RODRIGO DANTAS NUNES

ALGORITMO PARA MELHORAR O DESEMPENHO DE UMA MÉTRICA NÃO INTRUSIVA DE QUALIDADE DE VOZ

LAVRAS – MG

2017

RODRIGO DANTAS NUNES

ALGORITMO PARA MELHORAR O DESEMPENHO DE UMA MÉTRICA NÃO INTRUSIVA DE QUALIDADE DE VOZ

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 Engenharia de Sistemas e Automação, para a obtenção do título de Mestre.

Prof. DSc. Demóstenes Zegarra Rodríguez
Orientador

LAVRAS – MG
2017

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

Nunes, Rodrigo Dantas

Algoritmo para melhorar o desempenho de uma métrica não intrusiva de qualidade de voz / Rodrigo Dantas Nunes. – Lavras : UFLA, 2017.

70 p. : il.

Orientador: Prof. DSc. Demóstenes Zegarra Rodríguez.

Dissertação(mestrado acadêmico)–Universidade Federal de Lavras, 2017.

Bibliografia.

1. Avaliação da Qualidade de Voz 2. Qualidade de Experiência. 3. P. 563. I. Rodríguez, Demóstenes Zegarra. .II. Título.

RODRIGO DANTAS NUNES

ALGORITMO PARA MELHORAR O DESEMPENHO DE UMA MÉTRICA NÃO INTRUSIVA DE QUALIDADE DE VOZ
ALGORITHM TO IMPROVE THE PERFORMANCE OF A NONINTRUSIVE VOICE QUALITY METRIC IN REAL-TIME

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 Engenharia de Sistemas e Automação, para a obtenção do título de Mestre.

APROVADA em 28 de Março de 2017.

Profa. DSc. Renata Lopes Rosa	UFLA
Prof. DSc. Luiz Henrique Andrade Correia	UFLA
Prof. DSc. Rodrigo Tomás Nogueira Cardoso	CEFET-MG

Prof. DSc. Demóstenes Zegarra Rodríguez
Orientador

LAVRAS – MG
2017

À minha mãe, Regina e à minha finada avó, Léa.
À minha amada esposa, Fernanda e à minha filha, Maria Clara.
Aos meus irmãos, Marciel e José Geraldo.

AGRADECIMENTOS

A Deus, por todas as oportunidades as quais me permite desfrutar todos os dias.

À minha esposa e companheira de vida, Fernanda; à minha filha, Maria Clara e a minha querida mãe, pela paciência e compreensão por minha ausência, fruto das demandas do curso.

Ao professor e meu orientador, Demóstenes Zegarra Rodríguez, por acreditar em mim, pelos seus ensinamentos, paciência e compreensão, estando sempre presente durante as etapas do meu processo de formação.

Aos meus amigos e colegas de curso, Diogo Aranha Ribeiro e Rodrigo de Lima Cunha, pelas inúmeras assistências ao longo do curso.

Ao Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG) pelo suporte financeiro, oferecido por meio do Programa Institucional de Ajuda de Custo à Capacitação de Servidores.

Aos meus amigos e colegas de trabalho do Escritório de Projetos do CEFET-MG, pelo apoio e conselhos, em especial ao Cristiano Fraga Guimarães Nunes, pelo suporte matemático.

Ao Ulisses Cotta Cavalca, Secretário de Governança da Informação no CEFET-MG, pelo apoio e pela condição de trabalho flexibilizado, o que me possibilitou cumprir as disciplinas.

À Universidade Federal de Lavras (UFLA), em especial ao Programa de Pós-Graduação em Engenharia de Sistemas e Automação (PPGESISA), e ao Departamento de Engenharia (DEG), pela oportunidade de crescimento acadêmico.

À Fernanda, secretária do PPGESISA, pela disponibilidade nas diversas vezes em que foi solicitada.

A todos os professores do PPGESISA, pelas sugestões e ajuda no trabalho desenvolvido.

