Avaliação da Qualidade da Voz em Serviços de Comunicação usando *Deep Learning*

Diego Jose de Sousa Gouveia[1], **Renata Lopes Rosa**[1], **Demóstenes Zegarra Rodríguez**[1]

[1]Ciência da Computação – Universidade Federal de Lavras (UFLA)
Cep 37200-000 – Lavras – MG – Brazil

{diegosousa.st,renata.rosa,demostenes.zegarra}@dcc.ufla.br

Abstract. *The telephone services based on IP networks are very used around the world. However, the Packet Loss Rate (PLR) can occur on IP networks, affecting the users Quality of Experience (QoE), being necessary to perform the assessment of the speech quality. The determination of a methodology to predict a speech quality is relevant and necessary. Consequently, this paper introduces a novel non-intrusive speech quality model based on deep learning, in order to identify five speech quality classes. A speech database was built, in which different PLRs are applied and the index quality of each file was calculated. Experimental results of performance assessment show that the proposed model overcomes the ITU-T Recommendation P.563.*

Resumo. *Os serviços telefônicos baseados em redes IP são muito utilizados ao redor do mundo. No entanto, uma taxa de perda de pacotes (PLR) pode ocorrer em redes IP, afetando a qualidade de experiência (QoE) dos usuários, sendo necessário avaliar a qualidade da voz. A determinação de uma metodologia para prever uma qualidade da voz é relevante e necessária. Consequentemente, este artigo apresenta um modelo não intrusivo de classificação de qualidade de voz, baseado em aprendizagem profunda utilizando cinco classes. Foi construída uma base de dados, na qual diferentes PLRs são aplicadas e o índice de qualidade de cada arquivo foi calculado. Os resultados experimentais mostram que o desempenho do modelo proposto supera a Recomendação ITU-T P.563.*

1. Introdução

Os usuários de aplicativos de voz cresceram nos últimos anos [Montag et al. 2015], principalmente devido ao aumento do número de dispositivos móveis. No entanto, um canal de comunicação pode sofrer degradações, afetando a Qualidade de Experiência (QoE, *do inglês Quality of Experience*) do usuário [Rodríguez et al. 2014]. Portanto, pesquisas referentes à avaliação da qualidade de um sinal de voz é relevante para as áreas de redes e telecomunicações.

Em um cenário de rede, grandes distâncias conduzem a maiores probabilidades da atenuação do sinal e demais distúrbios em uma comunicação [Cremonezi et al. 2017], como a Taxa de Perda de Pacotes (PLR, *do inglês Packet Loss Rate*), a qual afeta a qualidade de sinal de voz ou vídeo recebido pelo usuário final.

O PLR é considerado um dos parâmetros de rede mais prejudiciais de acordo com algumas pesquisas [Rodríguez et al. 2014]. Assim, a detecção de problemas de rede é essencial para ajudar os sistemas adaptativos implementados pelas operadoras de sistemas

de comunicação. Portanto, a QoE do usuário nos serviços de comunicação de voz e demais serviços [Rodríguez et al. 2016] podem ser melhorados.

A análise da qualidade de fala recebida pode ser realizada por meio de métodos subjetivos e objetivos; os métodos subjetivos são frequentemente utilizados na avaliação da qualidade da voz, mas têm muitas limitações, como um número mínimo de participantes, o custo e o tempo necessário para a realização de testes. Por outro lado, os métodos objetivos que utilizam um algoritmo, são menos dispendiosos e menos demorados.

A Recomendação ITU-T P.862 [ITU-T Rec. P.862 2001], mais conhecida como Avaliação Perceptual da Qualidade da Fala (PESQ, *do inglês Perceptual Evaluation of Speech Quality*), é um método objetivo intrusivo que estima uma Pontuação de Opinião Média (MOS, *do inglês Mean Opinion Score*) para avaliação da qualidade de voz fim-a-fim nas redes telefônicas de banda estreita (NB, *do inglês Narrow Band*). Nos últimos anos, a Recomendação ITU-T P.863 foi aprovada, e o seu algoritmo estima a qualidade da voz de sistemas de telecomunicações de NB até banda larga (SWB, *do inglês Super Wide Band*) [ITU-T Rec. P.863 2014], mas o seu código-fonte não está disponível gratuitamente.

A Recomendação ITU-T P.563 é a métrica objetiva não-intrusiva mais representativa para NB, mas seu desempenho em redes com perdas não é satisfatório [Polacky and Pocta 2014]. Deve-se salientar que recentemente, o Grupo de Estudo 12 da ITU-T lançou um novo processo de padronização para um método objetivo não intrusivo que pretende atender as características tanto das redes NB quanto das SWB [ITU-T Temporary Document 2015].

O desenvolvimento de vários algoritmos de aprendizado de máquina, tais como, as Redes Neurais Artificiais (RNA) e algoritmos de aprendizagem profunda ou *Deep Learning* (DL) têm sido empregados para análise e reconhecimento de fala. Atualmente, a Rede Neural Convolucional (CNN, *do inglês Convolutional Neural Network*), a Máquina de Boltzmann Restrita (RBM, *do inglês Restricted Boltzmann Machine*) e suas variações [Chen et al. 2015] são métodos muito populares utilizados no reconhecimento da fala e imagem, com desempenhos bastante satisfatórios. RBM é uma RNA estocástica generativa, treinada de forma não supervisionada e para problemas que necessitam de classificação, é necessário adicionar um método de aprendizagem supervisionado, classificando as amostras com base nas características extraídas pelo RBM. Estudos [Bengio et al. 2012] sobre identificação de classes em sinais de voz demonstram uma acurácia superior da Máquina de Boltzmann Restrita Discriminativa (DRBM, *do inglês Discriminative Restricted Boltzmann Machine*) em relação ao clássico uso da máquina de vetor de suporte (SVM, *do inglês Support Vector Machine*). De acordo com [Bengio et al. 2012], técnicas padrões de aprendizagem de máquina possuem dificuldades em classificar assinaturas acústicas similares.

