entrada de áudio, pré processamento, extração de características, classificação e saída, que serão detalhados nas próximas sessões do trabalho.

O sistema funciona em duas etapas distintas: o treinamento e a validação. Tanto no treinamento quanto na validação, uma amostra de sinal de áudio passa por um pré-processamento que fragmenta o sinal para capturar as características da voz. Destes fragmentos, são extraí-

das as características do sinal de voz para o processo e estes são usados para treinar o modelo de classificação. Um processo muito semelhante é realizado na fase de validação, porém após as características do sinal de voz serem extraídas, o conjunto de características é confrontado com cada um dos modelos existentes e será classificado pelo modelo que apresenta a melhor coerência.

2.3.1 Características de Voz

De acordo com (RIBEIRO et al., 2014) e (JR et al., 2017) as principais características utilizadas para o reconhecimento da voz são a frequência fundamental conhecida como pitch, a energia e as propriedades relacionadas à duração da voz e pausas, enquanto que entre as características que descrevem a qualidade da voz encontram-se as frequências formantes, a distribuição espectral, representada por MFCC's (Mel Frequency Cepstral Coefficients) ou LFPC's (Log Frequency Power Coefficients).

2.3.1.1 Pitch

De acordo com (PENTEADO, 2009), Pitch é a sensação psicofísica da frequência fundamental ou forma como julgamos o som, no que diz respeito à sua intensidade, considerando-o mais grave ou mais agudo. Onde pode ser entendido como a altura (frequência) percebida em um sinal de áudio. O pitch, é considerado uma característica extremamente importante para o reconhecimento, porém (LIN; JENG, 2017) e (IRIYA, 2014) afirmam que os valores do pitch em si são menos relevantes do que suas alterações instantâneas, como por exemplo suas derivadas e sua energia (JR et al., 2017).

2.3.1.2 Energia

A energia, segundo (JR et al., 2017), é a intensidade sonora percebida pelo ouvido humano. O ouvido consegue distinguir a intensidade sonora em decibéis, entretanto é difícil medir a intensidade sonora diretamente. Por este motivo, o que é usado normalmente é a energia do sinal, capturada através da amplitude das amostras, onde a curva de energia depende de muitos fatores como os fonemas, o locutor, a cultura do locutor, como também do estado emocional do locutor.

2.3.1.3 Formantes

As formantes representam picos na resposta em frequência e podem ser entendidas como frequências de ressonância, onde parte das características das vogais é decorrente das duas primeiras formantes, o que as tornam extremamente importantes para o estudo da fonética e de reconhecimento de voz (JR et al., 2017).

2.3.1.4 Coeficientes de Potência em Escala Logarítmica

Os coeficientes de potência em escala logarítmica (LFPC) são coeficientes que refletem a distribuição espectral de energia do sinal. Segundo (JR et al., 2017) o cálculo destes coeficientes é feito passando-se o espectro do sinal por um banco de filtros que delimitam sub-bandas e calculam a média da energia nestas sub-bandas.

2.3.1.5 Coeficientes Mel Cepstrais

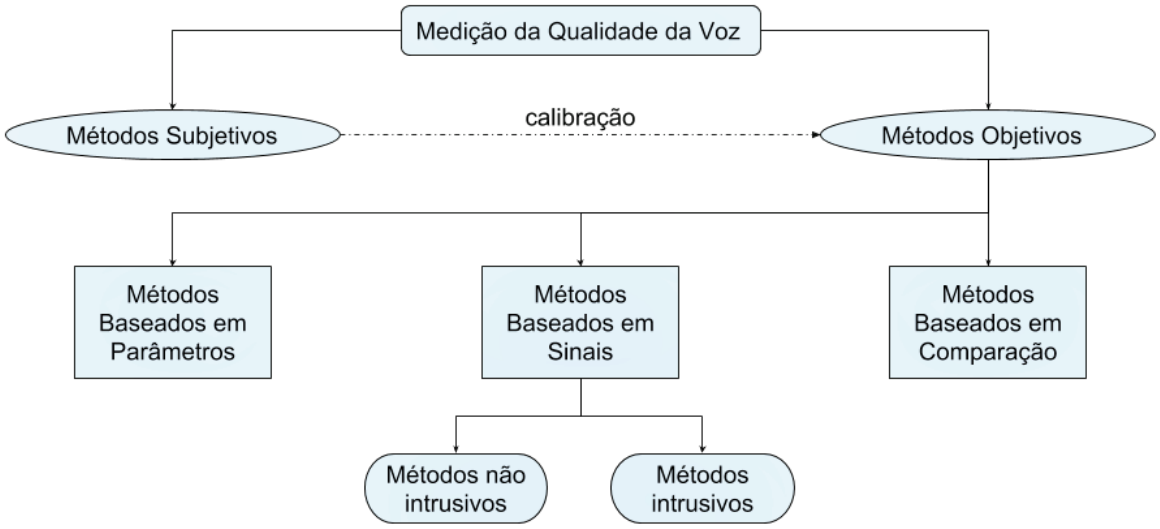
Os coeficientes mel cepstrais (MFCC) são uma representação paramétrica do espectro de frequências do sinal de voz (RIBEIRO et al., 2014). De acordo com (JR et al., 2017) a ideia dos coeficientes mel cepstrais vem do inverso do espectro do sinal, onde através da transformada de Fourier obtemos o espectro do sinal sobre o domínio do tempo, o cepstro é obtido através da transformada inversa de Fourier do log espectro (RIBEIRO et al., 2014).

A diferença entre o mel-cepstro e o cepstro é que para os coeficientes mel-cepstrais são aplicados filtros com bandas de frequências igualmente espaçadas na escala mel. Onde o uso da escala mel advém de estudos da Psicoacústica, pois a escala mel se aproxima melhor do sistema auditivo humano, que possui um comportamento não-linear na frequência (JR et al., 2017).

2.4 Métodos de avaliação de qualidade de Voz

Segundo as recomendações da ITU-T, os métodos de avaliação da qualidade da voz são divididos em dois métodos, os métodos objetivos, onde o índice MOS é obtido mediante o uso de um algoritmo, ou nos métodos subjetivos, onde há intervenção de indivíduos para obter o resultado da avaliação.

Figura 2.3 – Métodos de avaliação da qualidade da voz.



Fonte: Autor

Nos métodos subjetivos, os resultados são baseados em testes realizados em ambiente controlado, onde usuários classificam a qualidade da voz de acordo com suas experiências obtidas no experimento. Já no método objetivo existe a necessidade de um algoritmo predizer uma avaliação humana da qualidade do sinal da voz (REC, 1996). A Figura 2.3 ilustra a classificação dos métodos de avaliação da qualidade da voz.

Os métodos objetivos são divididos em 2 tipos, os não intrusivos e os intrusivos. Nos métodos intrusivos há a necessidade de um sinal de referência para comparar com o sinal no destino e avaliar a qualidade do sinal (REC, 1996), com isso os métodos intrusivos são mais confiáveis e usados como referência para avaliações objetivas.

Nos últimos anos, estudos vêm mostrando que as avaliações subjetivas podem ser realizadas de uma forma remota. Este método de avaliação é chamado de crowdsourcing, no qual os usuários localizados em qualquer lugar no mundo, cadastrados em plataformas que oferecem este serviço, realizam diversas tarefas, como avaliação da qualidade de imagens, vídeos ou áudios.

Por outro lado, os métodos não intrusivos são os métodos que precisam somente do sinal no destino ou em qualquer outro ponto para ser avaliado, são métodos rápidos, possibilitando o uso em serviços e aplicações em tempo real (ASSEMBLY, 2000).

Outra vantagem dos métodos intrusivos é o menor tempo na execução das avaliações, comparado com os métodos comuns que são realizados em laboratórios.

Para determinar a qualidade do arquivo de áudio de voz, são utilizadas as ferramentas ITU-T P.563 e P.862 para analisar o sinal da voz, tendo como resultado dessa avaliação um índice chamado de MOS (RODRÍGUEZ, 2009) e (RIX et al., 2006). O valor máximo de MOS atingido pelas análises das ferramentas das recomendações da ITU-T P.563 e P.862 é de 4.5, a Tabela 2.4 representa essa escala, onde define que o mínimo aceitável para uma comunicação é 3.6 para sinais de banda estreita.

Tabela 2.4 – Escala de índice MOS da recomendação ITU-T P.862

Nível satisfação do usuário	**MOS**	
Muito satisfeito	4.5	Desejável
Satisfeito	4.0	
Alguns usuários insatisfeitos	3.6	Aceitável
Muitos usuários insatisfeitos	3.1	Qualidade não aceitável
Praticamente todos os usuários insatisfeitos	2.6	
Não recomendado	1.5	

Fonte: Adaptado da Recomendação P.862

2.4.1 Recomendação ITU-T P.800

A recomendação ITU-T P.800 descreve métodos e procedimentos para condução e avaliação subjetiva da qualidade do sinal em uma transmissão, onde as avaliações de equipamentos e sistemas de telecomunicações devem ser conduzidos utilizando apenas ouvintes ou métodos de pesquisa de testes subjetivos (REC, 1996).

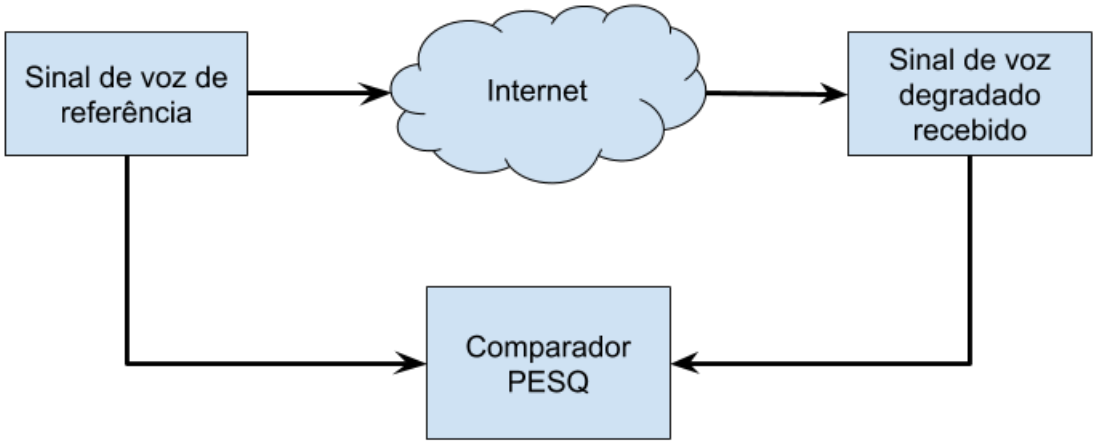
Esses métodos compreendem os testes de opinião onde se baseiam na opinião de vários indivíduos utilizando a escala de opinião para realizar a avaliação do sinal. O layout e a formulação das escalas de opinião que devem seguir o padrão alcançado através da experiencia adquirida pela ITU-T.

Nas escalas são avaliados diversos critérios onde cada um é avaliado recebendo uma pontuação, assim, ao final dos testes realizados, os resultados dos indivíduos participantes é somado e é realizada uma média aritmética dos valores, onde essa média é chamada de MOS.

2.4.2 Recomendação ITU-T P.862

A Recomendação ITU-T P.862 também conhecida como Avaliação Perceptiva da Qualidade de Voz (PESQ - *Perceptual Evaluation of Speech Quality*) é um método objetivo e intrusivo que é considerado como uma referência para a validação de outros métodos objetivos não intrusivos.

Figura 2.4 – Modelo PESQ.



Fonte: Adaptado da Recomendação ITU-T P.862

De acordo com a recomendação ITU-T P.862 (ASSEMBLY, 2000), o PESQ compara um sinal original com um sinal de áudio degradado resultante da passagem do áudio original pela rede de telecomunicação. O algoritmo também trata os atrasos durante os silêncios e durante a atividade de voz dos indivíduos (ASSEMBLY, 2000).

Segundo a recomendação ITU-T P.862 (ASSEMBLY, 2000), um modelo computacional é utilizado para comparar a saída do dispositivo sob teste com a entrada. Ao final, a representação interna dos sinais original e degradado é processada para considerar os efeitos de variações de ganho local e filtragem linear que podem, caso não sejam muito severos, ter pouco significado perceptivo. Efeitos mais graves, ou variações rápidas, são apenas parcialmente compensadas, de modo que um efeito residual permanece e contribui para o distúrbio perceptual global.

Isto permite que um pequeno número de indicadores de qualidade possa ser utilizado para modelar todos os efeitos subjetivos, os quais são combinados para dar qualidade objetiva de audição MOS (ASSEMBLY, 2000).

2.4.3 Recomendação ITU-T P.563

A recomendação ITU-T P.563, faz uso de características não intrusivas. Segundo (MALFAIT; BERGER; KASTNER, 2006), possui ótima vantagem para avaliar em tempo real o sinal de voz recebido, uma vez que não há uma necessidade de comparação com o sinal emitido. Consegue avaliar a qualidade de voz em aplicações de telefonia de 3,1 kHz de largura de faixa, porém possui desvantagens como áudios com intervalos de fala maiores que 200 ms, não são considerados naturais, o que pode interferir na pontuação MOS. Em sua análise pode haver um

erro, visto que a ferramenta não sabe diferenciar o silêncio natural de uma conversa com falhas na rede (RODRÍGUEZ, 2009), (NUNES, 2017). Seu uso é recomendado para monitorar as redes em tempo real e avaliar os sinais de voz em um dado ponto da conexão telefônica, levando em conta todas as classes de distorção que ocorrem nas redes.

O sistema de pontuação utilizado é baseado na escala MOS-LQO (*Mean Opinion Score-Listening Quality Objective*), que está de acordo com a recomendação ITU-T P.800. A pontuação calculada de um determinado sinal pode ser comparada com a qualidade percebida por um ouvinte humano que ouve o mesmo sinal com um dispositivo convencional de telefonia no mesmo ponto, e os valores dados vão de 1 a 5.

A validação da recomendação ITU-T P.563 inclui todos os experimentos disponíveis no processo de validação da recomendação ITU-T P.862. Além disso, seu algoritmo foi testado de forma independente com arquivos de voz desconhecidos, criados sob requisitos definidos estritamente para este fim por laboratórios independentes.

O escopo da recomendação ITU-T P.563 é dado com base nos resultados de referência do Grupo de Estudos 12 da ITU de 2003 e consiste em uma relação de fatores de testes, tecnologias de codificação e aplicações às quais esta recomendação se aplica.

A recomendação ITU-T P.563 cita que, embora haja uma correlação de, aproximadamente, 0.89 entre as pontuações objetivas e subjetivas, tanto para bases conhecidas quanto para bases desconhecidas, o algoritmo desta recomendação não pode substituir testes subjetivos. Porém, pode-se aplicar estas medições onde os testes subjetivos seriam muito caros.

A recomendação ITU-T P.563 também não fornece uma avaliação global da qualidade da transmissão, medindo apenas alguns efeitos da distorção da voz unidirecional e do ruído sobre a qualidade percebida da voz. O algoritmo pontua o sinal de voz da forma como é percebida por um ouvinte humano, usando um dispositivo telefônico convencional e com um nível de pressão sonora (SPL - Sound Pressure Level) de 79 dB no ponto de referência do ouvido (ERP - Ear Reference Point).

Assim somente os efeitos da perda de sonoridade, atrasos, eco e outras deficiências relacionadas à qualidade da voz não afetam a pontuação dada pelo algoritmo da recomendação ITU-T P.563. Porém de acordo com (NUNES, 2017) é importante ressaltar que o algoritmo da recomendação ITU-T P.563 foi projetado para predizer a qualidade da voz humana, não sendo recomendado para outros tipos de sinais de áudio não vocais, onde para isso os sinais de voz digitalizados devem seguir os seguintes requisitos:

- Frequência de amostragem de 8 kHz;
- Resolução de amplitude PCM linear de 16 bits;
- Atividade mínima de voz de 3 segundos;
- Mínimo de 25% e máximo de 75% de atividade de voz;
- Nível de voz entre –36 e –16 Decibel to Overload Point (dBov).

Esta recomendação realiza um ajuste do sinal para -26 dBov, apesar de estar dentro da faixa aceitável do último requisito, para evitar artefatos adicionais em função da baixa relação entre o sinal e o ruído (SNR - Signal-to-Noise Ratio) ou no corte de faixa de amplitude.

Funcionamento

Segundo a recomendação ITU-T P.563 (MALFAIT; BERGER; KASTNER, 2006), a abordagem dada por seu algoritmo deve ser vista pela perspectiva de um especialista que começa com o modelo do dispositivo receptor e, em seguida, um algoritmo de detecção de atividade de voz (VAD - Voice Activity Detector) é usado para identificar a voz que será avaliada.

Na etapa de pré-processamento várias análises são realizadas separadamente no sinal de voz, que detecta um conjunto de parâmetros do sinal. Estas análises serão aplicadas em primeiro lugar para todos os sinais. E então, baseando-se em um conjunto restrito de parâmetros chave, a principal classe de distorção é identificada e, juntos, eles são utilizados para ajustar o modelo de qualidade de voz no algoritmo.

De acordo com (NUNES, 2017), de forma resumida o processo de parametrização do sinal a ser submetido ao algoritmo da recomendação ITU-T P.563 pode ser dividido em três blocos funcionais independentes que correspondem às principais classes de distorção (MALFAIT; BERGER; KASTNER, 2006). São eles:

- Análise do trato vocal e artificialidade da voz, que se aplica a vozes masculinas, femininas e à robotização independente de gênero;
- Análise de ruídos aditivos fortes, que podem ser de baixa SNR estática ou baixa SNR segmentar;
- Interrupções, silêncios e recorte de tempo.

2.4.4 E-Model

Segundo (BERGSTRA; MIDDELBURG, 2003), o E-Model é um modelo computacional que avalia os efeitos combinados de variações em diversos parâmetros de transmissão, que afetam a qualidade da chamada telefônica. Resultando em um valor chamado de Fator R, que é uma derivação de atrasos e dos fatores de deterioração causados pelos equipamentos da rede, fazendo com que ele possa ser mapeado para um MOS estimado.

De acordo com (ASSEM et al., 2013), o E-Model original é muito complexo pois envolve diversos fatores fazendo com que o processamento de voz não esteja relacionado com o índice do Fator R. Por esta razão, uma versão simplificada do E-Model foi desenvolvida com o objetivo de focar nas partes mais importantes e para que possa ser usada em sistemas de monitoramento.

Para isso o modelo simplificado leva em conta somente o codec e as condições atuais da rede, que são os dois principais fatores que afetam a qualidade da voz. A equação do E-Model simplificado onde é calculado o valor de avaliação R é expressa pela Equação 2.1 (ASSEM et al., 2013).

$$R = R_0 - I_{codec} - I_{packetloss} - I_{delay} \tag{2.1}$$

Onde R_0 representa o sinal básico para relação de ruído, I_{delay} representa os atrasos introduzidos de ponta a ponta, I_{codec} é o fator de codec e o $I_{packetloss}$ é a taxa de perda de pacotes dentro de um tempo particular. Finalmente, o valor R é mapeado para a pontuação MOS (ASSEM et al., 2013).

Segundo (BERGSTRA; MIDDELBURG, 2003), o fator R varia de 100 (excelente) até 0 (péssimo) e, o MOS varia de 5 a 1. Um MOS estimado pode ser diretamente calculado do Fator R do E-Model, conforme demonstra a Tabela 2.5, onde os valores do Fator R do E-Model são mostrados à esquerda, com os valores correspondentes ao MOS à direita. O nível de satisfação dos ouvintes é apresentado na coluna do meio. E de acordo com (BERGSTRA; MIDDELBURG, 2003) e (WALKER; HICKS, 2002) para avaliar a qualidade da chamada VoIP é extremamente recomendado a utilização do E-Model.