A responsabilidade é uma atitude mental de não esperar que venha de fora a solução para o problema, mas contar consigo mesmo e com a sua atuação para a resolução das situações que se criarem.
(Djalma D. P. R. Oliveira)

RESUMO

Os sistemas telefônicos são uma importante ferramenta de comunicação, nos dias atuais e, como tal, têm sua qualidade constantemente questionada. A fim de prover mecanismos de avaliação, o Setor de Normalização em Telecomunicações da União Internacional das Telecomunicações (ITU-T) estabelece recomendações técnicas para determinar as métricas de avaliação da qualidade de experiência (QoE). Dentre elas se destacam as recomendações ITU-T P.862 e P.563, que por meio da Pontuação Média de Opinião (*Mean Opinion Score*, MOS), atribuem um valor para a amostra de áudio sendo avaliada. Desse modo, neste trabalho apresenta-se uma solução para avaliar a qualidade de voz em serviços de telecomunicações em tempo real, o que é útil para assinantes, provedores de serviços móveis (PSM) e autoridades reguladoras de telecomunicações (ART). O principal objetivo, nesta pesquisa, foi fornecer um modelo que melhore o desempenho do algoritmo proposto na Rec. ITU-T P.563. Para atingir esse objetivo, o modelo proposto considera dois fatores: uma melhor resposta em condições de rede com perda de pacotes e um tratamento mais adequado de segmentos de silêncio presentes no sinal de áudio. Assim, o modelo proposto dá origem a um Servidor de Qualidade de Voz (SQV) que é implementado em um serviço *Web*, a fim de fornecer condições para as ARTs e PSMs de monitorarem o estado da rede, especificamente os valores de índice MOS das chamadas de Voz Sobre IP (VoIP) estabelecidas. Adicionalmente, essa solução é implementada em formato de aplicativo para dispositivos móveis e denominada SQV_{mobile}. Tal abordagem permite, então, a realização de chamadas e a devida apresentação do valor de índice MOS ao usuário. Os resultados experimentais mostram que o desempenho do algoritmo da Rec. ITU-T P.563 foi realmente melhorado pelo modelo proposto, aproximando seus resultados aos valores da Rec. ITU-T P.862, atingindo um coeficiente de correlação de Pearson (*Pearson Correlation Coefficient*, PCC) de 0,9957 e um erro quadrático médio da raiz (*Root Mean Square Error*, RMSE) de 0,2983. Adicionalmente, o processamento e o consumo de energia correspondentes à aplicação são irrelevantes nos dispositivos móveis atuais. Além disso, a solução proposta fornece uma ferramenta importante para que os usuários monitorem a qualidade dos serviços contratados com as PSMs.

Palavras-chave: 1. Avaliação da Qualidade de Voz 2. Qualidade de Experiência. 3. P. 563. 4. Voip. 5. Perda de Pacotes.

ABSTRACT

Telephone systems are an important communication tool nowadays, and as such, have their quality constantly questioned. In order to provide evaluation mechanisms, the Telecommunication Standardization Sector of the International Telecommunication Union (ITU-T) establishes technical recommendations for determining the measure metrics of the quality of experience (QoE).The ITU-T recommendations P.862 and P.563 stand out among the others, and through the Mean Opinion Score (MOS) they assign a value for the audio sample being evaluated. Thus, the work introduces a solution for assessing voice quality in real-time telecommunication services, which is useful for subscribers, mobile service providers (MSP), and telecommunication regulatory authorities (TRA). The main objective of this research is to provide a model that improves the performance of the algorithm available in the ITU-T Rec. P.563. To accomplish this objective, the proposed model considers two factors: a better response in network conditions with packet loss and a more adequate treatment of silences segments present into the audio signal. Thus, the proposed model gives rise to a Voice Quality Server (VQS) that is implemented in a Web Service in order to provide conditions for the TRAs and MSPs to monitor the state of the network, specifically the values of the MOS index of the established Voice Over IP (VoIP) calls. In addition, this solution is implemented in mobile application format and is called VQS_{mobile}. Such an approach then allows for the making of calls and the proper presentation of the MOS index value to the user. The experimental results show that the P.563 algorithm performance was really improved by the proposed model, approximating its results to those given by ITU-T Rec. P.862 algorithm, reaching a Pearson correlation coefficient (PCC) of 0.9957 and a root-mean square error (RMSE) of 0.2983. Moreover, the processing and energy consumption corresponding to the application are almost negligible in current mobile devices. In addition, the proposed solution provides an important tool for users to monitor the quality of the services contracted with the MSP.