Atualmente, um número considerável de estudos que usam métodos de aprendizado profundo se concentram no reconhecimento de fala relacionado à abordagens de linguagem, mas não exploram o sinal de voz associado a degradações de qualidade causada pelos distúrbios que aconteceram nas redes.

Os parâmetros da rede são utilizados como entrada em modelos paramétricos de avaliação da qualidade da voz. Estudos [Lee and Chang 2016] desenvolveram um método

de emascaramento de perda de pacotes usando Deep Neural Networks (DNN), mas a qualidade perceptual do sinal de voz não é tratada. Outros estudos [Monika and Rama 2016] utilizam a Rede Neural conjuntamente com o Modelo Oculto de Markov (HMMs, *do inglês Hidden Markov Models*) na transmissão de voz, mas tampouco consideram os efeitos dos parâmetros de rede na qualidade da comunicação.

A principal contribuição deste artigo é demonstrar que um método classificador de qualidade não intrusivo baseado em DL pode alcançar resultados competitivos em comparação aos métodos intrusivos. Considerando as redes que apresentam perdas de pacotes e cinco classes de qualidade de voz. Para isso, um banco de dados de áudio é construído para ser utilizado como material de teste, no qual diferentes valores de PLR são aplicados. Modelos de distribuição de perdas de pacotes são aplicados em arquivos de voz originais, simulando cenários de rede IP; o PLR é modelado com base em [ITU-T Rec. G.107 2015]. Deve-se ressaltar que a atual recomendação não-intrusiva, a qual não tem resultados correlacionados com testes subjetivos, será utilizada para comparar o desempenho da solução proposta.

Adicionalmente, uma arquitetura de rede é proposta, em que, um Servidor de Qualidade de Voz (SQV) é apresentado. No servidor, um modelo de Classificador de Qualidade de Voz (CQV) é construído com base em uma DRBM.

O restante deste artigo está estruturado da seguinte forma. A seção 2 apresenta uma revisão da literatura sobre, modelos de avaliação de qualidade de voz, os parâmetros que caracterizam o sinal de voz e aprendizagem profunda ou DL; na seção 3 é apresentado o efeito da taxa de perda de pacotes no índice de qualidade da voz. A seção 4 apresenta o modelo de avaliação da qualidade da voz proposto. A seção 5 apresenta a avaliação experimental e os resultados. Finalmente, as conclusões são apresentadas na seção 6.

2. Revisão da Literatura

Nesta seção, em primeiro lugar, os métodos de avaliação de qualidade de voz são brevemente descritos; em seguida, os parâmetros mais utilizados na caracterização do sinal de voz são descritos. Em fim, o modelo de Máquina de Boltzmann Restrita Discriminativa (DRBM, *do inglês, Discriminative Restricted Boltzmann Machine*) é descrito com maior detalhe.

2.1. Métodos de Avaliação de Qualidade de Voz

Métodos de avaliação da qualidade de voz possuem como objetivo atribuir uma pontuação de qualidade a uma determinada comunicação. De uma forma geral, os métodos podem ser classificados em dois grupos principais, métodos subjetivos e objetivos.

Os métodos subjetivos descrevem em detalhes uma metodologia padronizada para realizar experimentos de qualidade de voz, a fim de reduzir a inserção de obliquidades e garantir a capacidade de repetição dos testes. Os resultados do teste subjetivo são determinados pelo índice de qualidade médio determinado por um grupo de avaliadores.

A recomendação ITU-T P.800 [ITU-T Rec. P.800 1996] se estrutura como um modelo para se realizar testes subjetivos de qualidade em laboratórios, objetivando indicar métodos e procedimentos adequados para a determinação da qualidade da fala em serviços de telefonia. Dentre os métodos recomendados estão o teste de conversação subjetiva, que

padroniza simulações de condições reais do serviço telefônico em laboratório, detalhando considerações a serem cumpridas tanto na escolha do ambiente a ser realizados os testes quanto na seleção de participantes, sendo utilizada uma escala composta de cinco pontuações de qualidade de áudio, na qual a média aritmética de qualquer conjunto destas pontuações resulta na *Mean Opinion Score of conversation* (MOSc). Também, trata sobre testes de audição, em que se estabelece o índice MOS, possui considerações menos rigorosas, contudo, possui controle mais rígido de certos parâmetros, suas escalas de opinião são: escala de audição humana, escala de esforço de audição e escala de preferência de audibilidade.