Tabela 2.5 – Medida de Qualidade da Chamada Telefônica.

Fator R	Nível de Satisfação	MOS
90 a 100	Muito satisfeitos	4.3 a 4.5
80 a 90	Satisfeitos	4.0 a 4.3
70 a 80	Alguns usuários insatisfeitos	3.6 a 4.0
60 a 70	Muitos usuários insatisfeitos	3.1 a 3.6
50 a 60	Praticamente todos insatisfeitos	2.8 a 3.1
0 a 50	Não recomendado	1.0 a 2.8

Fonte: Adaptado da recomendação ITU-T G.107 - E-Model.

Segundo (BERGSTRA; MIDDELBURG, 2003), os valores recomendados pelo E-Model sobre as condições de atraso e perda para que uma rede possua uma boa qualidade da chamada VoIP são:

- Atraso fim a fim deve ser menor que 150ms;
- Limite máximo de 50ms para a variação de atraso;
- Taxa de perda limitada a 3% (mas recomendável é ser inferior a 0,50%).

E o não respeito destes limites gera degradações da qualidade de voz perceptíveis ao usuário.

2.5 Qualidade de Experiência - QoE

De acordo com (LAGHARI; CONNELLY, 2012), o termo de QoE tem sido o principal tema de pesquisa da comunidade científica relacionada a telecomunicações, porém, apesar disso, este conceito pode ser aplicado a várias outras áreas do conhecimento.

O conceito de QoE está relacionado à satisfação do cliente com os serviços de telecomunicação em geral. Se o usuário de um serviço não sente que está recebendo um serviço proporcional ao valor que está pagando, ele pode reclamar do serviço com o provedor ou até mesmo cancelar seu contrato (FIEDLER; HOSSFELD; TRAN-GIA, 2010).

As métricas de qualidade de serviço são frequentemente utilizadas pelo técnicos para avaliar a qualidade de experiência dos serviços entregues (REICHL et al., 2015), porém não podem ser utilizadas para determinar a experiência dos usuários finais pois ela está diretamente relacionada às características das próprias redes dos usuários.

Os resultados da avaliação de QoE são apresentados como um valor escalar simples, tipicamente usando o índice MOS da ITU-T P.800 (REC, 1996), pois a definição de QoE tem

uma relação entre emoções, características de qualidade e percepção de qualidade (ARNDT et al., 2014), que, embora sejam úteis, suas limitações são evidentes para várias aplicações por utilizar somente a média aritmética (HOSSFELD et al., 2016). Com isso impede ao provedor de serviço de ter um número real de quantos usuários estão insatisfeitos. Porém apesar das limitações esta é a forma estabelecida pela ITU-T de se avaliar QoE em cenários de chamadas de voz (ARNDT et al., 2014).

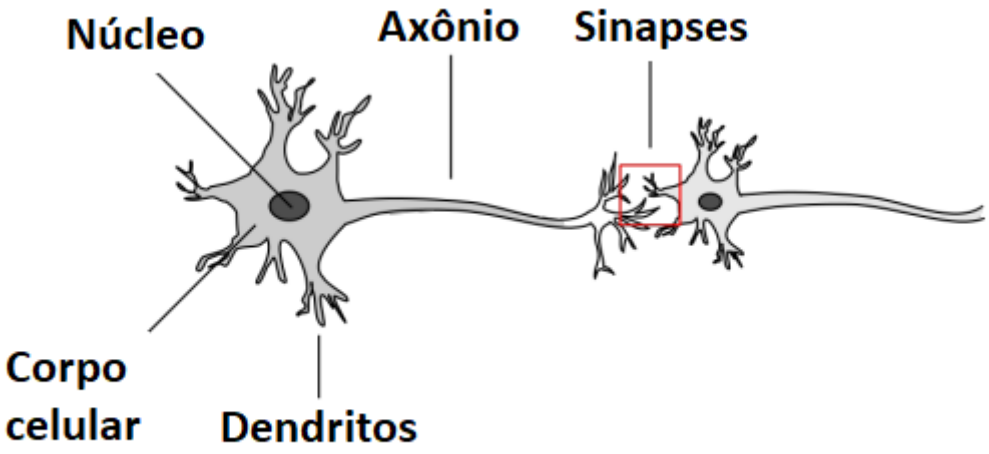
2.6 Redes Neurais Artificiais

As redes neurais artificiais (RNAs) podem ser definidas como um tipo de rede inspirada nos sistemas nervosos biológicos que processam informação (AZEVEDO, 2016), e estas redes podem ser utilizadas em vários tipos de aplicações, tais como reconhecimento de padrões, diagnósticos médicos, aplicações financeiras, mineração de dados, reconhecimento de gestos, de voz e escrita, e diversas outras aplicações.

De acordo com (AZEVEDO, 2016), o cérebro humano é composto de bilhões de células conhecidas como neurônios. Onde cada neurônio conecta-se a milhares de outros através de conexões conhecidas como sinapses, onde o cérebro contém dezenas de trilhões delas. O sistema nervoso humano forma-se através deste complexo conjunto de conexões, os neurônios humanos podem ser resumidos nos seguintes elementos: os dendritos, um corpo celular e os axônios. Onde os dendritos são responsáveis pela recepção das informações, o corpo celular é responsável por armazenar o núcleo da célula e os axônios realizam o transporte destas informações. Os sinais são transmitidos aos demais neurônios através da sinapse.

O modelo de um neurônio humano pode ser visto na Figura 2.5.

Figura 2.5 – Modelo de neurônio humano.



Fonte: Adaptado de (AZEVEDO, 2016).

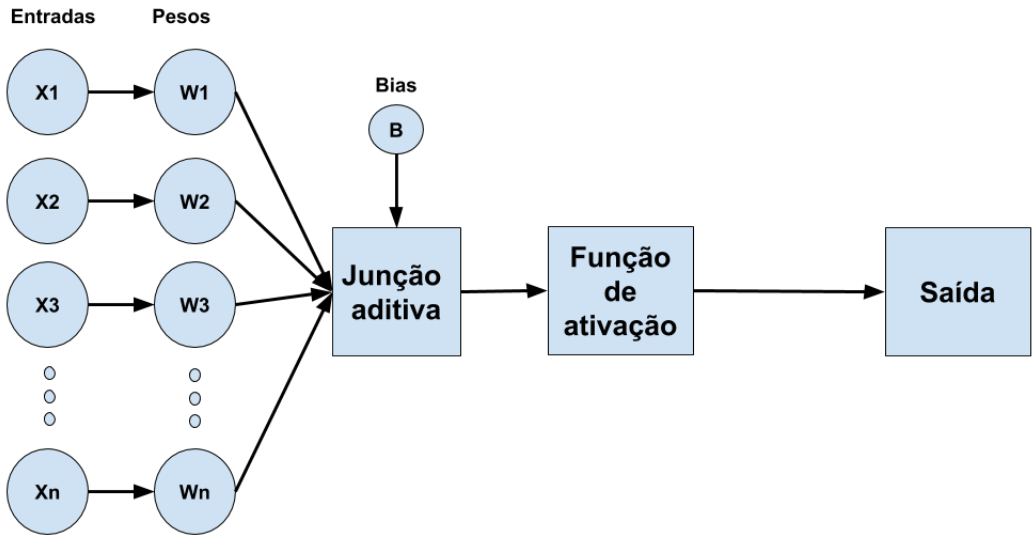
2.6.1 Neurônios artificiais

Em (AZEVEDO, 2016) afirma-se que o psiquiatra e neuroanatomista Warren S. McCulloch, juntamente ao lógico Walter Pittso desenvolveram na década de 1940 o modelo de rede neural artificial e que de acordo com (AZEVEDO, 2016), estas redes se assemelham ao cérebro humano pelo fato de que o conhecimento adquirido se dá através do processo de aprendizagem e é armazenado como forças de conexão entre seus neurônios, como valores dos pesos sinápticos. (AZEVEDO, 2016) e (HAYKIN, 2007) também descreveram que um neurônio artificial é composto pelos seguintes elementos:

- Um conjunto de entradas, contendo pesos ou forças. Os valores passados a estas entradas são multiplicados pelos pesos;
- Um somador (junção aditiva), responsável por computar a soma das entradas recebidas por este neurônio;
- Uma função de ativação, responsável por limitar o valor de saída do neurônio, em geral no intervalo [0,1] ou [1,1].

Onde estas funções são utilizadas como uma função de transferência de valores entre neurônios (SHARMA; RAI; DEV, 2012). À entrada de um neurônio artificial, pode-se ainda aplicar um valor adicional definido como bias (AZEVEDO, 2016). O modelo computacional de um neurônio contendo seus diferentes componentes é apresentado na Figura 2.6.

Figura 2.6 – Modelo de um neurônio artificial.



Fonte: Adaptado de (AZEVEDO, 2016)

(AZEVEDO, 2016) fala que as redes neurais são capazes também de realizar seu aprendizado através do conjunto de dados de entrada e saídas esperadas, tornando-as mais robustas e adaptáveis aos problemas propostos. De acordo com (MAIND; WANKAR et al., 2014), alguns dos principais benefícios deste tipo de rede são sua capacidade de aprender a realizar tarefas a partir de dados inseridos, além de criar sua própria representação da informação recebida.

2.6.2 Perceptrons

Segundo (AZEVEDO, 2016), o modelo perceptron se trata da mais simples rede neural artificial que existe, sendo utilizada para classificação de problemas linearmente separáveis. Foi inicialmente proposto por Frank Rosenblatt ao final da década de 1950 e é composto por um conjunto de entradas, um processador e uma saída (SCHIFFMAN, 2012). Os perceptrons utilizam como saída uma função degrau, na qual o retorno é +1 caso a soma de suas entradas for maior que um limiar predefinido, e -1 caso contrário (AZEVEDO, 2016).

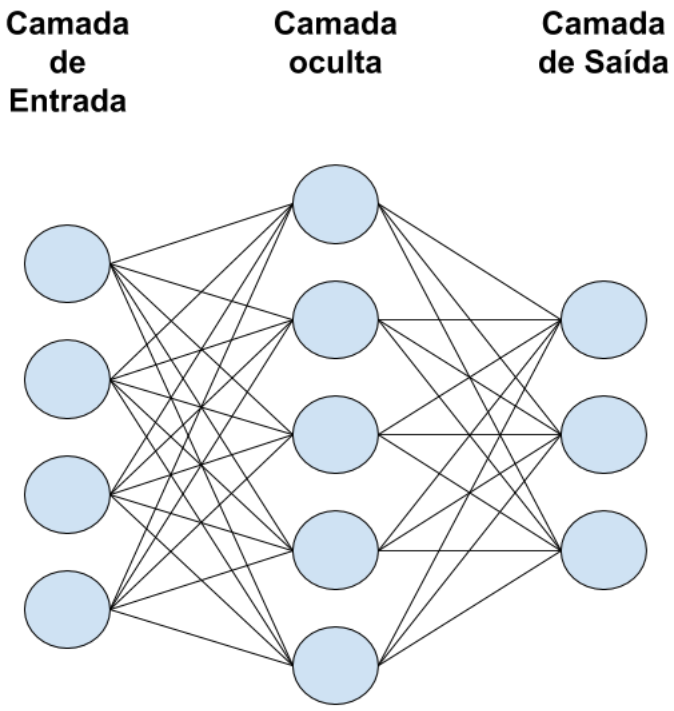
De acordo com (COPPIN, 2010), o processo de aprendizagem do perceptron pode ser descrito a partir do seguinte procedimento, os pesos inicialmente aleatórios são distribuídos a suas entradas, em seguida, um dado de treinamento é inserido e calculado. Caso a saída não seja

a esperada, estes pesos são ajustados de acordo com uma taxa de aprendizado predefinida. Os demais dados de treinamento são então inseridos e testados da mesma forma e caso haja erros, os pesos são novamente ajustados (AZEVEDO, 2016). Todo este processo é realizado de maneira contínua, até que não haja mais erros de validação para os dados da base de treinamento, onde o número de execuções do algoritmo de treinamento são denominadas épocas.

2.6.3 Perceptrons multicamadas

De acordo com (AZEVEDO, 2016), os perceptrons de múltiplas camadas (MLP - *Multi-layer Perceptron*) são redes neurais capazes de resolver problemas não linearmente separáveis, ou seja, problemas complexos. Estas redes podem ser consideradas como uma generalização dos perceptrons comuns e apresentam em sua estrutura uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. As camadas ocultas ficam entre as camadas de entrada e saída. Neste modelo de rede, todos os nós são conectados diretamente àqueles da camada seguinte, de forma que os sinais seguem um fluxo da entrada à saída e por isso, estas redes são conhecidas como redes do tipo feedfoward. Um exemplo de arquitetura das redes MLP pode ser visto através da Figura 2.7.

Figura 2.7 – Modelo de uma rede do tipo perceptron com múltiplas camadas.



Fonte: Adaptado de (AZEVEDO, 2016)

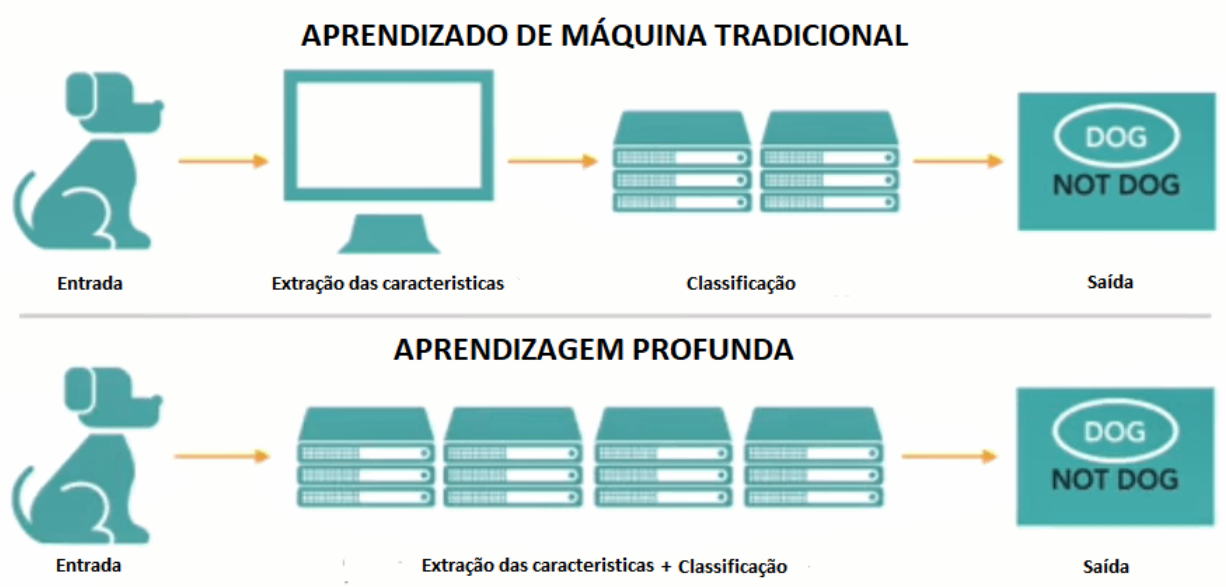
Segundo (AZEVEDO, 2016), o aprendizado neste tipo de rede ocorre através de um algoritmo interativo de retro-propagação (BackPropagation). Segundo (COPPIN, 2010), neste processo de aprendizagem, pesos aleatórios são atribuídos aos nós de entrada da rede, assim como ocorre nas redes perceptron comuns. Os dados de treinamento são então carregados à rede e percorrem seu caminho até a saída. Caso a saída contenha erros, este erro é calculado, propagado às camadas anteriores e o ajuste nos pesos é realizado em todos os neurônios (AZEVEDO, 2016). Este processo se repete enquanto o erro da saída for maior que um limiar predefinido. O algoritmo contém um parâmetro que representa a taxa de aprendizado que define a velocidade com a qual ela ocorre, onde em valores maiores, a aprendizagem ocorre de maneira mais rápida, já com valores menores, o processo é mais preciso.

2.7 Deep Learning

Técnicas de aprendizado de máquina (ML - *Machine Learning* permitiram com que computadores consigam reconhecer padrões do mundo real e inferir decisões que antes pareciam subjetivas (LECUN; BENGIO et al., 1995). No entanto, a facilidade que computadores possuem na resolução de problemas com regras bem definidas como cálculos matemáticos, jogos de tabuleiro ou cartas, contrasta com a dificuldade encontrada em atividades como reconhecimento de padrões em imagens, músicas, textos, etc.

Por conta disso, as dificuldades enfrentadas sugerem que sistemas de Inteligência Artificial (IA) precisam da capacidade de adquirir seu próprio conhecimento, extraindo padrões em dados brutos. No entanto, técnicas convencionais de ML são limitadas na sua habilidade de processar dados naturais em sua forma bruta (LECUN; BENGIO; HINTON, 2015).

Figura 2.8 – Machine Learning vs Deep Learning.



Fonte: Adaptado de (COPELAND, 2016).

Assim, os algoritmos de ML para reconhecimento de padrões necessitam bastante das técnicas de extração de características que vão transformar os dados brutos em uma representação eficaz desses dados para que um classificador possa inferir os padrões nos dados de entrada. Por isso, o extrator de características tem uma enorme influência no desempenho de algoritmos de ML.

Outra dificuldade em aplicações de ML é que muitos fatores influenciam na variação dos dados observados, tendo como exemplo identificar o modelo de um cachorro ou gato em diferentes iluminações e posições ou reconhecer o formato do corpo ao variar o ângulo de visualização. Desse modo, obter boas representações dos dados para reconhecimento de padrões é bastante dispendioso.

Aprendizagem de representação (RL - *Representation Learning*) é um conjunto de métodos que permite que as máquinas recebam dados brutos como entrada e, automaticamente, descubram as representações necessárias para detecção ou classificação (LECUN; BENGIO; HINTON, 2015).

De acordo com (DENG; YU et al., 2014), a arquitetura de Deep Learning (aprendizado profundo) consiste em um conjunto de técnicas de aprendizado de máquina, focados na utilização de diversas camadas de processamento de informação, sendo aplicado com sucesso em diversas áreas, tais como visão computacional e reconhecimento de objetos. Estas arquiteturas utilizam redes com diversas camadas intermediárias (TUSHAR, 2015), sendo que através delas,

segundo (WU et al., 2014), tornou-se então possível reconhecer objetos sem a necessidade da extração prévia de características.

Para (LECUN; BENGIO; HINTON, 2015), as redes de aprendizagem profunda baseiam-se na propriedade hierárquica dos sinais, sendo a rede neural convolutiva um tipo de rede de aprendizagem profunda que obteve sucesso por ser fácil de treinar e possuir melhor generalização se comparada às redes totalmente conectadas.

O Deep Learning (DL), uma subárea da ML, possui métodos de RL com múltiplos níveis de representação, obtidos pela composição de módulos simples, mas não lineares, que transformam a cada representação de um nível em uma representação de nível superior, ligeiramente mais abstrato. Ou seja, por meio de um conjunto de algoritmos tenta criar modelos de abstrações de alto nível em dados através de múltiplas camadas de processamento por meio de várias transformações não lineares (BENGIO; COURVILLE; VINCENT, 2013), (DENG; YU et al., 2014), (SCHMIDHUBER, 2015), (LECUN; BENGIO; HINTON, 2015).

Com a composição suficiente de tais transformações, funções muito complexas podem ser aprendidas. Dessa forma, para tarefas de classificação, as camadas mais altas da representação amplificam aspectos da entrada que são importantes para a discriminação e suprimem variações irrelevantes.

A título de exemplo, uma imagem surge na forma de uma matriz de valores de pixels, e as características aprendidas na primeira camada de representação tipicamente representam a presença ou a ausência de arestas em determinadas orientações e posições na imagem.

A segunda camada, tipicamente detecta formas através da identificação de arranjos particulares de arestas, independentemente de pequenas variações nas posições das bordas.