Keywords: 1. Speech Quality Evaluation. 2. Quality of Experience. 3. P. 563. 4. Voip. 5. Packet Loss.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O Setor de Normalização em Telecomunicações da União Internacional das Telecomunicações (ITU-T) é responsável pelo desenvolvimento de recomendações técnicas no domínio das telecomunicações. Estas recomendações descrevem padrões para várias finalidades, como, por exemplo, metodologias para avaliação da qualidade de sinais de voz e vídeo.

As recomendações da série P da ITU-T definem os estimadores de qualidade de voz, cujo objetivo é pontuar a qualidade das comunicações. Além disso, esses estimadores são de grande importância para o monitoramento da Qualidade de Experiência (*Quality of Experience*, QoE) vivida pelos usuários, que realizam chamadas de voz sobre IP (VoIP), sejam pelas operadoras de serviço telefônico ou diretamente pela Internet.

Segundo o *Institute of Electrical and Electronic Engineers* (IEEE, 2015), a QoE é o grau de satisfação ou de aborrecimento de um usuário de uma aplicação ou serviço que resulte do cumprimento de suas expectativas em relação à utilidade e/ou gozo da aplicação ou serviço, tendo em conta a personalidade do usuário e o seu estado atual de humor.

Os métodos utilizados para avaliar a qualidade de voz são divididos em dois grupos principais: subjetivo e objetivo. Os subjetivos envolvem indivíduos e são de alto custo por exigirem infraestrutura apropriada de laboratório e também pelo tempo gasto na realização dos testes. Esses métodos baseiam-se em audições e são descritos nas recomendações ITU-T P.800 (ITU-T, 1996) e ITU-T BS.1534 (ITU-T, 2015).

Em contrapartida, os métodos objetivos utilizam algoritmos e tentam predizer, de forma aproximada, a pontuação que seria dada em testes subjetivos por indivíduos. Além disso, as métricas objetivas são subdivididas em dois tipos, os métodos não intrusivos e os intrusivos.

Para avaliar a qualidade do sinal, os métodos intrusivos, também chamados de *double-ended*, precisam do sinal original como referência para comparar com o sinal degradado no ponto de avaliação. Para classificar as amostras, ambos os tipos de métodos usam um índice de qualidade de voz, mais conhecido como *Mean Opinion Score* (MOS). Assim, os resultados dos métodos intrusivos são considerados mais confiáveis, sendo a Rec. ITU-T P.862 (ITU-T, 2001) e sua sucessora, a Rec. ITU-T P.863 (ITU-T, 2014), as mais aceitas. Por outro lado, os métodos não intrusivos, conhecidos como *single-ended*, só precisam do sinal no ponto final, ou em um dado ponto onde o sinal deva ser avaliado. A Rec. ITU-T P.563 (ITU-T, 2004) é a mais utilizada entre os métodos objetivos não intrusivos.

Dessa forma, considerando que a comunicação de voz é um serviço de tempo real, o sinal de referência, geralmente, não está disponível para avaliação no ponto final e, por esse motivo, a forma mais adequada de avaliar os serviços de comunicação de voz é por meio da utilização de métodos não intrusivos, como a Rec. ITU-T P.563.

Alguns trabalhos, tais como os de Dubey e Kumar (2015), Mossavat, Petkov e Klejin (2011) e a própria Rec. ITU-T P.563 citam que essa métrica é considerada a solução mais adequada para a avaliação de QoE em cenários de comunicação de voz.

Entretanto, nenhum método de avaliação de voz é atualmente disponibilizado em serviços de comunicação comercial para seus usuários, os quais acabam não tendo essas informações relativas ao desempenho do serviço utilizado.

Além disso, a Rec. ITU-T P.563 tem duas principais limitações: uma diz respeito à perda de pacotes em redes de Protocolo de Internet (*Internet Protocol*, IP), como relatado nos trabalhos de Cherif et al. (2012), Dubey e Kumar (2015), Falk e Chan (2006) e Raja et al. (2007). A outra limitação está relacionada ao processamento de segmentos de silêncios naturais da fala humana na conversação, fator este não identificado na literatura como objeto de estudo.