Os métodos objetivos podem ser classificados em diferentes modelos de acordo com a informação empregada na entrada do algoritmo. Os modelos baseados em sinais de fala usam esse sinal para prever o índice de qualidade. Os modelos paramétricos utilizam diferentes fatores, tais como, parâmetros de rede, principalmente atraso e PLR, relacionados com um caminho de transmissão, a potência do ruído, codec de voz, entre outros. Entretanto, é importante ressaltar que um método objetivo não é capaz de reproduzir exatamente as pontuações médias obtidas pelo teste subjetivo. Ademais, os modelos baseados em sinais de fala são divididos em métodos intrusivos e não intrusivos. Os métodos intrusivos precisam de um sinal de referência para comparar com o sinal no ponto final. Já em contraste, os métodos não intrusivos só precisam do sinal de fala no ponto em que é avaliado. Os métodos intrusivos apresentam uma maior acurácia do que os métodos não intrusivos porque eles usam os sinais originais e degradadas. É importante destacar que os métodos não intrusivos são mais apropriados para avaliar a qualidade dos serviços em tempo real, como VoIP, pois o sinal de voz original não está disponível.

A recomendação P.862, mais conhecida como a Avaliação Perceptual de Qualidade da Fala (PESQ), descreve um método intrusivo para predizer a qualidade subjetiva do sinal de voz de banda estreita (0,3 - 3,4 kHz). Resumidamente, a PESQ compara um sinal com um sinal degradado que é o resultado da passagem de através de um sistema de comunicação. O resultado do PESQ é uma previsão da qualidade percebida por um indivíduo em um teste subjetivo, uma pontuação de qualidade do áudio escutado, semelhante a MOS, denominado *Mean Opinion Score Listening Quality Objective* (MOS-LQO).

Por outro lado, a recomendação P.563 é a métrica não intrusiva padronizada mais representativa. O índice de qualidade previsto pelo algoritmo P.563 está relacionado à qualidade percebida em qualquer ponto de uma comunicação. O algoritmo P.563 funciona identificando a classe de distorção principal do sinal degradado e, depois de aplicar um modelo de qualidade de fala, retorna o índice MOS, a MOS-LQO. Entretanto, deve-se ressaltar que o algoritmo P.563 não proporciona uma avaliação completa da qualidade, só se podem medir efeitos da distorção unidirecional da voz e do ruído sobre a qualidade sonora; assim, atrasos, eco do locutor, dentre outros que afetam as interações bidirecionais não influenciam as pontuações da P.563, além de ser projetado exclusivamente para avaliação da voz humana.

A recomendação ITU-T Rec. G.107, mais conhecida como E-Model, é o modelo paramétrico mais utilizado na avaliação de qualidade de voz, medindo os efeitos dos parâmetros de uma rede de transporte e das condições acústicas do meio no qual estão os locutores.

2.2. Caracterização do Sinal de Voz e Modelos de Classificação

A área de reconhecimento do sinal de fala na maioria das soluções envolve um sinal de referência a ser manipulado [Graves 2012], no entanto, a adaptabilidade à novas características da voz exige métodos mais complexos. As características e propriedades da voz geralmente são determinadas usando diferentes recursos. O sinal de voz é uma sequência de vogais e consoantes, e esse fato exibe mudanças rápidas no parâmetro de taxa de cruzamento por zero (ZCR, *do inglês Zero Crossing Rate*); então, para compreender a variabilidade de ZCR, o sinal precisa ser analisado no domínio temporal.

De acordo com [Saini and Kaur 2013], em um reconhecimento de sinal de voz, HMMs podem ser utilizados, para reconhecer a variabilidade temporal da fala, e os modelos de misturas de gaussianas (GMMs, *do inglês Gaussian mixture model*) são utilizados para modelar a densidade dos estados no HMM. No reconhecimento de voz, os HMMs são utilizados porque o sinal de voz pode ser observado, como um sinal estacionário por partes. No entanto, para modelar dados no espaço, o GMM é estatisticamente ineficiente de acordo com o estudo realizado em [Hinton et al. 2012].

Outros recursos populares são os coeficientes Cepstrais de Frequência Mel (*MFCC, do inglês Mel-Frequency Cepstrum Coefficients*), que fornece uma representação de sinal de voz compacta. Os parâmetros MFCC, *pitch*, amplitude e ZCR são utilizados como modelos de mistura gaussiana (GMM) e o melhor desempenho para discriminação de voz é alcançado pelos MFCCs com seu derivado de primeiro ordem. De acordo com [Liu et al. 2007], o MFCC, a Codificação Preditiva Linear (LPC, *do inglês Linear Prediction Coding*), ZCR, Pares de Linhas Espectrais (LSPs, *do inglês Line Spectral Pairs*), os recursos espectroscópicos, *rolloff* e flux são usados para a classificação de música e fala. No presente estudo, para caracterizar melhor o fluxo do sinal de voz, esses parâmetros são extraídos de um sinal de voz e são usados no modelo proposto. Para isso, utilizamos os sinais de voz disponibilizadas em 3 diferentes bases de dados, nos quais são aplicadas diferentes valores de taxas de perda de pacotes e modelos de distribuição de essas perdas.

Recentemente, a abordagem de aprendizagem não supervisionada tem sido utilizada em diversas aplicações [Andreoni Lopez et al. 2017], como no reconhecimento de voz. Seu paradigma geralmente visa construir representações de entrada que podem ser usadas para a predição e classificação de dados. A classificação pode ser realizada por agrupamento, estimativa de densidade e Análise de Componentes Principais, usando o estudo de componentes independentes.