A terceira camada pode montar formas em combinações maiores correspondendo a partes de objetos, e as camadas subsequentes detectariam objetos ou cenas a partir das combinações dessas partes, que demonstra as diversas camadas de um método de RL (LECUN; KAVUKCUOGLU; FARABET, 2010).

Dessa forma, o principal aspecto da DL é que as camadas de features são, na verdade, aprendidas a partir dos dados brutos usando um processo de aprendizagem de propósito geral.

Alternativamente, o termo DL tem sido também usado como referência a redes neurais artificiais (RNA) feed forward. Segundo (SCHMIDHUBER, 2015) e (GOODFELLOW; BENGIO; COURVILLE, 2016), a diferença entre as RNA's clássicas e uma Deep Neural Network (DNN) é a profundidade ou número de camadas, também denominado de cadeias de possível

aprendizagem (CAP - *Credit Assignment Path*). (BENGIO; COURVILLE; VINCENT, 2013) afirmam que variar o número e o tamanho de cada camada proporciona valores variados de abstração.

Dessa forma, os algoritmos de DL ajudam a separar as abstrações e escolher quais recursos são úteis para o aprendizado. Mesmo que os métodos de DL sejam usados com frequência em treinamento supervisionado (DENG; YU et al., 2014), (KRIZHEVSKY; SUTSKEVER; HINTON, 2012), (MOHAMED et al., 2011), (COLLOBERT; WESTON, 2008), alguns algoritmos de DL vêm sendo aplicados em treinamento não supervisionado também (GRAVES; MOHAMED; HINTON, 2013), (LI et al., 2016). A razão disso é que dados não rotulados são encontrados em maior abundância e isso se converte então em uma importante vantagem destes algoritmos.

É possível destacar também que algumas estruturas de DL tais quais a Deep Neural Networks (DNN), a Recurrent Neural Networks (RNN), a Convolutional Neural Networks (CNN) e Deep Belief Networks (DBN) vêm sendo aplicadas em diversas áreas, obtendo resultados muito melhores do que os resultados indicados até então em competições de ML (KRIZHEVSKY; SUTSKEVER; HINTON, 2012).

Em última análise, um importante progresso na IA pode ocorrer por meio da combinação de RL com outros métodos de representação de dados mais complexos e embora a DL venha sendo utilizada no reconhecimento de escrita e voz, por exemplo, novos paradigmas são necessários para substituir as regras baseadas em manipulação de expressões simbólicas por operações em grandes vetores (BOTTOU, 2014).

2.7.1 Redes Neurais Recorrentes

As Redes Neurais Recorrentes (RNN) possuem duas características específicas o processamento de neurônios e a topologia de redes. O processamento de um neurônio é descrito como uma equação diferencial com entrada externa sequencial no tempo. E a topologia da rede possui conexões assimétricas com loops de feedback entre cada neurônio. Essas duas características implicam que o RNN inclui a maioria das redes neurais convencionais, como redes neurais multicamadas e redes Hopfield (TAKASE; GOUHARA; UCHIKAWA, 1993).

2.7.2 Redes Neurais por Convolução

As Redes Neurais por Convolução (CNN) têm sido aplicadas com sucesso em problemas de classificação de imagem em diversas outras aplicações como rastreamento de objetos, estimativa de poses, detecção e reconhecimento de texto, detecção visual de saliência, reconhecimento de ações e rotulagem de cenas (GUO et al., 2017). Embora, apesar de poderosas as abordagens de classificação de imagens baseadas na CNN apresentadas na literatura sejam muito bem sucedidas, elas requerem uma grande quantidade de memória para funcionar (ÇALIK; DEMIRCI, 2018).

A rede neural convolucional é uma classe de rede neural profunda que explora a forte correlação local e espacial em imagens naturais, alcançando um ótimo desempenho na área de análise visual. Recentemente, as CNNs tem sido empregadas na área de processamento acústico e tem provado ser capazes de aprender o padrão espectro-temporal do som e diferenciá-lo para fins de classificação (SHU; SONG; ZHOU, 2018). Nos últimos anos, vários algoritmos de classificação de cenas acústicas foram propostos onde os algoritmos mais utilizados incluem Máquina de Vetores de Suporte (SVM - *Support-Vector Machine*), Modelos de Mistura de Gaussianas (GMM - *Gaussian Mixture Models*), Modelos Ocultos de Markov (HMM - *Hidden Markov Models*), ou a hierarquia desses métodos.

Como uma eficiente ferramenta de aprendizagem profunda, a rede neural convolucional demonstrou seu sucesso na área de processamento acústico para resolver o problema de classificação de som e apresentou seu desempenho superior em relação às abordagens tradicionais. Em geral, os dados brutos são manipulados no formato de frequência de tempo para alimentar a rede neural profunda. Diferentes resoluções de frequência temporal na extração de características de sinal têm diferentes efeitos na precisão da classificação (SHU; SONG; ZHOU, 2018).

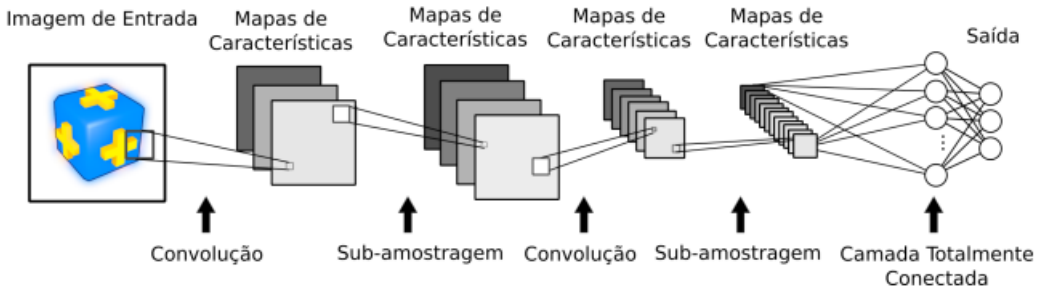
Nos últimos anos, temas como, o Codificador Automático, a Codificação Esparsa, a Máquina Boltzmann Restrita, Redes de Crenças Profundas e Redes Neurais Convolucionais são comumente usados em aprendizagem profunda. Entre diferentes tipos de modelos, as CNN tem demonstrado alto desempenho na classificação de imagens (GUO et al., 2017).

A classificação de imagens desempenha um papel importante na visão computacional, tem um significado muito importante em nosso dia a dia. A classificação de imagens é realizada através do pré-processamento de imagens, segmentação de imagens, extração de características-chave e identificação de correspondência (GUO et al., 2017). A extração de características e o classificador foram integrados a uma estrutura de aprendizado que supera o método tradicional

de dificuldades de seleção de recursos. A ideia da aprendizagem profunda é descobrir múltiplos níveis de representação, com a esperança de que recursos de alto nível representem uma semântica mais abstrata dos dados. Um dos principais ingredientes da aprendizagem profunda na classificação de imagens é o uso de arquiteturas convolucionais (GUO et al., 2017).

De acordo com (GUO et al., 2017), a inspiração do design da rede neural convolucional vem da estrutura do sistema visual dos mamíferos, o modelo de estrutura visual baseado no córtex visual do gato foi proposto por Hubel e Wiesel em 1962. Os tipos de camada de rede neural convolucional incluem principalmente três tipos, a Convolutional layers, a Pooling layers e a Fully Connected Layer. A Figura 2.9 mostra a arquitetura do LeNet-5, onde a LeNet-5 é uma arquitetura de uma rede CNN que foi introduzida por Yann LeCun.

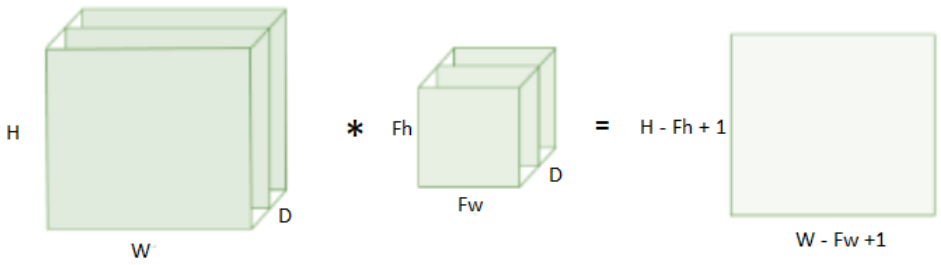
Figura 2.9 – Arquitetura de uma rede CNN.



Fonte: Adaptado de (GUO et al., 2017).

LeNet-5 é uma rede convolucional de 7 níveis feita por (LECUN; KAVUKCUOGLU; FARABET, 2010), que classifica dígitos, foi aplicada por vários bancos para reconhecer números escritos à mão em cheques (cheques) digitalizados em imagens de 32x32 pixels. A capacidade de processar imagens de alta resolução requer camadas maiores e mais convoluções, portanto, essa técnica pois uma limitação pela disponibilidade de recursos computacionais disponíveis.

Figura 2.10 – Arquitetura da LeNet-5.



Fonte: Adaptado de (LECUN; KAVUKCUOGLU; FARABET, 2010).

Como por exemplo na Figura 2.10, a LeNet-5 começa com uma imagem de 32 x 32 x 1, na primeira etapa, foram usados seis filtros 5 x 5 com passada 1 e foi obtido 28 x 28 x 6. Com um passo e sem preenchimento, a dimensão foi reduzida de 32 x 32 para 28 x 28. Então, utilizando uma média de agrupamento com um filtro de largura de 2 e a 2 passos e reduzindo a dimensão pelo fator de 2 e terminando com 14 x 14 x 6. Logo após foi usado outra camada convolucional com dezesseis filtros 5 x 5 e finalizando com 10 x 10 x 16. Em seguida, outra camada de pooling e finalizando com 5 x 5 x 16. Em seguida, a próxima camada é uma camada totalmente conectada (Full-Connected) com 120 neurônios. As camadas anteriores com os 400 parâmetros (5x5x16) se conectam a uma nova camada com 120 neurônios seguida deu uma outra camada com 84 neurônios. Utilizando essa rede era possível reconhecer os dígito de 0 a 9. Em uma versão moderna desta rede neural é utilizado a função softmax com uma saída de classificação de dez ondas (LECUN; KAVUKCUOGLU; FARABET, 2010).

- Convolutional layers: A camada convolucional é a parte central da rede neural convolucional, que tem conexões locais e pesos de características compartilhadas. O objetivo da camada Convolucional é aprender as representações das características das entradas. Como mostrado na Figura 2.9, a camada Convolucional é constituída por vários mapas de características. Cada neurônio do mesmo mapa de características é usado para extrair características locais de diferentes posições na camada anterior, mas para neurônios únicos, sua extração é a característica local das mesmas posições no antigo mapa de características. Para obter um novo recurso, os mapas de características de entrada são convolucionados com um kernel e, em seguida, os resultados são passados para uma função de ativação não linear, podendo obter diferentes mapas de características aplicando kernels diferentes de acordo com a Equação 2.2. A função típica de ativação é sigmoid, tanh e Relu (GUO et al., 2017).

$$(f * g)(t) \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} f(\tau)\, g(t-\tau)\, d\tau \tag{2.2}$$

- Pooling layers: O processo de amostragem é equivalente à filtragem difusa. A camada de pooling tem o efeito da extração de recurso secundário, onde pode-se reduzir as dimensões dos mapas de características e aumentar a robustez da extração de características. Geralmente é colocado entre duas camadas convolucionais. O tamanho dos mapas de características na camada de pool é determinado de acordo com a etapa de movimentação

dos kernels. As operações de pool típicas são pooling medium e maximum. Podemos extrair as características de nível alto de entradas empilhando várias camadas convolucional e de pooling (GUO et al., 2017)

- Flatten: E o processo de reorganizar o mapa de características em um vetor de uma coluna para que os dados possam ser inseridos na rede neural.

- Fully-connected layers: Em geral, o classificador da rede neural Convolucional é uma ou mais camadas totalmente conectadas. Eles pegam todos os neurônios da camada anterior e os conectam a cada neurônio da camada atual. Não há informações espaciais preservadas em camadas totalmente conectadas. A última camada totalmente conectada é seguida por uma camada de saída. Para tarefas de classificação, a regressão softmax é comumente usada por gerar uma distribuição de probabilidade bem executada das saídas. Outro método comumente usado é o SVM, que pode ser combinado com CNNs para resolver diferentes tarefas de classificação (GUO et al., 2017).

A convolução é um tipo especializado de operação linear e para ser considerada uma CNN é preciso que a rede (LECUN et al., 1998) detalhe o procedimento de treinamento de uma CNN em seu experimento e apesar da literatura não especificar uma arquitetura padrão, a quantidade de camadas necessárias está diretamente relacionada com o número de hierarquias de características importantes para representar e reconhecer um objeto do dataset utilizado. Um dataset com classes altamente distintas pode precisar de uma representação interna mais simples, enquanto que datasets com diferentes classes, muito parecidas, necessitam de um nível de representações internas maior para que a CNN seja capaz de distinguir uma classe da outra (AZEVEDO, 2016).

Devido à grande quantidade de parâmetros de uma CNN é possível a ocorrência de overfitting. Na tentativa de minimizar este problema comum, se utilizam técnicas de regularização, que tem como objetivo reduzir a quantidade de neurônios ativos quando uma determinada característica está presente na imagem. Essa redução da quantidade de neurônios ativos é similar ao comportamento do córtex visual dos mamíferos, onde o objetivo é que apenas pequenas porções de neurônios sejam ativadas (AZEVEDO, 2016). Essa ativação deve acontecer de acordo com as características observadas na imagem.

Segundo (AZEVEDO, 2016), a técnica proposta por (GEOFFREY et al., 2012), chamada Dropout, procura desativar aleatoriamente um conjunto de neurônios a cada iteração do

treinamento. Assim, com um número menor de ativações, é reduzido o problema de overfitting forçando cada camada da rede a se especializar em uma determinada característica de forma mais distinta.

De acordo com (AZEVEDO, 2016), o Dropout pode ser utilizado em qualquer camada da CNN, inclusive após as camadas full connected e após as camadas de convolução da rede ou em ambas as camadas. Com isso, é possível analisar que as CNN's proporcionam uma grande melhoria na representação e abstração dos dados quando comparadas com outras redes "Deep" com arquitetura mais simples.

3 TRABALHOS RELACIONADOS

Nesta seção são apresentados o desenvolvimento de alguns trabalhos que estão relacionados ao sistema proposto.

No trabalho de (LIU et al., 2014) a detecção e avaliação da tosse têm valor clínico crucial para doenças respiratórias. Utilizando redes neurais profundas (DNN) que foram aplicadas para modelar recursos acústicos na detecção de tosse, foi desenvolvido um sistema de detecção de tosse em duas etapas. O conjunto de dados experimentais utilizado para treinamento e validação contém gravações de áudio de 20 pacientes com duração de cerca de 24 horas por gravação. O desempenho do sistema proposto foram avaliados por sensibilidade, especificidade, medida F1 e média macro de recordação. Resultados experimentais mostraram que muitas das configurações DNN superam os Modelos de Mistura de Gaussianas (GMM) na sensibilidade, especificidade e medida F1 respectivamente.

(WROGE et al., 2018) utilizaram os biomarcadores derivados da voz humana que podem oferecer a visão de distúrbios neurológicos, como a doença de Parkinson (DP), por causa de sua função cognitiva e neuromuscular subjacente. A DP é um distúrbio neurodegenerativo progressivo que afeta cerca de um milhão de pessoas nos Estados Unidos, com aproximadamente sessenta mil novos diagnósticos clínicos realizados a cada ano (OBESO; OLANOW; NUTT, 2000). Historicamente, a DP tem sido difícil de quantificar e os médicos tendem a se concentrar em alguns sintomas enquanto ignoram os outros, baseando-se principalmente em escalas subjetivas de avaliação (LANGSTON, 2006). Devido à diminuição do controle motor que é a marca da doença, a voz foi utilizada como um meio para detectar e diagnosticar a DP. O trabalho desenvolvido por eles explora a eficácia do uso de algoritmos de classificação supervisionada, como redes neurais profundas, para diagnosticar com precisão indivíduos com a doença. A taxa de acurácia na precisão foi de 85% fornecido pelos modelos de aprendizado de máquina, excedendo a precisão do diagnóstico clínico médio de não especialistas que possuem uma acurácia de 73,8% e a precisão média dos especialistas em distúrbio de movimento que foi de 79,6% sem acompanhamento e 83,9% após acompanhamento (RIZZO et al., 2016).

De acordo com (MOHARIR et al., 2017) mesmo com o avanço rápido da tecnologia, ainda podemos observar uma quantidade significativa de mortes de crianças menores de cinco anos. A maioria dessas mortes em todo o mundo pode ser atribuída a várias condições médicas, das quais três são muito significativas: asfixia ao nascer, prematuridade e infecções. Asfixia ao nascer (asfixia perinatal) é uma condição médica que é caracterizada por padrões respiratór-

rios anormais em um recém-nascido que podem levar a danos irreversíveis ao cérebro ou, se negligenciados, podem ser fatais (MOHARIR et al., 2017). Na maioria dos casos, a condição é diagnosticada após o recém-nascido sofrer danos consideráveis. Visando diagnosticar a condição do recém-nascido foi construído um modelo de aprendizado de máquina pela qual a asfixia pode ser determinada em seus estágios iniciais, o modelo realiza a observação dos padrões no choro da criança e submetendo-a através de diferentes camadas de uma rede neural construída em um banco de dados de amostras previamente registradas de crianças afetadas. O software utilizado para a construção do modelo foi o NVIDIA DIGITS e a maior precisão alcançada pelo trabalho é de 94%.

De acordo com (TÜNDIK et al., 2017) os métodos de classificação automática são frequentemente utilizados no diagnóstico precoce de diferentes doenças que afetam a produção da fala. Esses métodos também podem ser aplicados para identificar amostras de fala de pacientes afetados por doença de Parkinson (DP) ou transtorno depressivo (DD). O trabalho desenvolvido tem como objetivo a aplicação de técnicas de detecção automática de tensões e fraseamento prosódico em amostras de fala patológica para avaliar em que medida essas ferramentas podem ser úteis para caracterizar de maneira não supervisionada os atributos prosódicos de amostras patológicas de indivíduos afetados por DP e DD e podendo também classificar as amostras como pertencentes a indivíduos saudáveis ou não saudáveis.

(MITTAG; MÖLLER, 2018) utilizaram as CNN's para detectar a quantidade de pacotes perdidos em sistemas de comunicação de voz. Para estimar a quantidade de pacotes perdidos, utilizaram espectrogramas dos sinais de fala transmitidos como entrada de uma rede neural convolucional, onde as interrupções causadas por pacotes perdidos podem ser claramente vistas no espectrograma do sinal degradado. O modelo proposto por eles permitiu estimar a taxa de perda de pacotes de um sistema de comunicação simplesmente usando o arquivo de fala gravado do lado do receptor, sem a necessidade do sinal de fala de referência que foi originalmente enviado através do canal. Os resultados mostraram que o modelo pode reduzir o erro de previsão em mais de 75% quando comparado a um modelo baseado em recursos do MFCC.

No trabalho de (AFFONSO et al., 2018), foram utilizados vários cenários de rede que consideram diferentes taxas de perda de pacotes (PLR) nos quais os sinais com problemas são avaliados usando o algoritmo descrito na recomendação ITU-T P.862. Os resultados mostraram uma relação entre os parâmetros de desvanecimento e PLR e o índice global de qualidade da fala. O objetivo principal do trabalho foi propor um modelo não intrusivo de classificação da

qualidade da fala baseado em uma Deep Belief Network (DBN) que considera as deficiências com fio e sem fio no sinal da fala. Os resultados demonstraram uma alta correlação entre o modelo proposto baseado no algoritmo DBN e o da recomendação ITU-T P.862. Para validação, foi utilizado o algoritmo não intrusivo da recomendação ITU-T P.563, o modelo proposto e o algoritmo da recomendação ITU-T P.862 atingiram uma acurácia média de 96,14% e 72,12%, respectivamente.