A quantificação vetorial fornece entradas discretas e é uma aplicação precoce de aprendizagem não supervisionada para análise de áudio [Räsänen et al. 2009]. Os modelos de treinamento não supervisionados funcionam com modelos iniciais, representando pequenas quantidades de dados transcritos, nos quais o modelo é usado para decodificar quantidades maiores de dados não transcritos. Assim, novos modelos são treinados usando parte ou todos esses dados rotulados automaticamente.

Os avanços do hardware dos computadores permitiram o desenvolvimento de DNN, que contém muitas camadas de unidades ocultas não-lineares e muitas camadas de saída. O DNN pode ser implementado usando técnicas não supervisionadas, semi-supervisionadas ou supervisionadas. O RBM, constituído basicamente de unidades visíveis e escondidas, pode aprender diversas características de discriminação para um

determinado problema [Pan et al. 2014], e eventualmente, melhorar o custo computacional e o tempo necessário para completar o processo de treinamento. A ideia fundamental é alimentar a rede com exemplos não classificados e depois reconstruir os dados de entrada. O trabalho de [Hinton et al. 2006] destaca o uso da Divergência Contrastiva (CD, *do inglês Contrastive Divergence*) como um método comumente usado para o aprendizado em RBMs, devido à sua eficiência e resultados confiáveis. O CD pretende ajustar os valores de entrada no modelo, trabalhando a aproximação da aprendizagem de máxima verossimilhança.

De acordo com [Jaitly and Hinton 2011], o RBM pode ser utilizado para modelar fragmentos de um sinal de voz. Geralmente, o RBM vem sendo empregado para a aprendizagem não supervisionada. Porém, o seu uso na aprendizagem supervisionada foi proposto pelo algoritmo de Máquina de Boltzmann Restrita Discriminativa (DRBM). A ideia do aprendizado supervisionado é incorporar a informação do rótulo/classe na camada visível (entrada) e, assim, calcular a distribuição conjunta dos dados de entrada e do rótulo correspondente de forma discriminatória, ou seja, calcular a probabilidade de cada classe ter determinada amostra.

Portanto, pretende-se utilizar o algoritmo DRBM para predizer uma qualidade de voz em cenários de perda de pacotes. Outros algoritmos, tais como o SVM sem uso de DL, serão utilizados com a finalidade de comparar o seus desempenhos com o modelo proposto. Em seguida o algoritmo DRBM é descrito.

2.3. Máquina de Boltzmann Restrita Discriminativa

Um RBM é uma rede neural estocástica e é capaz de gerar dados de acordo com uma distribuição de probabilidade. Assim, as unidades de um RBM são modeladas como variáveis aleatórias com a distribuição de probabilidade conjunta de acordo com (1).

$$P(y,\ \mathbf{x},\ \mathbf{j}) = \frac{\exp(-E(y, \mathbf{x}, \mathbf{j}))}{Z} \tag{1}$$

onde, $x = [x_1, ..., x_{nd}]^T$ e $j = [h_1, ..., h_{nd}]^T$ são vetores de estado da entrada e de variáveis ocultas, respectivamente. $y \in 1, ..., n_c$ é o rótulo (classe) correspondente ao vetor de entrada, $E(y, \mathbf{x}, \mathbf{j})$ é conhecido como função de energia global e Z é uma variável escalar utilizada para garantir que a soma de $P(y,\ \mathbf{x},\ \mathbf{j})$ no seu domínio seja igual a 1.

Os RBMs são geralmente treinados usando o algoritmo de divergência contrastiva (CD) para minimizar a função de perda generativa k_{gen}. A função não considera a variável j, focando na perda generativa para posteriormente ser usada no RBM.

$$k_{gen} = -\sum_{t=1}^{a_t} \log P(y^{(t)},\ \mathbf{x}^{(t)}) \tag{2}$$

onde a_t é o número de amostras de treinamento e (x^t, y^t) representa a t-ésima amostra de treinamento, constituída pela entrada $x^{(t)}$ e sua respectiva classe $y^{(t)}$. A busca eficiente dos mínimos desta função envolve o cálculo de seu gradiente em relação aos parâmetros do modelo que é resolvido considerando certas aproximações, do algoritmo de CD.

O algoritmo DRBM é treinado para minimizar a função de perda discriminativa k_d.

$$k_d = -\sum_{t=1}^{n_t} \log P(y^{(t)}|\mathbf{x}^{(t)}) \tag{3}$$

No presente trabalho, a implementação do DRBM foi desenvolvida na linguagem *Python* e é baseada na biblioteca *Theano*. O processador usado é um Intel Xeon com 4 núcleos físicos, com uma frequência de 2,4 GHz e 8 GB de RAM.

3. Efeito da Taxa de Perda de Pacotes no Índice de Qualidade da Voz

Para estudar o efeito do PLR na qualidade do sinal de voz, foi construído uma base de dados de arquivos de voz. Para ter amostras de voz mais representativas, foram utilizados arquivos de três bases de dados, cujas principais características estão descritas na Tabela 1.

No total, foram utilizados 80 arquivos diferentes. Conforme mencionado anteriormente, nosso estudo é sobre o sinal da banda estreita; portanto, a taxa de amostragem utilizada foi de 8 kHz.

Tabela 1. Características dos arquivos originais de voz.