Em (OOSTER; MEYER, 2019) explorou-se uma *Deep Machine Listening* para Estimating Speech Quality (DESQ), que consegue prevê a qualidade de fala percebida com base nas probabilidades posteriores do fonema obtidas de uma rede neural profunda. A degradação dos fonemas é quantificada com a medida de Gini baseada em entropia que é comparada com a distância temporal média (MTD - *Mean Temporal Distance*) proposta anteriormente. Como as longas pausas de fala e perdas de pacote podem ter um grande efeito na qualidade da fala, o objetivo do trabalho era verificar se uma detecção de atividade de voz (VAD - *Voice Activity Detection*) possuía um efeito benéfico ou prejudicial sobre o poder preditivo do modelo proposto. A avaliação é realizada correlacionando a saída do modelo e os valores de MOS dos ouvintes com audição normal que classificaram os sinais degradados por artefatos típicos de uma transmissão VoIP. Teve como resultados que a medida baseada em Gini e o MTD resultam em previsões muito semelhantes, porém com um custo computacional menor para a medida de Gini, para detectar as pausas e falhas na rede.

Em (XU; ZHANG, 2011) introduziu-se várias tecnologias de processamento de perda de pacotes usadas para chamadas VoIP para reduzir a distorção fonética causada pela perda de pacotes em uma transmissão VoIP ocasionando a perda da qualidade da voz na transmissão. Para isso, apresentou um esquema baseado na tecnologia interlace melhorada e tecnologia de correção de erros, o algoritmo de compensação de interpolação lagrangiana, que reduziu a distorção fonética causada pela perda de pacotes em uma transmissão VoIP, melhorando a capacidade de perda de quadros contínuos e resolvendo a recuperação e compensação de quadros perdidos, melhorando assim a qualidade de voz VoIP.

4 METODOLOGIA

Nesse capítulo será apresentado as ferramentas, métodos e etapas para a realização do projeto.

4.1 Ferramentas utilizadas

Keras: É uma API de alto nível para redes neurais, escrito em Python e usa como backend o TensorFlow, CNTK ou Theano. Ela foi projetada com foco no desenvolvimento rápido, sendo capaz de passar da ideia ao resultado com o menor esforço possível.

TensorFlow: É uma biblioteca de software de código aberto para computação numérica usando gráficos de fluxo de dados. Os nós no gráfico representam operações matemáticas, enquanto as bordas do gráfico representam os arrays de dados multidimensionais (tensores) comunicados entre eles. A arquitetura flexível permite implantar computação para uma ou mais CPUs ou GPUs em uma área de trabalho, servidor ou dispositivo móvel com uma única API.

LibROSA: É um pacote python para análise de música e áudio. Ele fornece os blocos de construção necessários para criar sistemas de recuperação de informações musicais.

Python: É uma linguagem de programação de alto nível, interpretada, de script, imperativa, orientada a objetos, funcional, de tipagem dinâmica e forte. Foi lançada por Guido van Rossum em 1991.

Dropbox: É um serviço para armazenamento e partilha de arquivos. É baseado no conceito de "computação em nuvem", e foi utilizado para armazenar as bases de dados do projeto e os algoritmos.

MATLAB R2017 (MATrix LABoratory): Trata-se de um software interativo de alta performance voltado para o cálculo numérico. O MATLAB integra análise numérica, cálculo com matrizes, processamento de sinais e construção de gráficos em ambiente fácil de usar onde problemas e soluções são expressos somente como eles são escritos matematicamente.

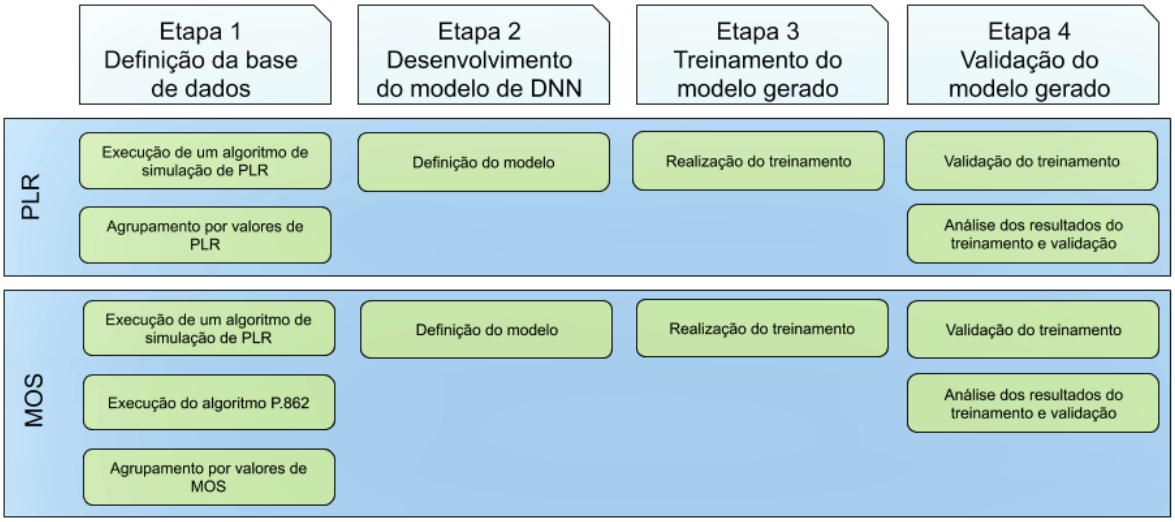
A linguagem de programação Python foi utilizada para o desenvolvimento do modelo Deep Learning pois o Python possui diversos pontos fortes para o fluxo de trabalho na Ciência de Dados e além disso é muito mais rápido e mais fácil de começar a usar em relação a outras linguagens disponíveis para trabalhos na Ciência de Dados.

A linguagem de programação MATLAB foi utilizada para desenvolvimento dos algoritmos de simulação de perda de pacotes e de execução automática dos algoritmos das recomendações ITU-T P.563 e P.862.

4.2 Fluxograma de desenvolvimento

Para a realização do trabalho foram realizadas diferentes etapas para chegar no resultado final, conforme Figura 4.1:

Figura 4.1 – Etapas dos procedimentos realizados na metodologia.



Fonte: Autor.

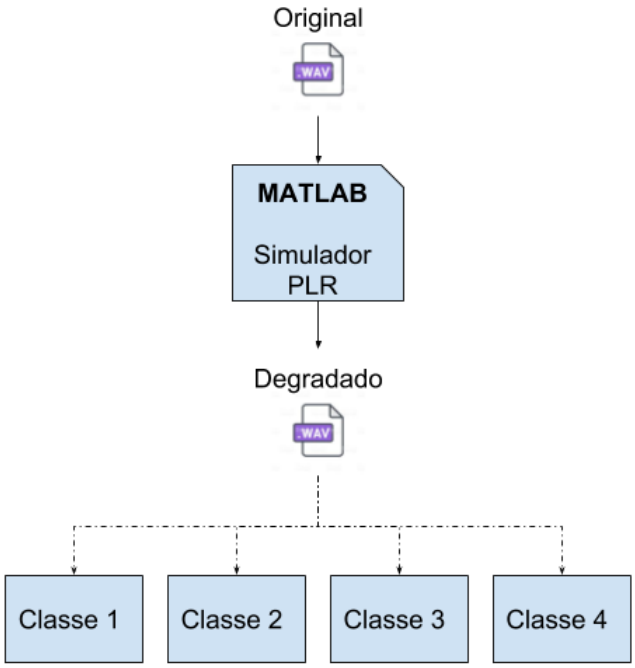
4.2.1 Etapa 1 - Definição da base de dados

A primeira etapa consiste em baixar os arquivos da recomendação da ITU-T P.501 que contém 28 arquivos de áudio no formato WAV com uma taxa de conversão de 16 kHz, onde contém dados de conversação em 7 idiomas diferentes, sendo que para cada idioma contém 2 arquivos com vozes masculinas e 2 arquivos com vozes femininas com uma duração total de 8 segundos.

Além disso, a base de dados contém arquivos de áudio que não possuem nenhuma degradação do sinal da voz e com isso foram utilizadas como referência para os algoritmos de classificação da qualidade da voz. Para isso, o projeto foi dividido em dois modelos, onde um classificou a qualidade da voz pela sua taxa de perda de pacotes, enquanto o outro modelo classificou a qualidade da voz através de seu índice MOS.

Com o objetivo de preparar a base de dados, foi desenvolvido um algoritmo no software MATLAB, que é capaz de simular diferentes taxas de perda de pacotes para um mesmo arquivo de áudio gerando assim um arquivo de áudio degradado correspondente. A Figura 4.2 mostra todo o processo realizado por esse algoritmo.

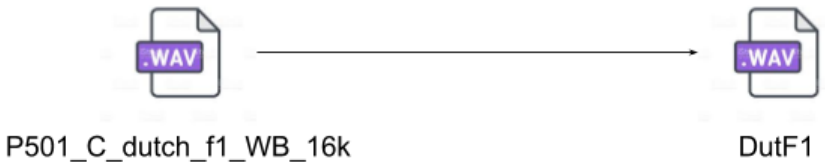
Figura 4.2 – Geração de sinais degradados de voz usando o fator de degradação PLR e classificação dos sinais degradados de voz.



Fonte: Autor.

Entretanto, para facilitar o processo de geração de arquivos de áudio e automatização de todo o processo, foi necessário realizar uma padronização do nome do arquivo original para que tenha 9 caracteres no total:

Figura 4.3 – Padronização do nome do arquivo etapa 1.



Fonte: Autor.

A etapa 1 representada pela Figura 4.3, consiste em renomear cada arquivo de áudio, para um padrão contendo a definição do idioma no arquivo e qual o sexo da pessoa.

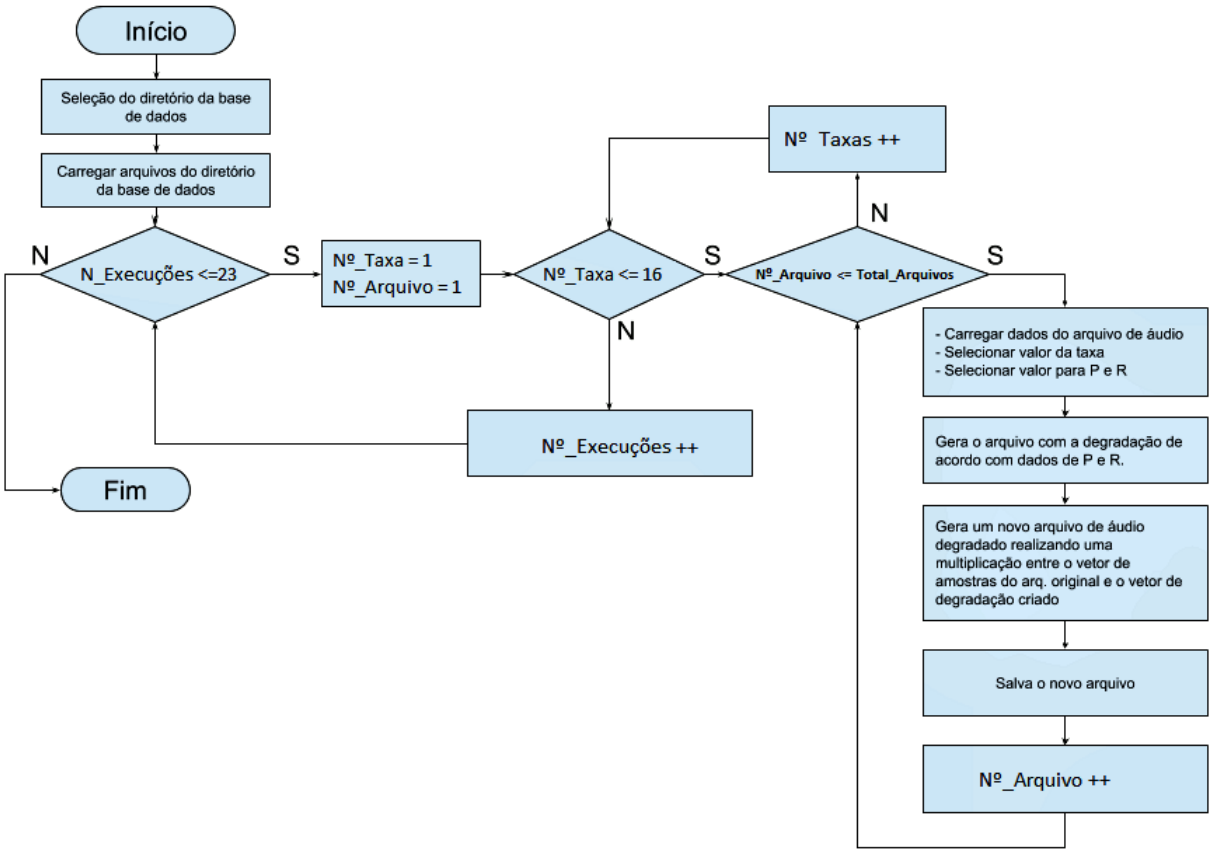
Figura 4.4 – Padronização do nome do arquivo etapa 2.

Fonte: Autor.

Na etapa 2 representada pela Figura 4.4, o padrão foi acrescido de um contador de iteração de execução do algoritmo e também do percentual de perda de pacotes, onde ”1” é a iteração do controle do algoritmo de simulação de PLR, ”Dut” representa o idioma presente no arquivo, o ”F1” representa o sexo da pessoa no áudio e o índice sendo que na base de dados para cada idioma contém 2 arquivos com vozes masculinas e 2 arquivos com vozes femininas, e o ”005” representa a taxa de perda de pacotes, onde 005 equivale a 0.5%.

A Figura 4.5 representa um fluxograma detalhado do processo de geração de arquivos degradados, onde as variáveis Nº_Execuções é o controle de quantas vezes o algoritmo será executado, a Nº_Taxa é a responsável por selecionar qual a taxa de PLR será utilizada para a degradação na execução, a variável Nº_Arquivos representa o arquivo selecionado na base dados original que irá sofrer a degradação pela taxa de perda de pacotes e a Total_Arquivos representa o total de arquivos contido na base de dados original da recomendação selecionada.

Figura 4.5 – Fluxograma completo de geração de arquivos de áudio degradado.



Fonte: Autor.

O Algoritmo 1 é a representação em pseudocódigo do algoritmo de geração de arquivos degradado do sinal de voz representado na Figura 4.5.

Algoritmo 1 Algoritmo para geração de arquivos de voz degradados

Entrada: Diretório com a base de dados
Saída: Arquivos de áudio degradado
início
1 - Carregar arquivos do diretório da base de dados;
2 - Inicializar: Vetor_de_Taxas com as 16 taxas definidas de PLR;
3 - Inicializar: Total_de_Arquivos, contando o número de arquivos na pasta original;
4 - Inicializar: Nº_da_Taxa = 1;
5 - Inicializar: Nº_de_Execução = 1;
repita
if *Nº_de_Execução <= 23* **then**
1 - Inicializar: Nº_da_Taxa = 1;
2 - Inicializar: Nº_do_Arquivo = 1;
if *Nº_da_Taxa <= 16* **then**
if *Nº_do_Arquivo <= Total_de_Arquivos* **then**
1 - Carregar dados do arquivo de áudio [Nº_do_Arquivo];
2 - Selecionar o valor da taxa no Vetor_de_Taxas[Nº_da_Taxa];
3 - Gerar o valor de perda de pacote, onde é definido pela Equação PLR = 1-r/(p+r). Onde P é a probabilidade de ir de um valor bom para um valor ruim e R é a probabilidade de ir de um valor ruim para um bom, podendo variar entre os números presentes no intervalo de 0.0000 a 1.000.;
4 - Gerar um arquivo texto de degradação contendo 0 e 1, onde 0 é a perda do sinal e 1 é a continuidade do sinal de voz.;
5 - Gerar um novo arquivo degradado da voz realizando a multiplicação do arquivo de degradação com o arquivo de áudio original gerando assim um novo arquivo degradado de voz.;
6 - Salvar o arquivo degradado no diretório da base de dados.;
7 - Incrementar o Nº_do_Arquivo.;
else
1 - Incrementar o Nº_da_Taxa.;
end
else
1 - Incrementar o Nº_de_Execução;
end
else
Para a execução do algoritmo;
end
até *até Nº_de_Execução = 23*;
fim

4.2.1.1 Determinação das classes de PLR

Para o modelo de classificação por perda de pacotes, foi realizado após o processamento de arquivos, a separação dos arquivos em quatro classes distintas de acordo com os valores de PLR. Onde as classes foram separadas da seguinte forma:

- Classe 1: PLR(0.5% + 1.0% + 1.5% + 2.0%) com um total de 1840 arquivos degradados;
- Classe 2: PLR(2.5% + 3.5% + 4.5% + 5.5%) com um total de 1840 arquivos degradados;
- Classe 3: PLR(7.0% + 8.0% + 9.0% + 10.0%) com um total de 1840 arquivos degradados;
- Classe 4: PLR(12.0% + 15.0% + 18.0% + 21.0%) com um total de 1840 arquivos degradados;

Para gerar cada classe, foi executado o seguinte procedimento, onde para cada um dos 20 arquivos de áudio da base original foram realizadas o processamento com as 16 taxas de perda de pacote resultando num total de 320 arquivos para cada execução do algoritmo, porém para aumentar o número de arquivos e a diversificação entre eles o algoritmo foi executado diversas vezes resultando em um total de 7360 arquivos de áudio degradados, e como cada classe possui o mesmo número de arquivos degradados, tem-se que cada classe é formada por 4 taxas de PLR diferentes, resultando em 1840 por classe.

4.2.1.2 Determinação das classes de MOS

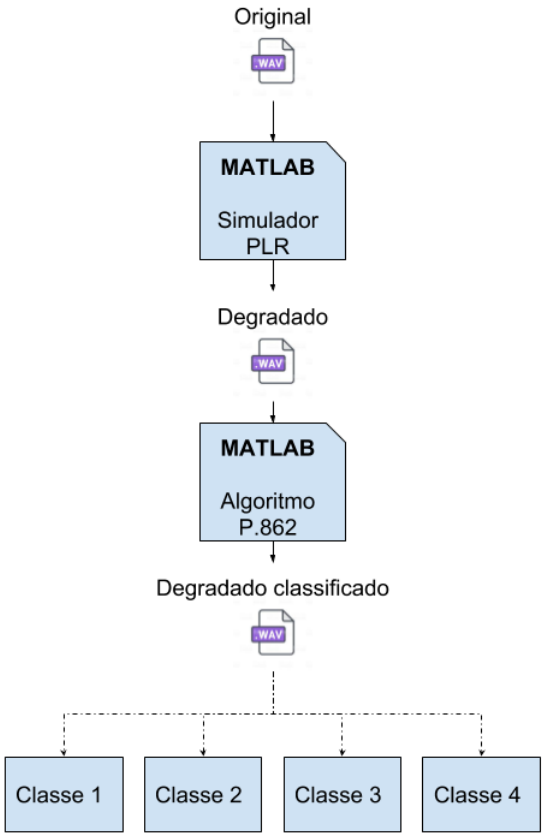
Para o modelo de classificação por índice MOS, foi executado após o processamento dos arquivos um novo algoritmo, também desenvolvido no MATLAB, capaz de automatizar a execução do algoritmo P.862 para todos os arquivos de áudio degradado, comparando-os com seus respectivos arquivos de áudio original, tendo como resultado o índice MOS referente à taxa de perda de pacotes encontrada no arquivo de áudio degradado.