Base de dados de áudios	Número de Arquivos	Comprimento Min. / Max.	Silêncio Médio
ITU-T Rec. P. Sup. 23 [ITU-T Rec. Sup. 23 1998]	27	7 / 10 s.	46%
ITU-T Rec. P.862 [ITU-T Rec. P.862 2001]	20	8 / 10 s.	41%
ANITA [EADS Telecom 2003]	33	7 / 10 s.	29%

Diferentes modelos de distribuições de PLR são implementadas com base no modelo Gilbert-Elliot representado por (1):

$$prob = P(q_t = R|q_{t-1} = B) \qquad prob_2 = P(q_t = B|q_{t-1} = R) \tag{4}$$

Onde, $prob$ é a probabilidade de passar de um estado ruim (R) que indica a perda de pacotes para um estado bom (B) que representa a entrega de pacotes; $prob_2$ é a probabilidade de passar do estado B para o estado R; e q_t e q_{t-1} representam os estados nos instantes t e $t-1$, respectivamente. Assim, com a variação de $prob$ e $prob_2$ é possível obter distribuições de perdas de pacotes diferentes, para o mesmo valor PLR, que é determinado por:

$$PLR = \frac{prob}{prob + prob_2} \tag{5}$$

Foram utilizadas três modelos de distribuições de perdas de pacotes representadas como baixa, moderada e alta; de acordo com (4) e (5). Os valores de PLR variaram de 0,5% até 20% com passos de 0,5%. Foi considerado como cenário mais degradante, um PLR de 20%, pois valores de PLR maiores são pouco prováveis de acontecer em redes reais. Assim, foram utilizados 40 valores de PLR diferentes e 120 cenários de PLR que

foram aplicados em cada arquivo de voz original 100 vezes. No total, 960.000 arquivos degradados de voz foram criados e avaliados pelo algoritmo PESQ. Neste trabalho, cada pacote perdido de 10 ms de comprimento foi substituído por um segmento de silêncio e não é considerado um algoritmo de emascaramento de perdas de pacotes.

É importante salientar que o nosso método proposto considera o índice de qualidade, e não o valor de PLR, para classificar as amostras de voz. O PLR é uma função de probabilidade e o número de pacotes perdidos não é o mesmo em todos os testes, e também as perdas podem ocorrer durante um segmento de voz ou sem voz, então o índice MOS pode ser diferente para o mesmo PLR. Para provar isso, foram criados dois cenários de teste, o primeiro foi configurado para 80% das perdas atingirem apenas os segmentos de voz e o segundo foi configurado para 80% das perdas atingirem apenas os segmentos sem voz.

A Tabela 2 apresenta ambos os cenários, em que cada valor MOS, do PESQ, corresponde à média de 20 arquivos selecionados de [ITU-T Rec. P.862 2001]. Pode-se observar que no caso de perdas de pacotes afetarem o segmento de voz, o valor de MOS diminui drasticamente.

Tabela 2. Perdas de pacotes em segmentos sem voz ou com voz.

Valor de PLR (%)	MOS com PLR em segmentos sem voz	MOS com PLR em segmentos com voz
1	4,4	4,0
5	3,7	2,9
10	3,4	2,2
15	3,3	2,0
20	3,1	1,6

4. Método de Avaliação da Qualidade do Sinal de Voz Proposto

A solução proposta para classificar o sinal de voz de acordo com sua qualidade é representada na Figura 1. A solução pode ser resumida da seguinte maneira, em primeiro lugar, uma fase de treinamento é realizada no Servidor de Qualidade de Voz (SQV) e, como resultado, um modelo de CQV é estabelecido em relação a base de dados previamente construída.

O SQV está programado para receber chamadas telefônicas periodicamente a partir de um sistema automático gerenciado pelo provedor de serviços e usando a mesma infra-estrutura de rede. Portanto, a base de dados inicial pode ser atualizada com amostras de voz que contenham diferentes tipos de degradações que as consideradas inicialmente. É importante destacar que as amostras de áudio em determinados momentos da comunicação já servem como base de dados, onde será efetuada a classificação. O modelo de aprendizagem é treinado novamente, realizando a extração de parâmetros e obtendo um modelo CQV aprimorado. O novo treinamento é importante porque a rede pode apresentar diferentes degradações. Uma vez que o CQV é determinado, ele é enviado para um cliente em um modo *off-line*. O aplicativo do cliente, representado por 1 e 2, analisa o sinal de voz de entrada e usa o CQV para determinar a classe de qualidade do sinal, indicada por A, B, C, D e E; note que a saída CQV não é um valor do índice

MOS. Neste trabalho, o SQV foi modelado para NB, mas pode ser facilmente estendido para sistemas de comunicação SWB.

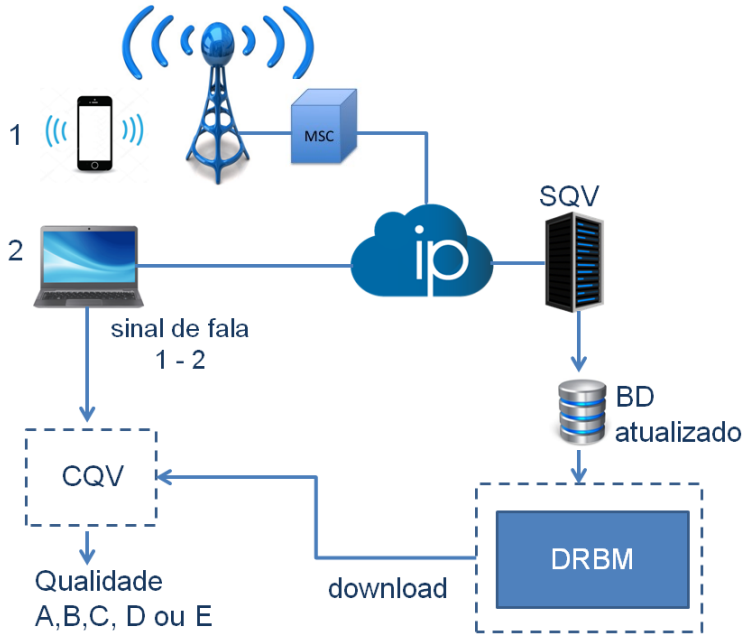

Figura 1. Esquema de rede do modelo da solução proposta para classificação de qualidade de voz.

Consideramos, neste trabalho, cinco classes de qualidade do sinal de voz com base na escala de qualidade de classificação de categoria absoluta de 5 pontos (ACR, *do inglês Absolute Category Rating*), a qual é descrita na recomendação ITU-T P.800. A Tabela 3 apresenta as classes propostas da qualidade do sinal de voz.

Tabela 3. Classes de qualidade do sinal de voz e seus valores de índice MOS

Classe de qualidade de voz	Qualidade percebida escala ACR	Valores de índice MOS
Classe-A	Excelente	5,00-4,00
Classe-B	Bom	3,99-3,00
Classe-C	Razoável	2,99-2,00
Classe-D	Ruim	1,99-1,00
Classe-E	Muito ruim	inferior a 1,00

Como afirmado anteriormente, o processo de treinamento ocorre no SQV, no qual os parâmetros dos arquivos que contem o sinal de voz são extraídos pelo DRBM, para ser em seguida classificados.

A Fig. 2 mostra as etapas envolvidas, no qual o DRBM efetua a extração de características do um sinal de voz, para logo classificar em uma das cinco classes de qualidade definidas na Tabela 2.

Quando as amostras do sinal de voz tenham comprimentos mais longos, elas podem ser divididas em segmentos de voz menores para serem treinadas no modelo de aprendizagem proposto.

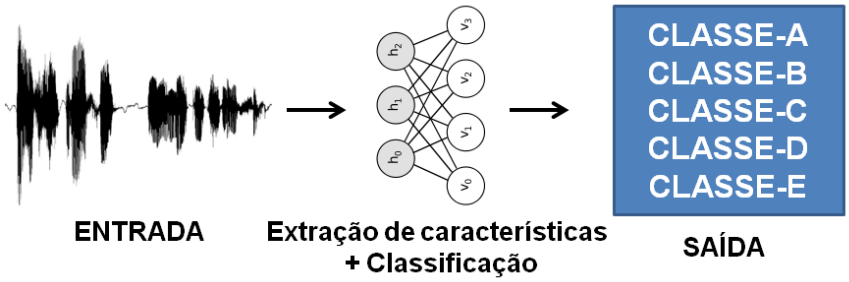

Figura 2. Etapas do modelo de classificador de qualidade da voz.

5. Avaliação Experimental e Resultados

Esta seção trata sobre a metodologia dos experimentos realizados e os respectivos resultados obtidos.

5.1. Topologia do modelo de aprendizagem

Nos experimentos, a topologia do DRBM apresenta uma taxa de aprendizado de 0,01, o treinamento foi efetuado até um máximo de 500 épocas, e os melhores resultados foram obtidos com 100 épocas. O DRBM foi testado com valores de 10, 25, 50, 100 e 200 de *hidden units*, porém a melhor desempenho alcançado foi com o valor de 50. Esta topologia é utilizada porque alcançou o melhor desempenho em relação a outras topologias testadas.

Os 80 arquivos contendo sinal de voz originais da base de dados são divididos aleatoriamente para as fases de treinamento e validação considerando 80% e 20%, respectivamente. Assim, 768.000 amostras de voz, criadas a partir de 64 amostras são usadas para treinamento e um conjunto de dados separado composto por 192.000 amostras de voz são utilizados para validação.

A extração de sessenta e três características do sinal de voz em quadros de 25 ms com sobreposição de 10 ms é realizada. Essas características são rotuladas com o classificador após passar pelo DRBM, que gerou os valores estimados para cada uma das amostras de voz degradadas. Os parâmetros extraídos correspondem ao centróide espectral, deslocamento espectral, fluxo espectral, vinte FFT *Power Spectrum*, ZCR e MFCCs que consideram 13 características estáticas do MFCC (12 MFCCs e *log energy*) e as derivadas de primeira e segunda ordem das características estáticas.

Adicionalmente, o algoritmo SVM também foi implementado, considerando os mesmos parâmetros, a fim de comparar seus valores de acurácia com os valores alcançados pelo DRBM.

5.2. Avaliação do desempenho do modelo de qualidade do sinal de voz proposto

Em testes preliminares, foram comparados diferentes classificadores, como o clássico SVM sem o uso do DRBM. Os resultados experimentais demonstraram que o DRBM apresentou melhores resultados, conforma mostra Tabela 4 e, portanto, foi utilizado no modelo CQV.

Os resultados da Tabela 4 apresentam as precisões das classes obtidas pelos classificadores SVM sem DRBM e o DRBM. Nessas experiencias, foram utilizados os cenários

de 0%, 5%, 10%, 15% e 20% de PLR nos sinais de voz utilizados como material de teste.

Tabela 4. Precisão do SVM comparado com o DRBM.