Após adquirir estes dados, os arquivos foram separados em 5 classes de acordo com seu índice MOS, utilizando como referência a tabela do Fator-R do E-Model na sessão 2.4.4.

Porém, foi detectado que não foram gerados arquivos com índice MOS para a classe 5 onde o índice MOS está entre 4,3 e 4,5, com isso foi adotado que os arquivos com índices maiores que 4,0 seriam somente uma classe, assim resultando em 4 classes para os índices encontrados conforme a Figura 4.6.

- Classe 1: MOS(1,00 a 3,09) com um total de 798 arquivos degradados;
- Classe 2: MOS(3,10 a 3,59) com um total de 3106 arquivos degradados;
- Classe 3: MOS(3,60 a 3,99) com um total de 2838 arquivos degradados;
- Classe 4: MOS(4,00 a 4,50) com um total de 618 arquivos degradados;

Figura 4.6 – Obtenção de valores MOS dos sinais degradados de voz usando o algoritmo da recomendação ITU-T P.862 e classificação dos valores MOS dos sinais degradados de voz.



Fonte: Autor.

4.2.1.3 Determinação da base de dados de arquivos adicionais

Para uma validação extra do modelo foram preparadas duas bases de dados adicionais, com arquivos da base de dados da ”VoxCeleb”, onde foram realizados a simulação de PLR em cada um e depois obtido o índice MOS de cada arquivo, e finalmente foram selecionados aleatoriamente 1000 arquivos para cada base.

O VoxCeleb é um conjunto de dados áudio visuais que consiste em pequenos trechos de fala humana, extraídos de vídeos de entrevistas enviados para o YouTube, ela contém 100.000

arquivos de áudio no formato WAV com mais de 1.251 pessoas abrangendo diferentes etnias, sotaques, profissões e idades. Como os dados são extraídos de vídeos do YouTube, eles não possuem uma boa qualidade visto que os vídeos não são criados em ambientes preparados, com isso, a acurácia do modelo pode ser um pouco reduzida.

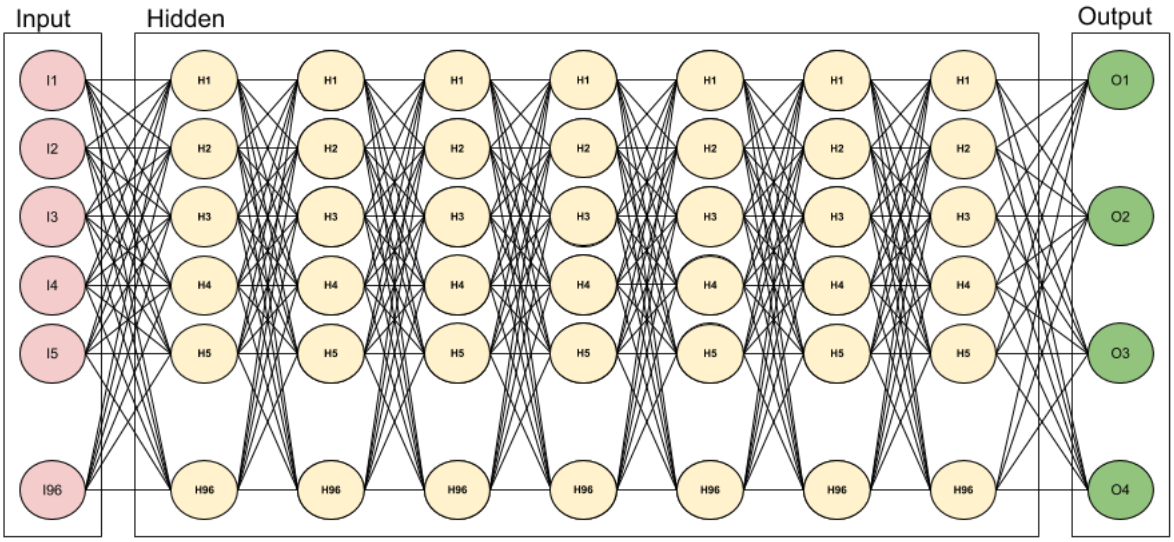
Para preparar a base de dados pela taxa de PLR foram selecionados 1000 arquivos aleatoriamente com as mesmas taxas de PLR utilizadas na subseção 4.2.1.1 resultando em 4 classes com 250 arquivos em cada. Para preparar a base de dados pelo MOS foi utilizado o mesmo processo que foi utilizado na subseção 4.2.1.2, porém as classes contém a mesma quantidade de arquivos.

4.2.2 Etapa 2 - Modelo proposto

Nesta sessão do trabalho, será apresentada a arquitetura da Deep Learning utilizada no treinamento e validação do projeto.

A partir do processamento das características dos arquivos na entrada da rede neural, foi definido um modelo da rede neural, onde a figura 4.7 representa a arquitetura do modelo de Deep Learning gerada:

Figura 4.7 – Arquitetura do modelo de Deep Learning.



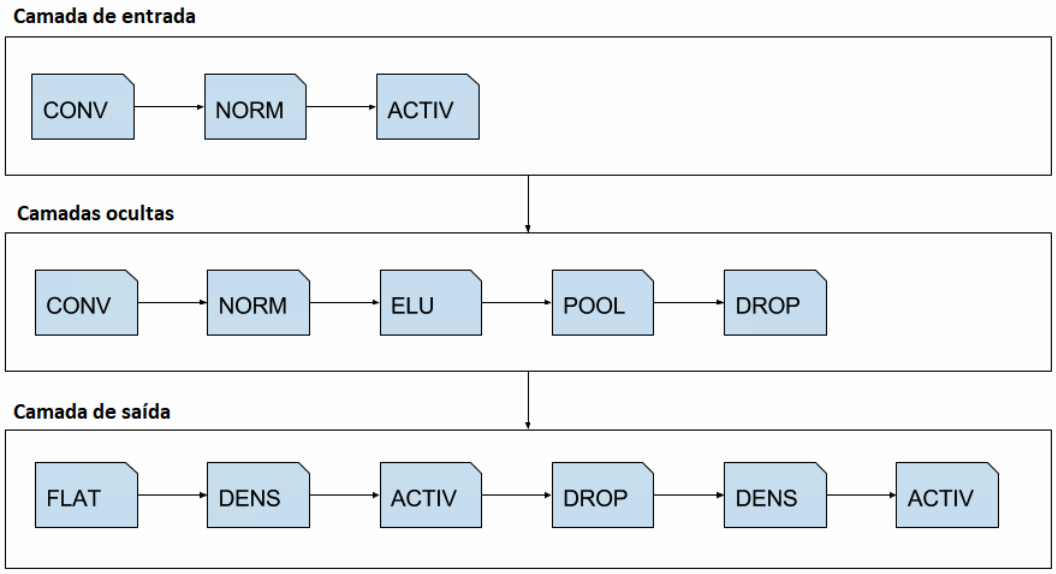
Fonte: Autor.

O modelo gerado possui as seguintes características:

- Possui uma camada de entrada (I) com 96 neurônios, onde 96 é o número de características que serão utilizadas na rede, definidas automaticamente de acordo com a saída de parâmetros da biblioteca LibROSA.
- Possui sete camadas ocultas (H) com 96 neurônios cada, ligadas de forma recorrente de acordo com as RNN's.
- Possui uma camada de saída (O) com 4 neurônios, onde cada um representa uma saída de uma classe.

Estrutura interna do modelo Cada camada definida no modelo possui as seguintes características e funções que são representadas pela Figura 4.8.

Figura 4.8 – Estrutura interna do modelo de Deep Learning.



Fonte: Elaborada pelo autor.

A Figura 4.8 representa a estrutura interna de cada camada da rede, onde cada camada é responsável por realizar as seguintes funções:

1 - Camada de entrada (Input Layer): A camada de entrada é responsável por realizar a entrada do arquivo de áudio na rede neural e com isso realizar uma leitura do número de características que a biblioteca Librosa pode analisar, definido pela Tabela 2.2 na sessão 2.3. Ela é utilizada para fazer o preprocessamento do arquivo de entrada e utilizar os dados do preprocessamento como entrada nos neurônios na camada escondida.

2 - Camada oculta (Hidden Layer): As camadas ocultas são responsáveis por analisar as características extraídas do arquivo de áudio para que seja possível realizar o reconhecimento e classificação das suas características representadas na Tabela 2.3 localizada na sessão 2.3.

3 - Camada de saída (Output Layer): A camada de saída é responsável por conectar todas as características e gerar o modelo final de classificação do sinal de áudio da rede.

Na Figura 4.8, cada etapa de execução é definida pelos seguintes itens:

- **Conv:** é a realização de uma convolução;
- **Norm:** é a normalização dos dados;
- **Activ:** é a realização de uma função de ativação;
- **Elu:** é a realização de uma função de ativação Elu;
- **Pool:** é a realização da função Pooling para subdividir a imagem;
- **Drop:** é a chamada da função de Dropout, onde é desligado o percentual de neurônios da rede;
- **Flat:** é a realização da operação de Flatten, onde organiza os blocos separados pela operação de Pooling em um vetor linear;
- **Dens:** é a operação responsável por reduzir o número de neurônios agrupando os neurônios responsáveis por representar cada classe.

4.2.3 Etapa 3 - Testes

Nessa seção será apresentado como foi realizado o procedimento para os testes dos modelos avaliados pela taxa de perda de pacotes e pelo índice MOS de cada arquivo executado na simulação para a realização do treinamento do modelo.

Para a realização dos testes foram utilizados 80% dos arquivos de áudios para o treinamento e 20% para validação dos modelos onde a separação dos arquivos para treinamento e validação foi aleatória a cada execução de época do algoritmo e foram definidos conforme Tabelas 4.3 e 4.4.

Tabela 4.1 – Separação das classes para treinamento e validação por valores de PLR

Classes	**Taxa de PLR (%)**	**Nº de Arquivos**	**Total**	**Utilização**
1	(0.5),(1.0),(1.5),(2.0)	1840	7360	Treinamento 80% 5888
2	(2.5),(3.5),(4.5),(5.5)	1840		
3	(7.0),(8.0),(9.0),(10.0)	1840		Validação 20% 1472
4	(12.0),(15.0),(18.0),(21.0)	1840		

Fonte: Autor.

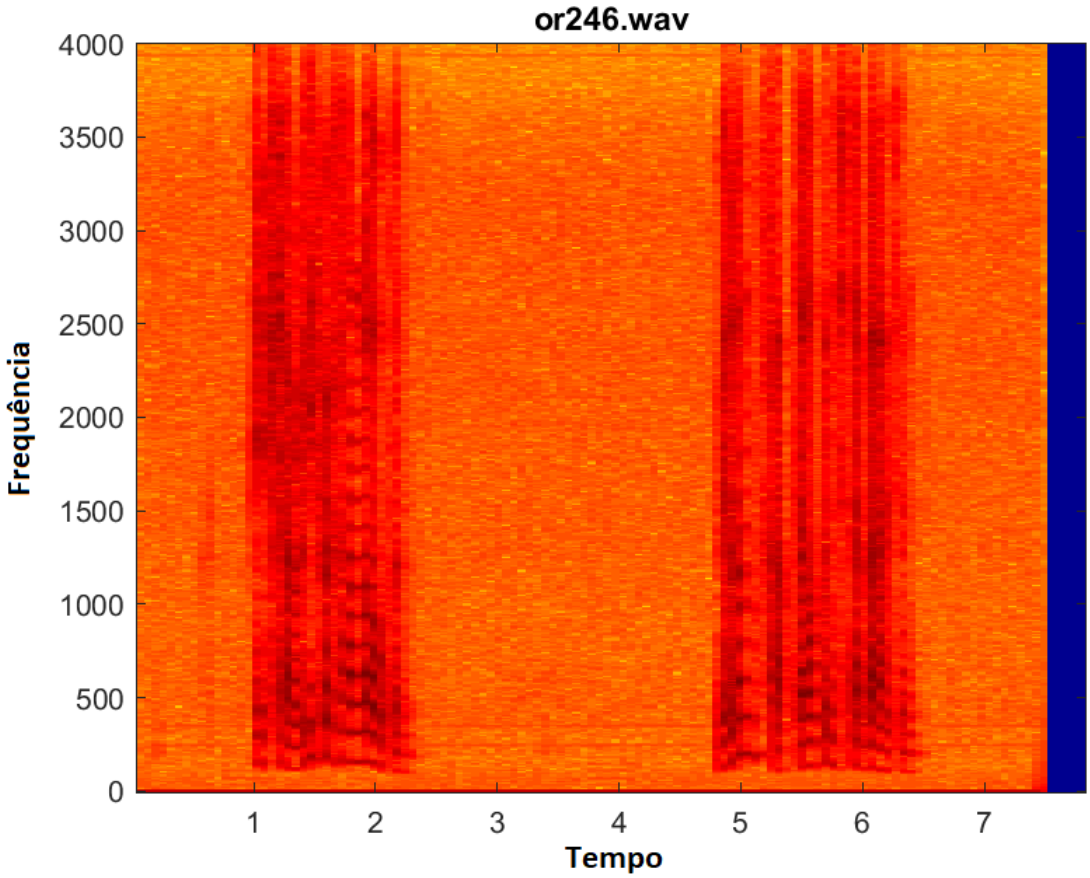
Tabela 4.2 – Separação das classes para treinamento e validação por índice MOS

Classes	**Indice MOS**	**Nº de Arquivos**	**Total**	**Utilização**
1	1.00 à 3.09	798	7360	Treinamento 80% 5888
2	3.10 à 3.59	3106		
3	3.60 à 3.99	2838		Validação 20% 1472
4	4.00 à 4.50	618		

Fonte: Autor.

Pode-se verificar que na Tabela 4.3, a base de dados utilizada para o treinamento e validação está com suas classes balanceadas enquanto na Tabela 4.4 as classes estão totalmente desbalanceadas. As classes estão desbalanceadas devido ao resultado da execução do algoritmo da recomendação ITU-T P.862, que ao analisar os arquivos de áudio degradado comparando com os seus respectivos arquivos de áudio original, e como o algoritmo de simulação de perda de pacotes é executado de forma aleatória, a degradação gerada no arquivo pode ou não ser inserida em período de silêncio da conversação. Quando a degradação é inserida nos períodos de silêncio o índice MOS consequentemente é elevado, e quando a degradação é inserida em períodos normais com sinal de voz, o índice MOS é reduzido. Por isso gera-se o desbalanceamento das classes.

Figura 4.9 – Espectrograma do arquivo de áudio original.

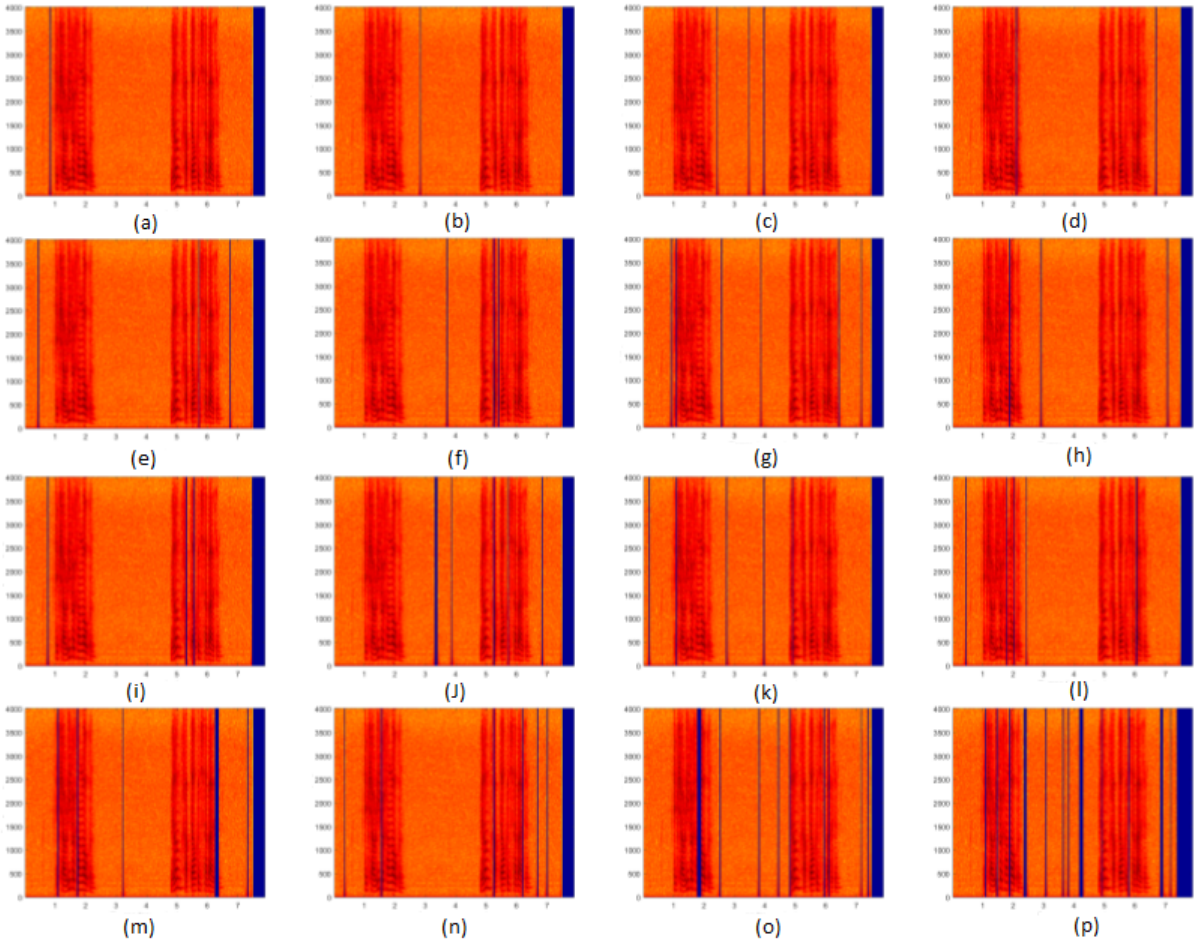


Fonte: Autor.

A Figura 4.9 representa um espectrograma de um arquivo original, onde pode-se verificar que o sinal da voz não sofreu nenhum tipo de degradação, onde a degradação é representada pela cor azul. Porém, a cor azul não representa somente a perda do sinal da voz, ela também representa o silêncio na conversação.

Porém, como visto anteriormente na Sessão 2.4.2, o algoritmo da recomendação ITU-T P.862 possui a capacidade de diferenciar os silêncios e as perdas no sinal da voz, entretanto o algoritmo da recomendação ITU-T P.563 visto na Sessão 2.4.3 não possui esta capacidade.

Figura 4.10 – Espectrogramas dos arquivos de áudio degradados para os valores de PLR. Onde as letras de ”a” a ”p” representam respectivamente os valores de PLR 0.5%, 1.0%, 1.5%, 2.0%, 2.5%, 3.5%, 4.5%, 5.5%, 7.0%, 8.0%, 9.0%, 10.0%, 12.0%, 15.0%, 18.0% e 21.0%.



Fonte: Autor.

A Figura 4.10 representa os espectrogramas gerados para o mesmo arquivo de áudio representado pela Figura 4.9, porém como podemos ver a degradação ocorrida pelas taxas de perdas de pacotes definidas, representaram uma degradação do sinal da voz para as diferentes taxas, e a medida que o percentual de perda de pacote aumenta podemos verificar que ocorre um aumento dos silêncios no sinal da voz.

5 RESULTADOS

Neste capitulo serão apresentados os resultados obtidos a partir dos testes realizados no Capítulo 4.

Para a geração dos resultados cada modelo foi treinado e validado 10 vezes e os resultados da sessão 5.1 e 5.2 foram gerados através da média dos resultados obtidos em cada execução.