Packet Loss Rate (%)	SVM	DRBM
Classe-A	87,73	91,45
Classe-B	86,13	90,28
Classe-C	86,01	90,05
Classe-D	84,19	88,13
Classe-E	83,51	87,79

A Tabela 5 apresenta os valores médios de precisão da avaliação de desempenho CQV e ITU-T P.563 para a avaliação da classe de qualidade de voz nos testes de validação, usando o formato da matriz de confusão. Cada índice MOS da P.563 é atribuído a uma classe de qualidade de acordo com a Tabela 3, e a saída CQV é o classificador de qualidade.

Tabela 5. Matriz de Confusão para Predição de Classe de Qualidade do sinal de Voz (em Porcentagem) usando Algoritmo CQV e P.563

Classe de Qualidade do sinal	CQV / P.563 Classe-A	CQV / P.563 Classe-B	CQV / P.563 Classe-C	CQV / P.563 Classe-D	CQV / P.563 Classe-E
Classe-A	91,45 / 44,36	8,55 / 18,85	0,0 / 32,20	0,0 / 4,59	0,0 / 0,0
Classe-B	5,11 / 0,30	90,28 / 57,17	4,61 / 33,25	0,0 / 9,28	0,0 / 0,0
Classe-C	0,0 /0,69	6,93 / 3,87	90,05 / 84,42	3,02 / 11,02	0,0 / 0,0
Classe-D	0,0 / 0,0	0,0 / 0,0	2,74 / 2,90	88,13 / 85,89	9,13 / 11,21
Classe-E	0,0 / 0,0	0,0 / 0,00	0,0 / 6,65	12,21/ 10,33	87,79 / 83,02

Além disso, testes subjetivos de avaliação de qualidade do sinal de voz foram realizados em um ambiente controlado. No total, 42 voluntários participaram dos testes subjetivos, que consistiam de 17 mulheres e 25 homens, com idade entre 18 e 52 anos. Todos os avaliadores relataram não ter experiência em testes de qualidade de sinal de voz. Todas as avaliações foram realizadas no mesmo local durante um período de 7 semanas.

O material de teste foi composto por 10 arquivos originais de sinal de voz para cada classe de qualidade e os valores de qualidade desses arquivos de teste foram distribuídos homogeneamente para cobrir o intervalo de qualidade de cada classe; dessa forma, se garante a existência de arquivos com valores MOS próximos aos limiares de cada classe.

Cada arquivo de áudio foi avaliado e classificado por pelo menos 15 voluntários. A escala de qualidade utilizada nos testes foi a mesma apresentado na Tabela 3. O comprimento de cada arquivo de áudio era de 16 segundos. Adotou-se arquivos de 16 segundos para verificar a eficiência do modelo na identificação de arquivos de maior duração. Os resultados experimentais demonstraram que a proposta com uso do CQV atingiu uma precisão de 91,45 % em relação a Classe A.

Finalmente, para demonstrar a importância de considerar diferentes modelos de distribuições de PLR, dois cenários de testes adicionais foram implementados. Em primeiro lugar, a fase de treinamento foi realizada com apenas uma distribuição de PLR (TR-A), em seguida, uma segunda distribuição de PLR foi incorporada (TR-B) e o modelo foi

treinado novamente. O número de arquivos para validação foi de 20 %, diferentemente dos arquivos utilizados na fase de treinamento.

A Tabela 6 apresenta o desempenho alcançado pelo CQV proposto em cada cenário, com os valores médios de precisão, utilizando a matriz de confusão.

Tabela 6. Matriz de confusão do desempenho do CQV (em porcentagem) considerando os dois modelos de distribuição PLR

Classe de Qual. do sinal de voz	TR-B / TR-A Classe-A	TR-B / TR-A Classe-B	TR-B / TR-A Classe-C	TR-B / TR-A Classe-D	TR-B / TR-A Classe-E
Classe-A	88,13 / 57,11	8,55 / 13,78	3,32 / 29,11	0,0 / 0,00	0,0 / 0,0
Classe-B	6,78 / 0,30	87,45 / 63,29	5,77 / 36,41	0,0 / 2,11	0,0 / 0,0
Classe-C	0,0 /0,0	9,13 / 2,91	87,35 / 82,21	3,52 / 14,88	0,0 / 0,0
Classe-D	0,0 / 0,0	0,0 / 0,0	5,63 / 3,22	86,18 / 82,33	8,19 / 14,45
Classe-E	0,0 / 0,0	0,0 / 0,0	0,0 / 2,57	14,33/ 15,99	85,67 / 81,44

Os resultados apresentados na Tabela 5 e 6 destacam a utilidade e adaptabilidade da solução apresentada na Figura 1, em que o CQV é atualizado com novas amostras capturadas pelo SQV que melhor representa o status atual da rede, pela variabilidade que pode acontecer com os modelos de distribuição de PLR.