Para realizar as etapas de treinamento e validação foi utilizado um computador com um processador I5 de 2.3 Ghz, 8GB de memória RAM e uma GPU GTX 1050 TI com 4GB de memória de vídeo. A etapa de treinamento foi realizada em 8 horas para 1000 épocas e a validação foi realizada em 3 minutos.

5.1 Resultados utilizando o modelo PLR

Os resultados obtidos nos testes com a base de dados preparada com os valores de PLR foram conseguidos através da validação do modelo utilizando 20% do total de arquivos da base de dados, ou seja, 1472 arquivos sendo 368 arquivos para cada classe definida. Os resultados são mostrados em matrizes de confusão geradas para os números de épocas 10, 50, 100, 500 e 1000, e são apresentados nas Tabelas 5.1, 5.2, 5.3, 5.4 e 5.5.

Tabela 5.1 – Resultado da matriz de confusão para 10 épocas para valores de PLR

	Classe 1	Classe 2	Classe 3	Classe 4
Classe 1	351	17	-	-
Classe 2	11	317	40	-
Classe 3	-	50	278	40
Classe 4	-	-	72	296

Fonte: Autor.

A Tabela 5.1 apresenta os dados de acertos e erros dos testes realizados para 10 épocas de treinamento, através dela podemos perceber que na validação do modelo treinado com apenas 10 épocas obteve-se um total de 230 erros de classificação de classes e um total de 1242 acertos na validação, obtendo uma acurácia de 84,375% na validação do modelo.

Tabela 5.2 – Resultado da matriz de confusão para 50 épocas para valores de PLR

	Classe 1	Classe 2	Classe 3	Classe 4
Classe 1	353	15	-	-
Classe 2	5	328	35	-
Classe 3	-	30	306	32
Classe 4	-	-	58	310

Fonte: Autor.

A Tabela 5.2 apresenta os dados de acertos e erros dos testes realizados para 50 épocas de treinamento, através dela podemos perceber que na validação do modelo treinado com 50 épocas obteve-se um total de 175 erros de classificação de classes e um total de 1297 acertos na validação, obtendo uma acurácia de 88,111%.

Tabela 5.3 – Resultado da matriz de confusão para 100 épocas para valores de PLR

	Classe 1	Classe 2	Classe 3	Classe 4
Classe 1	356	12	-	-
Classe 2	10	319	39	-
Classe 3	-	20	321	27
Classe 4	-	-	49	319

Fonte: Autor.

A Tabela 5.3 apresenta os dados de acertos e erros dos testes realizados para 100 épocas de treinamento, através dela podemos perceber que na validação do modelo treinado com 100 épocas obteve-se um total de 157 erros de classificação de classes e um total de 1315 acertos na validação, obtendo uma acurácia de 89,334%.

Tabela 5.4 – Resultado da matriz de confusão para 500 épocas para valores de PLR

	Classe 1	Classe 2	Classe 3	Classe 4
Classe 1	359	9	-	-
Classe 2	11	325	32	-
Classe 3	-	21	319	28
Classe 4	-	-	48	320

Fonte: Autor.

A Tabela 5.4 apresenta os dados de acertos e erros dos testes realizados para 500 épocas de treinamento, através dela podemos perceber que na validação do modelo treinado com 500 épocas obteve-se um total de 149 erros de classificação de classes e um total de 1323 acertos na validação, obtendo uma acurácia de 89,877%.

Tabela 5.5 – Resultado da matriz de confusão para 1000 épocas para valores de PLR

	Classe 1	Classe 2	Classe 3	Classe 4
Classe 1	357	7	-	-
Classe 2	6	332	17	-
Classe 3	-	14	321	22
Classe 4	-	-	23	329

Fonte: Autor.

A Tabela 5.5 apresenta os dados de acertos e erros dos testes realizados para 1000 épocas de treinamento, através dela podemos perceber que na validação do modelo treinado com 1000 épocas obteve-se um total de 89 erros de classificação de classes e um total de 1383 acertos na validação, obtendo uma acurácia de 93,954%.

Foram realizados testes com a base de dados de arquivos adicionais preparada pela taxa de PLR para verificar a acurácia do modelo de PLR, onde o mesmo obteve aproximadamente 86,958% de acurácia na determinação das classes após o treinamento do modelo com 1000 épocas.

5.2 Resultados utilizando o modelo MOS

Os resultados obtidos nos testes com a base de dados preparada com os valores do índice MOS foram obtidos através da validação do modelo utilizando 20% do total de arquivos da base de dados, ou seja, 1472 arquivos, porém como a base de dados para índice MOS não está balanceada, o número de arquivos de cada classe na validação é representado pela Tabela 5.6.

Tabela 5.6 – Cálculo de número de arquivos em cada classe para validação

Total de arquivos por classe	Percentual no total de arquivos	Total de arquivos na validação
798	10,84%	160
3106	42,20%	621
2838	38,56%	568
618	8,40%	124

Fonte: Autor.

O número de arquivos de cada classe foram calculados aproximadamente, de acordo com o seu percentual no número total de arquivos, conforme a Tabela 5.6, ao verificar cada classe obtivemos seu percentual de representação no total de arquivos da base de dados, assim, conseguimos determinar um número aproximado de arquivos na base utilizada para a validação.

Os resultados são mostrados em matrizes de confusão geradas para os números de épocas 10, 50, 100, 500 e 1000, e são apresentados nas Tabelas 5.7, 5.8, 5.9, 5.10 e 5.11.

Tabela 5.7 – Resultado da matriz de confusão para 10 épocas para valores de MOS

	Classe 1	Classe 2	Classe 3	Classe 4
Classe 1	148	12	-	-
Classe 2	9	587	25	-
Classe 3	-	42	507	19
Classe 4	-	-	52	72

Fonte: Autor.

A Tabela 5.7 apresenta os dados de acertos e erros dos testes realizados para 10 épocas de treinamento, através dela podemos perceber que na validação do modelo treinado com 10 épocas obteve-se um total de 159 erros de classificação de classes e um total de 1313 acertos na validação, obtendo uma acurácia de 89,198%.

Tabela 5.8 – Resultado da matriz de confusão para 50 épocas para valores de MOS

	Classe 1	Classe 2	Classe 3	Classe 4
Classe 1	150	10	-	-
Classe 2	12	560	49	-
Classe 3	-	25	525	18
Classe 4	-	-	47	77

Fonte: Autor.

A Tabela 5.8 apresenta os dados de acertos e erros dos testes realizados para 50 épocas de treinamento, através dela podemos perceber que na validação do modelo treinado com 50 épocas obteve-se um total de 161 erros de classificação de classes e um total de 1311 acertos de classificação, obtendo uma acurácia de 89,0625%.

Tabela 5.9 – Resultado da matriz de confusão para 100 épocas para valores de MOS

	Classe 1	Classe 2	Classe 3	Classe 4
Classe 1	143	17	-	-
Classe 2	13	565	43	-
Classe 3	-	20	517	31
Classe 4	-	-	49	75

Fonte: Autor.

A Tabela 5.9 apresenta os dados de acertos e erros dos testes realizados para 100 épocas de treinamento, através dela podemos perceber que na validação do modelo treinado com 100

épocas obteve-se um total de 173 erros de classificação de classes e um total de 1299 acertos de classificação, obtendo uma acurácia de 88,247%.

Tabela 5.10 – Resultado da matriz de confusão para 500 épocas para valores de MOS

	Classe 1	Classe 2	Classe 3	Classe 4
Classe 1	148	12	-	-
Classe 2	11	580	30	-
Classe 3	-	15	521	32
Classe 4	-	-	42	82

Fonte: Autor.

A Tabela 5.10 apresenta os dados de acertos e erros dos testes realizados para 500 épocas de treinamento, através dela podemos perceber que na validação do modelo treinado com 500 épocas obteve-se um total de 142 erros de classificação de classes e um total de 1330 acertos de classificação, obtendo uma acurácia de 90,353%.

Tabela 5.11 – Resultado da matriz de confusão para 1000 épocas para valores de MOS

	Classe 1	Classe 2	Classe 3	Classe 4
Classe 1	149	11	-	-
Classe 2	9	585	27	-
Classe 3	-	17	521	30
Classe 4	-	-	39	85

Fonte: Autor.

A Tabela 5.11 apresenta os dados de acertos e erros dos testes realizados para 1000 épocas de treinamento, através dela podemos perceber que na validação do modelo treinado com 1000 épocas obteve-se um total de 133 erros de classificação de classes e um total de 1339 acertos de classificação, obtendo uma acurácia de 90,965%.

Também foram realizados testes com a base de dados de arquivos adicionais preparadas por índice MOS para verificar a acurácia do modelo MOS, onde o mesmo obteve aproximadamente 83,285% de acurácia na determinação das classes após o treinamento do modelo com 1000 épocas.

5.3 Comparação dos algoritmos

Nessa sessão será apresentado a comparação dos resultados obtidos a partir dos testes realizados com os algoritmos P.563 e P.862.

Tabela 5.12 – Resultados do processamento do algoritmo da recomendação ITU-T P.563 para a base de dados da recomendação ITU-T P.862 para as taxas de PLR 0.5, 1.0, 1.5, 2.0, 2.5, 3.5, 4.5 e 5.5%.

Índice MOS para as taxas de PLR obtidos pelo algoritmo P.563								
Arquivo	**0,50%**	**1,00%**	**1,50%**	**2,00%**	**2,50%**	**3,50%**	**4,50%**	**5,50%**
or105.wav	2,554	2,702	1,059	2,344	2,470	2,270	1,970	2,104
or109.wav	2,627	2,460	2,383	2,190	2,034	2,017	1,668	1,679
or114.wav	3,222	2,857	2,682	2,420	2,346	2,158	1,829	1,953
or129.wav	2,119	2,131	1,909	1,753	1,921	1,288	1,686	1,324
or134.wav	1,992	1,965	1,764	1,585	1,462	1,428	1,112	1,480
or137.wav	2,174	1,967	1,626	1,849	1,942	1,466	1,629	1,528
or145.wav	3,299	3,054	3,059	2,873	2,088	2,693	2,218	2,161
or149.wav	2,469	2,499	2,268	2,254	2,010	1,882	1,928	1,970
or152.wav	3,028	2,691	2,086	2,523	2,510	2,102	2,106	2,160
or154.wav	2,108	1,867	1,869	1,749	1,658	1,571	1,700	1,459
or155.wav	2,694	2,494	2,371	2,103	2,058	2,270	1,877	2,004
or161.wav	2,732	2,600	2,542	2,447	2,233	2,155	2,017	1,950
or164.wav	2,614	2,590	2,585	2,319	2,395	2,363	1,982	1,857
or166.wav	2,532	2,395	2,394	2,329	2,183	2,146	2,144	1,891
or170.wav	2,444	2,302	2,373	1,930	2,182	2,107	1,905	2,008
or179.wav	2,636	1,632	1,667	2,046	1,706	1,457	1,772	1,627
or221.wav	2,570	2,282	1,772	2,293	1,987	1,801	1,000	1,589
or229.wav	2,571	2,386	2,367	2,279	2,195	2,212	2,117	1,960
or246.wav	2,633	2,464	2,202	2,265	2,091	2,173	2,113	1,989
or272.wav	2,386	2,150	2,161	2,143	2,056	2,022	1,912	1,764

Fonte: Autor.

Tabela 5.13 – Resultados do processamento do algoritmo da recomendação ITU-T P.563 para a base de dados da recomendação ITU-T P.862 para as taxas de PLR 7.0, 8.0, 9.0, 10.0, 12.0, 15.0, 18.0 e 21.0%.

Índice MOS para as taxas de PLR obtidos pelo algoritmo P.563								
Arquivo	**7,00%**	**8,00%**	**9,00%**	**10,00%**	**12,00%**	**15,00%**	**18,00%**	**21,00%**
or105.wav	1,398	1,989	1,896	1,834	1,836	1,394	1,497	1,450
or109.wav	1,544	1,329	1,389	1,203	1,252	1,105	1,034	1,052
or114.wav	1,826	1,826	1,483	1,662	1,540	1,456	1,273	1,102
or129.wav	1,345	1,447	1,293	1,227	1,408	1,192	1,043	1,000
or134.wav	1,108	1,072	1,032	1,361	1,000	1,000	1,000	1,000
or137.wav	1,414	1,271	1,406	1,112	1,225	1,126	1,000	1,000
or145.wav	2,160	1,899	1,572	1,734	1,832	1,364	1,584	1,392
or149.wav	1,690	1,671	1,569	1,650	1,517	1,327	1,334	1,000
or152.wav	1,842	1,943	1,407	1,609	1,684	1,574	1,149	1,000
or154.wav	1,455	1,325	1,452	1,303	1,320	1,080	1,077	1,027
or155.wav	1,773	1,749	1,626	1,522	1,401	1,362	1,237	1,045
or161.wav	1,843	1,674	1,750	1,706	1,490	1,435	1,369	1,159
or164.wav	2,170	1,986	1,780	1,891	1,502	1,538	1,450	1,095
or166.wav	2,006	1,901	1,730	1,772	1,610	1,672	1,486	1,401
or170.wav	1,880	1,888	1,583	1,832	1,565	1,661	1,564	1,199
or179.wav	1,515	1,406	1,502	1,304	1,000	1,000	1,000	1,000
or221.wav	1,562	1,405	1,498	1,445	1,288	1,282	1,000	1,019
or229.wav	1,949	1,845	1,836	1,770	1,727	1,661	1,420	1,336
or246.wav	1,735	1,744	1,878	1,680	1,487	1,527	1,419	1,221
or272.wav	1,733	1,715	1,565	1,494	1,474	1,481	1,330	1,264

Fonte: Autor.

Tabela 5.14 – Primeira parte dos resultados do processamento do algoritmo da recomendação ITU-T P.563 para a base de dados da recomendação ITU-T P.501.

Índice MOS para as taxas de PLR obtidos pelo algoritmo P.563								
Arquivo	**0,50%**	**1,00%**	**1,50%**	**2,00%**	**2,50%**	**3,50%**	**4,50%**	**5,50%**
DutF1.wav	1,578	1,585	1,513	1,531	1,634	1,345	1,174	1,184
DutF2.wav	2,280	2,036	1,971	1,961	1,662	1,656	1,353	1,382
DutM1.wav	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
DutM2.wav	1,000	1,061	1,000	1,000	1,000	1,000	1,000	1,000
EngF1.wav	1,790	1,189	1,472	1,517	1,000	1,000	1,369	1,362
EngF2.wav	2,168	2,173	2,129	2,211	1,894	1,996	1,933	1,955
EngM1.wav	1,560	1,691	1,281	1,148	1,500	1,149	1,074	1,110
EngM2.wav	1,282	1,245	1,200	1,172	1,087	1,079	1,000	1,000
FinF1.wav	2,453	2,198	2,248	1,708	1,958	1,778	1,300	1,669
FinF2.wav	1,477	1,351	1,340	1,211	1,172	1,191	1,000	1,000
FinM1.wav	1,717	1,638	1,340	1,324	1,188	1,262	1,085	1,000
FinM2.wav	1,176	1,161	1,145	1,115	1,000	1,011	1,000	1,000
FreF1.wav	1,651	1,423	1,566	1,526	1,418	1,510	1,338	1,451
FreF2.wav	2,627	2,437	2,498	2,032	2,301	2,045	2,158	1,895
FreM1.wav	1,360	1,475	1,058	1,143	1,000	1,063	1,000	1,072
FreM2.wav	1,320	1,406	1,360	1,229	1,177	1,187	1,026	1,006
GerF1.wav	1,733	1,823	1,733	1,735	1,683	1,446	1,074	1,153
GerF2.wav	2,407	2,081	1,986	1,887	1,727	1,678	1,519	1,672
GerM1.wav	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
GerM2.wav	1,406	1,478	1,064	1,312	1,212	1,000	1,000	1,012
ItaF1.wav	2,556	2,424	2,136	2,302	2,387	1,917	2,027	1,691
ItaF2.wav	2,600	2,444	2,183	2,226	1,892	1,965	2,015	1,829
ItaM1.wav	1,273	1,268	1,282	1,043	1,000	1,058	1,000	1,000
ItaM2.wav	1,954	1,650	1,452	1,405	1,367	1,273	1,209	1,000
JapF1.wav	3,295	2,975	2,893	2,783	2,648	2,350	2,261	2,054
JapF2.wav	3,107	2,920	2,797	2,608	2,524	2,391	2,219	1,917
JapM1.wav	2,600	2,288	2,183	2,129	1,889	1,989	1,877	1,853
JapM2.wav	1,463	1,394	1,418	1,134	1,079	1,095	1,134	1,091

Fonte: Autor.

Tabela 5.15 – Segunda parte dos resultados do processamento do algoritmo da recomendação ITU-T P.563 para a base de dados da recomendação ITU-T P.501.

Índice MOS para as taxas de PLR obtidos pelo algoritmo P.563								
Arquivo	**7,00%**	**8,00%**	**9,00%**	**10,00%**	**12,00%**	**15,00%**	**18,00%**	**21,00%**
DutF1.wav	1,208	1,130	1,000	1,000	1,064	1,000	1,000	1,000
DutF2.wav	1,531	1,261	1,269	1,134	1,216	1,184	1,071	1,000
DutM1.wav	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
DutM2.wav	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
EngF1.wav	1,268	1,031	1,263	1,206	1,125	1,222	1,000	1,098
EngF2.wav	1,713	1,719	1,600	1,556	1,498	1,632	1,329	1,174
EngM1.wav	1,121	1,032	1,117	1,000	1,000	1,000	1,000	1,000
EngM2.wav	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
FinF1.wav	1,318	1,239	1,110	1,187	1,106	1,000	1,000	1,000
FinF2.wav	1,000	1,000	1,000	1,014	1,000	1,000	1,000	1,000
FinM1.wav	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
FinM2.wav	1,003	1,000	1,000	1,000	1,000	1,000	1,000	1,000
FreF1.wav	1,331	1,318	1,328	1,206	1,176	1,062	1,000	1,000
FreF2.wav	1,835	1,652	1,635	1,730	1,539	1,338	1,117	1,181
FreM1.wav	1,029	1,010	1,033	1,000	1,000	1,000	1,000	1,000
FreM2.wav	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
GerF1.wav	1,246	1,206	1,000	1,040	1,000	1,000	1,000	1,000
GerF2.wav	1,382	1,355	1,285	1,062	1,064	1,000	1,000	1,000
GerM1.wav	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
GerM2.wav	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
ItaF1.wav	1,353	1,260	1,508	1,216	1,452	1,061	1,191	1,000
ItaF2.wav	1,764	1,661	1,447	1,411	1,421	1,323	1,091	1,194
ItaM1.wav	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
ItaM2.wav	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
JapF1.wav	2,127	2,020	1,997	1,798	1,612	1,457	1,538	1,379
JapF2.wav	1,794	1,986	1,689	1,889	1,824	1,716	1,665	1,550
JapM1.wav	1,660	1,657	1,674	1,435	1,453	1,357	1,174	1,253
JapM2.wav	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000

Fonte: Autor.

Tabela 5.16 – Tabela com os resultados do processamento do algoritmo da recomendação ITU-T P.862 para a base de dados da recomendação ITU-T P.862 para as taxas de PLR 0.5, 1.0, 1.5, 2.0, 2.5, 3.5, 4.5 e 5.5%.

Classes do Índice MOS para as taxas de PLR obtidos pelo algoritmo P.862								
Arquivo	**0,05%**	**0,10%**	**0,15%**	**0,20%**	**0,25%**	**0,35%**	**0,45%**	**0,55%**
or105.wav	3,587	3,459	3,308	3,220	3,162	3,130	2,895	2,922
or109.wav	3,579	3,425	3,408	3,230	3,109	3,132	2,936	2,907
or114.wav	3,768	3,310	3,086	3,163	2,954	2,996	2,824	2,666
or129.wav	3,708	3,559	3,463	3,372	3,443	3,295	3,295	3,249
or134.wav	3,683	3,451	3,428	3,373	3,369	3,290	3,208	3,296
or137.wav	3,673	3,532	3,544	3,414	3,440	3,313	3,233	3,164
or145.wav	3,434	3,178	3,140	3,014	3,002	2,800	2,632	2,457
or149.wav	3,890	3,881	3,648	3,609	3,413	3,402	3,303	3,368
or152.wav	3,776	3,410	3,365	3,369	3,159	3,041	2,946	2,983
or154.wav	3,771	3,438	3,601	3,437	3,592	3,230	3,194	3,159
or155.wav	3,982	3,817	3,632	3,506	3,438	3,399	3,336	3,369
or161.wav	3,825	3,813	3,668	3,602	3,451	3,471	3,407	3,299
or164.wav	3,844	3,666	3,423	3,330	3,349	3,194	2,996	3,013
or166.wav	3,675	3,429	3,398	3,346	3,349	3,167	3,174	3,014
or170.wav	3,733	3,626	3,479	3,265	3,306	3,183	3,077	3,067
or179.wav	3,856	3,653	3,643	3,442	3,478	3,283	3,323	3,321
or221.wav	3,718	3,397	3,223	3,262	3,191	3,038	2,938	2,917
or229.wav	3,750	3,484	3,388	3,306	3,176	3,154	3,150	3,046
or246.wav	3,748	3,522	3,725	3,547	3,377	3,423	3,278	3,228
or272.wav	3,803	3,504	3,313	3,217	3,407	3,141	3,062	3,090

Fonte: Autor.

Tabela 5.17 – Tabela com os resultados do processamento do algoritmo da recomendação ITU-T P.862 para a base de dados da recomendação ITU-T P.862 para as taxas de PLR 7.0, 8.0, 9.0, 10.0, 12.0, 15.0, 18.0 e 21.0%.

Classes do Índice MOS para as taxas de PLR obtidos pelo algoritmo P.862								
Arquivo	**0,70%**	**0,80%**	**0,90%**	**1,00%**	**1,20%**	**1,50%**	**1,80%**	**2,10%**
or105.wav	2,810	2,700	2,771	2,707	2,629	2,471	2,465	2,382
or109.wav	2,850	2,778	2,773	2,718	2,618	2,548	2,437	2,424
or114.wav	2,654	2,572	2,462	2,480	2,354	2,281	2,200	2,113
or129.wav	3,210	3,145	3,060	3,016	3,025	2,943	2,837	2,752
or134.wav	3,125	3,184	3,161	3,133	3,056	2,946	2,882	2,767
or137.wav	3,137	3,102	3,041	3,057	2,999	2,894	2,849	2,768
or145.wav	2,496	2,393	2,388	2,310	2,161	2,038	1,965	1,882
or149.wav	3,127	3,168	3,069	3,112	2,986	2,965	2,813	2,685
or152.wav	2,888	2,797	2,781	2,714	2,654	2,550	2,454	2,356
or154.wav	2,923	3,069	2,968	2,960	2,847	2,774	2,713	2,545
or155.wav	3,172	3,111	3,188	3,018	2,942	2,942	2,690	2,707
or161.wav	3,263	3,173	3,128	3,080	3,012	2,980	2,891	2,787
or164.wav	2,934	2,941	2,880	2,903	2,790	2,736	2,562	2,553
or166.wav	3,009	2,960	2,926	2,930	2,778	2,724	2,626	2,602
or170.wav	3,011	2,941	2,849	2,917	2,807	2,699	2,671	2,551
or179.wav	3,152	3,112	3,063	3,074	3,027	2,920	2,771	2,739
or221.wav	2,829	2,818	2,744	2,670	2,661	2,525	2,500	2,383
or229.wav	2,933	2,960	2,896	2,885	2,763	2,666	2,614	2,555
or246.wav	3,150	3,140	3,095	3,045	2,928	2,888	2,823	2,639
or272.wav	2,874	2,951	2,895	2,853	2,762	2,680	2,573	2,501

Fonte: Autor.

Tabela 5.18 – Primeira parte dos resultados do processamento do algoritmo da recomendação ITU-T P.862 para a base da recomendação ITU-T P.501.

Índice MOS para as taxas de PLR obtidos pelo algoritmo P.862								
Arquivo	**0,50%**	**1,00%**	**1,50%**	**2,00%**	**2,50%**	**3,50%**	**4,50%**	**5,50%**
DutF1.wav	2,882	2,697	2,602	2,437	2,318	2,270	2,258	2,134
DutF2.wav	2,929	2,602	2,652	2,439	2,343	2,205	2,111	2,059
DutM1.wav	2,992	2,820	2,502	2,412	2,320	2,251	2,240	2,173
DutM2.wav	2,535	2,070	1,863	1,806	1,684	1,530	1,471	1,452
EngF1.wav	3,031	2,631	2,588	2,411	2,296	2,221	2,022	2,023
EngF2.wav	2,777	2,449	2,299	2,159	2,121	1,959	1,902	1,813
EngM1.wav	3,026	2,570	2,426	2,377	2,322	2,068	1,949	1,966
EngM2.wav	2,925	2,610	2,505	2,415	2,335	2,232	2,091	2,081
FinF1.wav	2,565	1,956	1,960	1,826	1,805	1,701	1,651	1,576
FinF2.wav	2,340	2,085	1,952	1,906	1,778	1,612	1,622	1,559
FinM1.wav	2,796	2,507	2,229	2,200	2,170	1,855	1,913	1,806
FinM2.wav	2,946	2,564	2,398	2,288	2,190	2,091	2,002	1,917
FreF1.wav	3,260	2,945	2,795	2,601	2,640	2,516	2,378	2,309
FreF2.wav	2,879	2,614	2,390	2,323	2,232	2,076	1,963	1,889
FreM1.wav	3,070	2,708	2,654	2,555	2,416	2,211	2,130	2,037
FreM2.wav	3,128	2,990	2,784	2,677	2,559	2,430	2,406	2,339
GerF1.wav	2,799	2,663	2,514	2,408	2,355	2,238	2,159	2,147
GerF2.wav	2,563	2,466	2,249	1,989	1,889	1,826	1,747	1,676
GerM1.wav	2,708	2,148	2,154	2,011	1,939	1,808	1,797	1,707
GerM2.wav	2,308	2,135	1,944	1,824	1,783	1,718	1,633	1,606
ItaF1.wav	2,883	2,691	2,542	2,488	2,350	2,197	2,196	2,087
ItaF2.wav	2,526	2,217	2,121	1,969	1,999	1,829	1,763	1,728
ItaM1.wav	2,688	2,380	2,100	2,087	1,935	1,877	1,729	1,690
ItaM2.wav	2,827	2,563	2,374	2,282	2,265	2,137	2,133	2,001
JapF1.wav	2,727	2,319	2,367	2,114	2,116	1,965	1,826	1,776
JapF2.wav	3,261	2,958	2,857	2,709	2,656	2,533	2,431	2,361
JapM1.wav	3,034	2,808	2,648	2,557	2,536	2,408	2,350	2,266
JapM2.wav	2,934	2,581	2,445	2,317	2,314	2,175	2,040	1,892

Fonte: Autor.

Tabela 5.19 – Segunda parte dos resultados do processamento do algoritmo da recomendação ITU-T P.862 para a base da recomendação ITU-T P.501.

Índice MOS para as taxas de PLR obtidos pelo algoritmo P.862								
Arquivo	**7,00%**	**8,00%**	**9,00%**	**10,00%**	**12,00%**	**15,00%**	**18,00%**	**21,00%**
DutF1.wav	2,026	1,955	1,904	1,922	1,864	1,811	1,744	1,674
DutF2.wav	1,999	1,874	1,828	1,766	1,721	1,644	1,594	1,501
DutM1.wav	2,044	2,038	2,052	2,000	1,964	1,907	1,854	1,802
DutM2.wav	1,338	1,292	1,292	1,252	1,227	1,151	1,116	1,067
EngF1.wav	1,915	1,875	1,832	1,770	1,733	1,641	1,572	1,550
EngF2.wav	1,722	1,666	1,616	1,554	1,530	1,395	1,334	1,248
EngM1.wav	1,891	1,786	1,763	1,671	1,611	1,499	1,423	1,353
EngM2.wav	1,937	1,960	1,896	1,816	1,749	1,668	1,605	1,563
FinF1.wav	1,491	1,466	1,427	1,399	1,366	1,328	1,315	1,263
FinF2.wav	1,512	1,424	1,421	1,374	1,379	1,349	1,271	1,252
FinM1.wav	1,709	1,654	1,578	1,585	1,522	1,481	1,422	1,372
FinM2.wav	1,868	1,849	1,784	1,732	1,760	1,681	1,599	1,566
FreF1.wav	2,216	2,124	2,130	2,103	2,020	1,943	1,889	1,809
FreF2.wav	1,786	1,740	1,740	1,625	1,582	1,466	1,426	1,331
FreM1.wav	1,959	1,963	1,921	1,819	1,791	1,701	1,579	1,578
FreM2.wav	2,243	2,173	2,139	2,126	2,053	2,000	1,906	1,885
GerF1.wav	2,017	2,006	1,952	1,970	1,883	1,808	1,763	1,679
GerF2.wav	1,629	1,563	1,534	1,499	1,454	1,398	1,359	1,303
GerM1.wav	1,620	1,601	1,542	1,481	1,477	1,462	1,366	1,331
GerM2.wav	1,555	1,511	1,517	1,497	1,481	1,380	1,312	1,351
ItaF1.wav	1,972	1,939	1,955	1,861	1,761	1,761	1,675	1,588
ItaF2.wav	1,664	1,584	1,557	1,477	1,489	1,434	1,352	1,342
ItaM1.wav	1,593	1,580	1,525	1,477	1,484	1,395	1,334	1,286
ItaM2.wav	2,003	1,965	1,862	1,843	1,821	1,722	1,663	1,587
JapF1.wav	1,685	1,650	1,614	1,579	1,506	1,415	1,378	1,296
JapF2.wav	2,300	2,266	2,200	2,207	2,100	1,974	1,929	1,870
JapM1.wav	2,154	2,147	2,097	2,062	1,993	1,949	1,901	1,865
JapM2.wav	1,799	1,794	1,739	1,710	1,651	1,556	1,483	1,422

Fonte: Autor.

As Tabelas 5.12, 5.13, 5.14 e 5.15 representam o índice MOS obtido pelo algoritmo P.563 para as duas bases de dados, como pode ser verificado devido o P.563 não conseguir diferenciar silêncio de perda de pacote, todos os índices MOS resultantes representam uma baixa QoE para o usuário, onde alguns receberam o índice MOS de menor qualidade 1.0.

Como pode ser visto nas Tabelas 5.12, 5.13, 5.14, 5.15, 5.16, 5.17, 5.18 e 5.19 contém os resultados dos processamentos da execução dos algoritmos das recomendações ITU-T P.563 e ITU-T P.862, ao compararmos os resultados das Tabelas 5.12 e 5.13 com as Tabelas 5.14 e 5.15 e as Tabelas 5.16 e 5.17 com as Tabelas 5.18 e 5.19 podemos verificar que os resultados para as execuções com o algoritmo da recomendação ITU-T P.563 são muito inferiores aos resultados

das execuções com o algoritmo da recomendação ITU-T P.682. Através da análise das tabelas podemos perceber que grande parte dos dados das Tabelas 5.12 e 5.13 se encontram entre as classes 1 que possui os valores entre 1.00 à 3.09 e classe 2 possuindo valores entre 3.10 à 3.59 da tabela do Fator-R.

Ao utilizar os resultados do índice MOS processamento do algoritmo da recomendação ITU-T P.563 para treinamento e validação do modelo, pode-se realizar uma separação das classes de forma incorreta, pois os resultados para as execuções com o algoritmo intrusivo da recomendação ITU-T P.862 são melhores devido a sua capacidade de diferenciar silêncios de perdas de pacote e ao realizar o treinamento e validação dos modelos com as classes incorretas, os resultados não seriam válidos para o objetivo deste trabalho.

Porém, ao utilizar os resultados do algoritmo da recomendação ITU-T P.862 para realizar o treinamento e validação do modelo para reconhecer e classificar cada uma das quatro classes com resultados válidos devido a separação correta das classes, o modelo final resultante treinado a partir das classes se tornou um novo método não intrusivo capaz de classificar a taxa perda de pacotes no sinal de voz em qualquer arquivo.

Tabela 5.20 – Resultado da matriz de confusão da comparação dos resultados do índice MOS das execuções dos algoritmos ITU-T P.563 e ITU-T P.862.

	Classe 1 - P.862	Classe 2 - P.862	Classe 3 - P.862	Classe 4 - P.862
Classe 1 - P.563	3894	2779	615	25
Classe 2 - P.563	0	20	25	0
Classe 3 - P.563	0	0	2	0
Classe 4 - P.563	0	0	0	0

Fonte: Autor.

Ao realizar a classificação dos índice MOS do Fator-R com os resultados dos algoritmos das recomendações ITU-T P.563 e ITU-T P.862, obtivemos de acordo com a Tabela 5.20 uma média de 53,21% de acerto quando comparamos os resultados obtidos do algoritmo P.563 com os obtidos do algoritmo P.862. Onde o algoritmo P.563 classificou de forma correta as classes de 3916 arquivos e errou as classes de 3444 arquivos da base de dados.

Ao comparar os resultados alcançados do algoritmo P.563 com os resultados do obtidos pelo modelo, onde o modelo foi treinado e validado de acordo com os resultados do MOS obtido pelo algoritmo da recomendação ITU-T P.862, podemos verificar que o modelo obteve um melhor desempenho que o algoritmo P.563, onde foi obtido uma acurácia de 91% para o modelo e de 53,21% para o algoritmo P.563.

6 CONCLUSÃO

Através dos resultados do treinamento e validação do modelo proposto para a base de dados para valores de PLR na sessão 5.1, podemos perceber que após ser executado com 1000 épocas, conforme os dados da Tabela 5.5, foi obtido uma taxa de acurácia de aproximadamente 94% no reconhecimento da classe por taxa de perda de pacotes. Através dessa taxa podemos perceber que a quantidade de arquivos disponíveis na base de dados foi suficiente para fornecer todas as características para o treinamento do modelo, e também que o número de épocas está diretamente relacionado ao resultado da acurácia obtido.

Através dos resultados do treinamento e validação do modelo proposto para a base de dados para valores de índice MOS na sessão 5.2, podemos perceber que a medida que fomos aumentando o número de épocas vimos que o número de erros de validação foi aumentado, porém após executar o modelo com um número de épocas igual a 500 e 1000, percebe-se que a quantidade de erros reduziu. Com isso podemos dizer que para o modelo, utilizando a base de dados preparada com valores de índice MOS, o número de épocas pode ser um fator limitante ao resultado final alcançado, porém pode-se obter um resultado melhor que o obtido ao aumentar o número de épocas para o modelo.

Através dos resultados alcançados pode-se concluir que o modelo proposto atende o objetivo principal do projeto que é classificar a taxa de perda de pacotes e identificar o índice MOS através de um algoritmo Deep Learning, onde a sua taxa de acerto foi em média de 94% para o modelo treinado pela taxa de perda de pacotes e de 91% para o modelo treinado pelo índice MOS dos arquivos degradados de áudio. Podemos destacar também que através dos resultados do modelo proposto com os resultados do algoritmo da recomendação ITU-T P.563, percebe-se que o modelo proposto obteve uma maior eficiência que o algoritmo da recomendação ITU-T P.563, se tornando um modelo de análise da qualidade de voz de forma não intrusiva com maior eficiência, podendo através dele monitorar as transmissões VoIP reconhecendo a taxa de perda de pacotes em tempo real e executando alguma ferramenta de correção para que a experiência do usuário não seja afetada.

Para trabalhos futuros podemos utilizar o modelo definido e seus resultados para identificar perdas de pacotes e outros problemas na rede como a variação do atraso dos pacotes (PDV - Packet Delay Variation) que é o atraso unidirecional de ponta a ponta entre os pacotes enviados na rede, também futuramente podemos treinar o modelo para classificar e reconhecer o Jitter, principalmente em comunicações que necessitam de uma boa QoE para aplicações em tempo

real, e com isso possibilitar o desenvolvimento de novas soluções com o objetivo de identificar novos problemas na rede que ocasionaram a perda de pacote e e solucioná-los, buscando melhorar a QoE do usuário em redes com uma QoS não aceitável.

REFERÊNCIAS

ABBAS, S.; MOSBAH, M.; ZEMMARI, A. Itu-t recommendation g. 114,“one way transmission time. In: CITESEER. **In International Conference on Dynamics in Logistics 2007 (LDIC 2007), Lect. Notes in Comp. Sciences**. 1996. Disponível em: <https://www.itu.int/rec/T-REC-G.114>.

AFFONSO, E. T. et al. Speech quality assessment in wireless voip communication using deep belief network. **IEEE Access**, IEEE, v. 6, p. 77022–77032, 2018. Disponível em: <https://ieeexplore.ieee.org/document/8513822>.

AMODEI, D. et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In: **International conference on machine learning**. [s.n.], 2016. p. 173–182. Disponível em: <https://pdfs.semanticscholar.org/8e0e/acf11a22b9705a262e908f17b1704fd21fa7.pdf>.

ARNDT, S. et al. Using electroencephalography to measure perceived video quality. **IEEE Journal of Selected Topics in Signal Processing**, IEEE, v. 8, n. 3, p. 366–376, 2014. Disponível em: <https://ieeexplore.ieee.org/document/6777327/>.

ASSEM, H. et al. Monitoring voip call quality using improved simplified e-model. In: IEEE. **2013 International Conference on Computing, Networking and Communications (ICNC)**. 2013. p. 927–931. Disponível em: <https://ieeexplore.ieee.org/document/6504214>.

ASSEMBLY, T. **ITU-T P. 862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs**. [S.l.], 2000. Disponível em: <https://www.itu.int/rec/T-REC-P.862>.

AZEVEDO, L. P. d. **Aplicação de redes neurais artificiais no processo de classificação de orquídeas do gênero Cattleya.** 2016. Monografia (Bacharel em Sistema de Informção), IFMG (Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais), Bagé, Brazil. Disponível em: <https://www.ifmg.edu.br/sabara/biblioteca/trabalhos-de-conclusao-de-curso/tcc-documentos/TCCLucasAzevedo.pdf>.

BEHDADFAR, M.; FAGHIHI, E.; SADEGHI, M. E. Qos parameters analysis in voip network using adaptive quality improvement. In: IEEE. **2015 Signal Processing and Intelligent Systems Conference (SPIS)**. 2015. p. 73–77. Disponível em: <https://ieeexplore.ieee.org/document/7422315>.

BENGIO, Y.; COURVILLE, A.; VINCENT, P. Representation learning: A review and new perspectives. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, v. 35, n. 8, p. 1798–1828, 2013. Disponível em: <https://ieeexplore.ieee.org/document/6472238>.

BERGSTRA, J. A.; MIDDELBURG, C. Itu-t recommendation g. 107: The e-model, a computational model for use in transmission planning. Citeseer, 2003. Disponível em: <https://www.itu.int/rec/T-REC-G.107>.

BOTTOU, L. From machine learning to machine reasoning. **Machine learning**, Springer, v. 94, n. 2, p. 133–149, 2014. Disponível em: <https://arxiv.org/pdf/1102.1808>.

BUENO, L. E. P. Codificação de áudio para transmissão de voz em tempo real. **Universidade Federal do Paraná**, 2008. Disponível em: <http://www.eletrica.ufpr.br/marcelo/TE072/022008/Luis-CodecsFala.pdf>.

ÇALIK, R. C.; DEMIRCI, M. F. Cifar-10 image classification with convolutional neural networks for embedded systems. In: IEEE. **2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA)**. 2018. p. 1–2. Disponível em: <https://ieeexplore.ieee.org/document/8612873/>.