6. Conclusão

Os resultados dos testes preliminares enfatizam que o PLR nem sempre é correlacionado com o índice MOS, porque, as perdas de pacotes podem alcançar segmentos com ou sem voz. Além disso, o mesmo valor de PLR, com diferentes modelos de distribuição, impacta a qualidade da voz de maneira diferente. Com base nesses resultados, uma base de dados composta de 960.000 arquivos de fala foi construída, a qual considera diferentes valores e modelos de PLR. Cada arquivo de voz é avaliado pelo algoritmo PESQ e classificado em uma das cinco classes de qualidade definidas. O CQV é baseado no DRBM que extrai características dos sinais analisados. Os resultados demonstraram o alto desempenho do CQV proposto, atingindo uma classificação de 91,45% a 87,79 % de precisão no teste de validação superando notoriamente os resultados obtidos pelo algoritmo P.563. Além disso, testes subjetivos foram realizados para avaliar mais 40 amostras de voz, nos quais o CQV proposto atingiu 91 % de precisão. Além disso, a arquitetura de rede que inclui o CQV foi apresentada destacando-se a sua adaptabilidade a mudanças de comportamento da rede. Por fim, a solução apresenta escalabilidade e potencial de aplicabilidade para auxiliar prestadores de serviços de telefonia em tarefas de administração de redes, tendo como objetivo principal a melhora da QoE do usuário final.

7. Agradecimentos

Este trabalho possui o suporte da Fundacao de Amparo a Pesquisa do Estado de Minas Gerais (FAPEMIG).

Referências

Andreoni Lopez, M., Lobato, A., Mattos, D., Alvarenga, I. B., Duarte, O. C., and Pujolle, G. (2017). Um Algoritmo Nao Supervisionado e Rápido para Selecao de Caracteristicas em Classificacao de Tráfego . In *Simposio Brasileiro de Redes de Computadores e Sistemas Distribuidos*, pages 573–586, Belem,Para.

Bengio, Y., Chapados, N., Delalleau, O., Larochelle, H., Saint-Mleux, X., Hudon, C., and Louradour, J. (2012). Detonation classification from acoustic signature with the restricted boltzmann machine. *Computational Intelligence*, 28(2):261–288.

Chen, C. L. P., Zhang, C. Y., Chen, L., and Gan, M. (2015). Fuzzy restricted boltzmann machine for the enhancement of deep learning. *IEEE Trans. on Fuzzy Systems*, 23(6):2163–2173.

Cremonezi, B. M., Vieira, A. B., Nogueira, M., and Nacif, J. A. M. (2017). Um protocolo de alocacao dinamica de canais para ambientes medicos sob multiplas estacoes base. In *Simposio Brasileiro de Redes de Computadores e Sistemas Distribuidos*, pages 272–285, Belem,Para.

EADS Telecom (2003). Audio enhancement in telecom. applications: Anita reference database description.

Graves, A. (2012). *Supervised Sequence Labelling with Recurrent Neural Networks*, volume 385 of *Studies in Computational Intelligence*. Springer.

Hinton, G., Deng, L., Yu, D., Dahl, G., rahman Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., and Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition. *Signal Processing Magazine*.

Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computing*, 18(7):1527–1554.

ITU-T Rec. G.107 (2015). The E-model: a computational model for use in transmission planning.

ITU-T Rec. P.800 (1996). Methods for subjective determination of transmission quality.

ITU-T Rec. P.862 (2001). Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs.

ITU-T Rec. P.863 (2014). Perceptual objective listening quality assessment (POLQA).

ITU-T Rec. Sup. 23 (1998). Coded-speech database.

ITU-T Temporary Document (2015). Technical requirement specification proposals for scope of single-ended perceptual evaluation of listening quality (P.SPELQ).

Jaitly, N. and Hinton, G. E. (2011). Learning a better representation of speech soundwaves using restricted boltzmann machines. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal*, pages 5884–5887, Prague, Czech Republic.

Lee, B. K. and Chang, J. H. (2016). Packet loss concealment based on deep neural networks for digital speech transmission. *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 24(2):378–387.

Liu, C., Xie, L., and Meng, H. (2007). Classification of music and speech in mandarin news broadcasts. In *National Conf. on Man-Machine Speech Communication*, pages 17–20, Anhui, China.

Monika, S. and Rama, A. (2016). An efficient digital speech transmission using neural network with HMM (Hidden Markov Model). In *Proc. Int. Conf. on Emerging Engineering Trends and Science*, pages 34–43, Tamilnadu, India.

Montag, C., Błaszkiewicz, K., Sariyska, R., Lachmann, B., Andone, I., Trendafilov, B., Eibes, M., and Markowetz, A. (2015). Smartphone usage in the 21st century: who is active on whatsapp? *BMC Research Notes*, 8(1):331–336.

Pan, G., Qiao, J., Chai, W., and Dimopoulos, N. (2014). An improved RBM based on bayesian regularization. In *Proc. Int. Joint Conf. on Neural Networks*, pages 2935–2939, Beijing, China.

Polacky, J. and Pocta, P. (2014). An analysis of the impact of packet loss, codecs and type of voice on internal parameters of p.563 model. In *Proc. IEEE Int. Conf. on Digital Technlogies*, pages 281–284, Zilina, Slovakia.

Räsänen, O. J., Laine, U. K., and Altosaar, T. (2009). Self-learning vector quantization for pattern discovery from speech. In *INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009*, pages 852–855.

Rodríguez, D. Z., Rosa, R. L., Alfaia, E. C., Abrahão, J. I., and Bressan, G. (2016). Video quality metric for streaming service using DASH standard. *TBC*, 62(3):628–639.

Rodríguez, D. Z., Wang, Z., Rosa, R. L., and Bressan, G. (2014). The impact of video-quality-level switching on user quality of experience in dynamic adaptive streaming over HTTP. *EURASIP J. Wireless Comm. and Networking*, 2014:216–226.

Saini, P. and Kaur, P. (2013). Automatic speech recognition: A review. *International journal of Engineering Trends & Technology*, pages 132–136.