CARVALHO, J. L. A.; DANILO, D. Técnicas de codificação de voz aplicadas em sistemas móveis celulares. **ENE/FT/UNB**, 2000. Disponível em: <http://www.ene.unb.br/joaoluiz/pdf/carvalho-dias2000_vocoders.pdf>.

CAVALCANTE, A. d. V. **Análise dos Efeitos de Codecs de áudio na Avaliação de Desvios Vocais**. Dissertação (Mestrado), 2018. Disponível em: <http://repositorio.ifpb.edu.br/xmlui/handle/177683/344>.

CHETLUR, S. et al. cudnn: Efficient primitives for deep learning. **arXiv preprint arXiv:1410.0759**, 2014. Disponível em: <https://arxiv.org/pdf/1410.0759.pdf>.

CHOUDHURY, S.; GIBSON, J. D. Payload length and rate adaptation for multimedia communications in wireless lans. **IEEE Journal on Selected Areas in Communications**, IEEE, v. 25, n. 4, p. 796–807, 2007. Disponível em: <https://ieeexplore.ieee.org/document/4205061>.

COLLOBERT, R.; WESTON, J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In: ACM. **Proceedings of the 25th international conference on Machine learning**. 2008. p. 160–167. Disponível em: <https://ronan.collobert.com/pub/matos/2008_nlp_icml.pdf>.

COPELAND, M. What's the difference between artificial intelligence. **Machine Learning, and Deep Learning**, 2016. Disponível em: <https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/>.

COPPIN, B. **Inteligência Artificial**. LTC, 2010. Disponível em: <https://www.amazon.com.br/Intelig%C3%AAncia-Artificial-Ben-Coppin/dp/8521617291>.

CUNY, R.; LAKANIEMI, A. Voip in 3g networks: An end-to-end quality of service analysis. In: IEEE. **The 57th IEEE Semiannual Vehicular Technology Conference, 2003. VTC 2003-Spring.** 2003. v. 2, p. 930–934. Disponível em: <https://ieeexplore.ieee.org/document/1207762/>.

DENG, L.; YU, D. et al. Deep learning: methods and applications. **Foundations and Trends® in Signal Processing**, Now Publishers, Inc., v. 7, n. 3–4, p. 197–387, 2014. Disponível em: <https://www.nowpublishers.com/article/DownloadSummary/SIG-039>.

EBERLE, W.; PENDERS, J.; YAZICIOGLU, R. F. Closing the loop for deep brain stimulation implants enables personalized healthcare for parkinson's disease patients. In: IEEE. **2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society**. 2011. p. 1556–1558. Disponível em: <https://ieeexplore.ieee.org/document/6090453/>.

FIEDLER, M.; HOSSFELD, T.; TRAN-GIA, P. A generic quantitative relationship between quality of experience and quality of service. **IEEE Network**, IEEE, v. 24, n. 2, p. 36–41, 2010. Disponível em: <https://ieeexplore.ieee.org/document/5430142>.

GARZÓN, N. V. O. **Análise de desempenho de uma proposta de transmissão oportunista sem fio em canais com desvanecimento rayleigh e na presença de interferência de Co-Canal para diferentes esquemas de modulação**. Dissertação (Mestrado) — Campinas/Universidade Estadual de Campinas/2014, 2014. Disponível em: <http://repositorio.unicamp.br/jspui/handle/REPOSIP/258816>.

GEOFFREY, H. et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. **IEEE Signal Processing Magazine**, v. 29, n. 6, p. 82–97, 2012. Disponível em: <https://ieeexplore.ieee.org/document/6296526/>.

GOODE, B. Voice over internet protocol (voip). **Proceedings of the IEEE**, IEEE, v. 90, n. 9, p. 1495–1517, 2002. Disponível em: <https://ieeexplore.ieee.org/document/1041060>.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. Deep learning. book in preparation for mit press. **deeplearningbook**, 2016. Disponível em: <http://www.deeplearningbook.org>.

GRAVES, A.; MOHAMED, A.-r.; HINTON, G. Speech recognition with deep recurrent neural networks. In: IEEE. **2013 IEEE international conference on acoustics, speech and signal processing**. 2013. p. 6645–6649. Disponível em: <https://ieeexplore.ieee.org/document/6638947>.

GUO, T. et al. Simple convolutional neural network on image classification. In: IEEE. **2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)**(. 2017. p. 721–724. Disponível em: <https://ieeexplore.ieee.org/document/8078730>.

HAN, K. et al. Learning spectral mapping for speech dereverberation and denoising. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, IEEE, v. 23, n. 6, p. 982–992, 2015. Disponível em: <https://ieeexplore.ieee.org/document/7067387>.

HAYKIN, S. **Redes neurais: princípios e prática**. Bookman Editora, 2007. Disponível em: <https://www.amazon.com.br/Redes-Neurais-Princ%C3%ADpios-Simon-Haykin/dp/8573077182>.

HOSSFELD, T. et al. Formal definition of qoe metrics. **arXiv preprint arXiv:1607.00321**, 2016. Disponível em: <https://www.arxiv.org/abs/1607.00321v1>.

INDEX, C. V. N. Forecast and methodology 2017–2022. **Cisco: San Jose, CA, USA**, 2019. Disponível em: <https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html>.

IRIYA, R. **Análise de sinais de voz para reconhecimento de emoções.** Tese (Doutorado) — Universidade de São Paulo, 2014. Disponível em: <http://www.teses.usp.br/teses/disponiveis/3/3142/tde-14042015-160249/>.

JR, J. R. et al. Reconhecimento automático de emoções através da voz. Florianópolis, SC, 2017. Disponível em: <https://repositorio.ufsc.br/bitstream/handle/123456789/182186/reconhecimento_automatico_emocoes_voz_final_pdfa.pdf?sequence=1>.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: **Advances in neural information processing systems**. [s.n.], 2012. p. 1097–1105. Disponível em: <https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf>.

LAGHARI, K. U. R.; CONNELLY, K. Toward total quality of experience: A qoe model in a communication ecosystem. **IEEE Communications Magazine**, IEEE, v. 50, n. 4, p. 58–65, 2012. Disponível em: <https://ieeexplore.ieee.org/document/6178834>.

LANGSTON, J. W. The parkinson's complex: parkinsonism is just the tip of the iceberg. **Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society**, Wiley Online Library, v. 59, n. 4, p. 591–596, 2006. Disponível em: <https://www.researchgate.net/publication/7213310_The_Parkinson's_Complex_Parkinsonism_Is_Just_the_Tip_of_the_Iceberg>.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **nature**, Nature Publishing Group, v. 521, n. 7553, p. 436, 2015. Disponível em: <https://www.nature.com/articles/nature14539>.

LECUN, Y.; BENGIO, Y. et al. Convolutional networks for images, speech, and time series. **The handbook of brain theory and neural networks**, v. 3361, n. 10, p. 1995, 1995. Disponível em: <http://yann.lecun.com/exdb/publis/pdf/lecun-bengio-95a.pdf>.

LECUN, Y. et al. Neural networks: Tricks of the trade. **Springer Lecture Notes in Computer Sciences**, v. 1524, n. 5-50, p. 6, 1998. Disponível em: <https://www.researchgate.net/publication/321613343_Neural_Networks_Tricks_of_the_Trade_Second_Edition>.

LECUN, Y.; KAVUKCUOGLU, K.; FARABET, C. Convolutional networks and applications in vision. In: IEEE. **Proceedings of 2010 IEEE International Symposium on Circuits and Systems**. 2010. p. 253–256. Disponível em: <https://ieeexplore.ieee.org/document/5537907>.

LECUN, Y.; RANZATO, M. Deep learning tutorial. In: CITESEER. **Tutorials in International Conference on Machine Learning (ICML'13)**. 2013. p. 1–29. Disponível em: <https://cs.nyu.edu/~yann/talks/lecun-ranzato-icml2013.pdf>.

LEE, H. et al. Enhancing voice over wlan via rate adaptation and retry scheduling. **IEEE Transactions on Mobile Computing**, IEEE, v. 13, n. 12, p. 2791–2805, 2013. Disponível em: <https://ieeexplore.ieee.org/abstract/document/6517180/>.

LI, J. et al. Deep reinforcement learning for dialogue generation. **arXiv preprint arXiv:1606.01541**, 2016. Disponível em: <https://arxiv.org/abs/1606.01541>.

LIN, C.-S.; JENG, W. Using content analysis in lis research: Experiences with coding schemes construction and reliability measures. **Qualitative and Quantitative Methods in Libraries**, v. 4, n. 1, p. 87–95, 2017. Disponível em: <https://ieeexplore.ieee.org/document/6999220>.

LIU, J.-M. et al. Cough detection using deep neural networks. In: IEEE. **2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)**. [S.l.], 2014. p. 560–563.

MAIND, S. B.; WANKAR, P. et al. Research paper on basic of artificial neural network. **International Journal on Recent and Innovation Trends in Computing and Communication**, v. 2, n. 1, p. 96–100, 2014. Disponível em: <https://www.academia.edu/7197728/Research_Paper_on_Basic_of_Artificial_Neural_Network>.

MALFAIT, L.; BERGER, J.; KASTNER, M. P. 563 — the itu-t standard for single-ended speech quality assessment. **IEEE Transactions on Audio, Speech, and Language Processing**, IEEE, v. 14, n. 6, p. 1924–1934, 2006. Disponível em: <https://ieeexplore.ieee.org/document/1709882>.

MANOUSOS, M. et al. Voice-quality monitoring and control for voip. **IEEE Internet Computing**, IEEE, v. 9, n. 4, p. 35–42, 2005. Disponível em: <https://ieeexplore.ieee.org/document/1463159>.

MATIAS, A. B. **Avaliação de QoS em sistema de comunicação VoIP utilizando plataforma embarcada.** 2010. Trabalho de graduação em Engenharia Elétrica, Publicação PPGENE. DM, Departamento de Engenharia Elétrica-Faculdade de Tecnologia, Universidade de Brasília, Brasília, DF. Disponível em: <http://bdm.unb.br/bitstream/10483/972/1/2008_Andr%C3%A9BisinotoMatias.pdf>.

MITTAG, G.; MÖLLER, S. Non-intrusive estimation of packet loss rates in speech communication systems using convolutional neural networks. In: IEEE. **2018 IEEE International Symposium on Multimedia (ISM)**. 2018. p. 105–109. Disponível em: <https://ieeexplore.ieee.org/iel7/8603129/8603241/08603267.pdf>.

MOHAMED, A.-r. et al. Deep belief networks using discriminative features for phone recognition. In: **ICASSP**. [s.n.], 2011. p. 5060–5063. Disponível em: <https://ieeexplore.ieee.org/document/5947494/>.

MOHARIR, M. et al. Identification of asphyxia in newborns using gpu for deep learning. In: IEEE. **2017 2nd International Conference for Convergence in Technology (I2CT)**. 2017. p. 236–239. Disponível em: <https://ieeexplore.ieee.org/iel7/8168943/8226083/08226127.pdf>.

NUNES, R. D. **Algoritmo para melhorar o desempenho de uma métrica não intrusiva de qualidade de voz.** Dissertação (Mestrado) — Universidade Federal de Lavras, 2017. Disponível em: <http://repositorio.ufla.br/jspui/bitstream/1/12703/2/DISSERTA%c3%87%c3%83O_Algoritmo%20para%20melhorar%20o%20desempenho%20de%20uma%20m%c3%a9trica%20n%c3%a3o%20intrusiva%20de%20qualidade%20de%20voz.pdf>.

OBESO, J. A.; OLANOW, C. W.; NUTT, J. G. **Levodopa motor complications in Parkinson's disease**. Elsevier, 2000. Disponível em: <https://www.ncbi.nlm.nih.gov/pubmed/11052214>.

OOSTER, J.; MEYER, B. T. Improving deep models of speech quality prediction through voice activity detection and entropy-based measures. In: IEEE. **ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. 2019. p. 636–640. Disponível em: <https://ieeexplore.ieee.org/document/8682754>.

ORTEGA, M. O.; ALTAMIRANO, G. C.; ABAD, M. F. Evaluation of the voice quality and qos in real calls using different voice over ip codecs. In: IEEE. **2018 IEEE Colombian Conference on Communications and Computing (COLCOM)**. 2018. p. 1–6. Disponível em: <https://ieeexplore.ieee.org/document/8466727/>.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in python. **Journal of machine learning research**, v. 12, n. Oct, p. 2825–2830, 2011. Disponível em: <http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf>.

PENTEADO, R. Z. Locução em propagandas: uma releitura da caracterização de vozes profissionais. **Impulso**, v. 19, n. 48, p. 85–94, 2009. Disponível em: <https://www.researchgate.net/publication/276259555_Locucao_em_Propagandas_Uma_Releitura_da_Caracterizacao_de_Vozes_Profissionais>.

REC, I. P. 800: Methods for subjective determination of transmission quality. **International Telecommunication Union, Geneva**, p. 22, 1996. Disponível em: <https://www.itu.int/rec/T-REC-P.800-199608-I/en>.

REC, I. Itu-t recommendation g. 131,"talker echo and its control". **International Telecommunications Union, Geneva, Switzerland**, 2003. Disponível em: <https://www.itu.int/rec/T-REC-G.131/en>.

REICHL, P. et al. Towards a comprehensive framework for qoe and user behavior modelling. In: IEEE. **2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)**. 2015. p. 1–6. Disponível em: <https://ieeexplore.ieee.org/abstract/document/7148138/>.

RIBEIRO, G. et al. Análise mel-cepstral na discriminação de patologias laríngeas. XXIV Congresso Brasileiro de Engenharia Biomédica – CBEB, 2014. Disponível em: <http://www.canal6.com.br/cbeb/2014/artigos/cbeb2014_submission_825.pdf>.

RIX, A. W. et al. Objective assessment of speech and audio quality—technology and applications. **IEEE Transactions on Audio, Speech, and Language Processing**, IEEE, v. 14, n. 6, p. 1890–1901, 2006. Disponível em: <https://ieeexplore.ieee.org/document/1709879>.

RIZZO, G. et al. Accuracy of clinical diagnosis of parkinson disease: a systematic review and meta-analysis. **Neurology**, AAN Enterprises, v. 86, n. 6, p. 566–576, 2016. Disponível em: <https://www.researchgate.net/publication/290443265_Accuracy_of_clinical_diagnosis_of_Parkinson_disease_A_systematic_review_and_meta-analysis>.

RODRÍGUEZ, D. Z. **Algoritmo para determinação da taxa de transmissão em uma rede IP.** Tese (Doutorado) — Universidade de São Paulo, 2009. Disponível em: <https://www.teses.usp.br/teses/disponiveis/3/3142/tde-30032010-153910/publico/dissertacao_demostenes.pdf>.

SANCHEZ-IBORRA, R.; CANO, M.-D.; GARCIA-HARO, J. Performance evaluation of qoe in voip traffic under fading channels. In: IEEE. **2013 World congress on computer and information technology (WCCIT)**. 2013. p. 1–6. Disponível em: <http://ieeexplore.ieee.org/document/6618721>.

SANTOS, V. E. L. dos et al. Análise de qualidade de voz de chamadas voip para diferentes codecs em links terrestre e satélite. 2014. Disponível em: <http://monografias.poli.ufrj.br/monografias/monopoli10012240.pdf>.

SCHIFFMAN, D. The nature of code: Simulating natural systems with processing. **The Nature of**, 2012. Disponível em: <https://www.amazon.com.br/Nature-Code-Simulating-Natural-Processing/dp/0985930802>.

SCHMIDHUBER, J. Deep learning in neural networks: An overview. **Neural networks**, Elsevier, v. 61, p. 85–117, 2015. Disponível em: <https://arxiv.org/abs/1404.7828>.

SHANMUGAN, K. S. Simulation-based estimate of qos for voice traffic over wcdma radio links. In: IEEE. **2009 5th International Conference on Wireless Communications, Networking and Mobile Computing**. 2009. p. 1–4. Disponível em: <https://ieeexplore.ieee.org/document/5301913>.

SHARMA, V.; RAI, S.; DEV, A. A comprehensive study of artificial neural networks. **International Journal of Advanced research in computer science and software engineering**, v. 2, n. 10, 2012. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.468.9353&rep=rep1&type=pdf>.

SHU, H.; SONG, Y.; ZHOU, H. Time-frequency performance study on urban sound classification with convolutional neural network. In: IEEE. **TENCON 2018-2018 IEEE Region 10 Conference**. 2018. p. 1713–1717. Disponível em: <https://ieeexplore.ieee.org/document/8650428/>.

SINAM, T. et al. A technique for classification of voip flows in udp media streams using voip signalling traffic. In: IEEE. **2014 IEEE International Advance Computing Conference (IACC)**. 2014. p. 354–359. Disponível em: <https://ieeexplore.ieee.org/document/6779348>.

TAKASE, H.; GOUHARA, K.; UCHIKAWA, Y. Time sequential pattern transformation and attractors of recurrent neural networks. In: IEEE. **Proceedings of 1993 International Conference on Neural Networks (IJCNN-93-Nagoya, Japan)**. 1993. v. 3, p. 2319–2322. Disponível em: <https://ieeexplore.ieee.org/document/714189>.

TRAORE, A. B. et al. Improvement of voip service quality based on an adaptive congestion control method. In: IEEE. **2018 14th IEEE International Conference on Signal Processing (ICSP)**. 2018. p. 265–269. Disponível em: <https://ieeexplore.ieee.org/document/8652492>.

TÜNDIK, M. Á. et al. Assessment of pathological speech prosody based on automatic stress detection and phrasing approaches. In: IEEE. **2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)**. 2017. p. 000067–000072. Disponível em: <https://ieeexplore.ieee.org/document/8268218>.

TUSHAR, A. Making sense of hidden layer information in deep networks by learning hierarchical targets. **arXiv preprint arXiv:1505.00384**, 2015. Disponível em: <https://arxiv.org/abs/1505.00384>.

WALKER, J. Q.; HICKS, J. T. The essential guide to voip implementation and management. **NetIQ Corporation**, Citeseer, 2002. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.460.5504&rep=rep1&type=pdf>.

WAND, M.; SCHULTZ, T. Pattern learning with deep neural networks in emg-based speech recognition. In: IEEE. **2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society**. 2014. p. 4200–4203. Disponível em: <https://ieeexplore.ieee.org/document/6944550/>.

WISNEVSKI, F. L.; FAGUNDES, R. D. R.; COSSIO, L. P. Codificador de voz baseado na qualidade perceptual. **V Mostra de Pesquisa da Pós-Graduação**, PUCRS, 2010. Disponível em: <http://www.pucrs.br/edipucrs/Vmostra/V_MOSTRA_PDF/Engenharia_Eletrica/83133-FLAVIO_LUIS_WISNEVSKI.pdf>.

WROGE, T. J. et al. Parkinson's disease diagnosis using machine learning and voice. In: IEEE. **2018 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)**. 2018. p. 1–7. Disponível em: <https://ieeexplore.ieee.org/document/8615607/>.

WU, M. et al. Characteristics of drug combination therapy in oncology by analyzing clinical trial data on clinicaltrials. gov. In: WORLD SCIENTIFIC. **Pacific Symposium on Biocomputing Co-Chairs**. 2014. p. 68–79. Disponível em: <https://www.ncbi.nlm.nih.gov/pubmed/25592569>.

XU, J.; ZHANG, C. Research of an improved packet loss compensation algorithm in voip. 05 2011. Disponível em: <https://ieeexplore.ieee.org/document/6014207/>.

YUHE, S.; JIE, X. New solutions of voip on multi-hop wireless network. In: IEEE. **2009 IITA International Conference on Control, Automation and Systems Engineering (case 2009)**. 2009. p. 199–202. Disponível em: <https://ieeexplore.ieee.org/document/5194425>.