Deve-se observar que esta última limitação ocorre devido ao fato de que, no algoritmo da Rec. ITU-T P.563, apenas enunciados de voz separados por blocos inferiores a 200 milissegundos são unidos por seu algoritmo Detector de Atividade de Voz (*Voice Activity Detector*, VAD). Portanto, se o usuário em uma chamada, por exemplo, parar de falar por um momento durante a conversa, o silêncio naturalmente produzido ali será classificado na avaliação como uma interrupção do sinal e o algoritmo da Rec. ITU-T P.563 atribuirá uma pontuação baixa.

Sendo assim, a hipótese a ser confirmada por testes com arquivos degradados consiste em verificar se é possível melhorar o algoritmo da Rec. ITU-T P.563, de forma que os valores de índices MOS sejam mais próximos dos valores dados pela Rec. ITU-T P.862, quando aplicados em ambientes VoIP, e considerando o fator dos silêncios naturais da conversação humana durante a fala. Este processo será realizado com base na inferência de modelos matemáticos para os dois cenários.

1.1 Objetivo

O objetivo, neste trabalho, é determinar um algoritmo que melhore o desempenho do algoritmo da Rec. ITU-T P.563 em cenários com perda de pacotes em redes IP, e em avaliações de amostras com silêncios naturais da fala.

Dessa forma, é desejável que os valores de índice MOS sejam mais próximos dos valores dados pelo algoritmo da Rec. ITU-T P.862 que, por ser um método objetivo intrusivo, é considerado uma referência aos métodos objetivos não intrusivos.

Adicionalmente, o algoritmo será implementado em dois formatos, denominados Servidor de Qualidade de Voz (SQV) e Servidor de Qualidade de Voz Móvel (SQV_{mobile}), sendo o primeiro em um *Web Service*, que permitirá o acesso por qualquer aplicação cliente, seja ela desenvolvida na plataforma *desktop*, *web* ou móvel, e o segundo, diretamente em um dispositivo móvel, por meio de um aplicativo *Android*.

O *Web Service* criado poderá também viabilizar o monitoramento em tempo real da qualidade do sinal de voz das comunicações por parte dos provedores de serviços móveis (PSM) e autoridades reguladoras de telecomunicações (ART), no caso específico do Brasil, a Agência Nacional de Telecomunicações (ANATEL), como forma de garantir o atendimento de suas normas de qualidade.

1.1.1 Objetivos específicos

Especificamente, os objetivos, neste trabalho, foram os seguintes:

1. criar um modelo matemático para melhorar o algoritmo da Rec. ITU-T P.563 com base em um cenário de rede que emule perda de pacotes;

2. desenvolver algoritmo de VAD externo ao algoritmo da Rec. ITU-T P.563, a fim de aperfeiçoar a avaliação de silêncios naturais da fala;

3. implementar o algoritmo proposto em uma aplicação *Android* para dispositivos móveis e em um *Web Service*, e avaliar o consumo de processamento requerido no dispositivo móvel.

1.2 Justificativa

Segundo Meirelles (2016), espera-se que o número de *smartphones* em operação no Brasil chegue a um total de 236 milhões, em 2018. Adicionalmente, o autor cita, para efeito de comparação, que, em 2010, havia apenas um dispositivo para cada dois brasileiros.

Com relação ao uso destes dispositivos, houve um aumento de 276,1%, de 2014 para 2015, totalizando 9,9% dos acessos totais do serviço móvel pessoal (SMP) que utilizaram a

tecnologia *Long-Term Evolution* (LTE), conhecida comercialmente como quarta geração (4G) (BRASIL, 2016).

Considerando o crescimento do número de dispositivos em operação, e também do uso da tecnologia 4G, todo esforço científico a fim de prover recursos que possibilitem o monitoramento da rede é válido e muito importante. Além disso, a solução proposta aqui poderia viabilizar às ARTs um melhor controle da qualidade dos serviços das operadoras, e o atendimento aos seus níveis mínimos de qualidade do serviço prestado por essas empresas.

Deve-se observar, também, que as métricas não intrusivas são as mais adequadas para avaliações de qualidade de voz em comunicações de tempo real, métricas essas utilizadas como base do algoritmo proposto.

1.3 Organização do trabalho

Este trabalho está dividido em três partes, das quais a primeira parte traz o referencial teórico, em que se abordam os diversos temas relacionados a esta dissertação. Na segunda parte apresenta-se a metodologia utilizada, bem como as diversas ferramentas que auxiliaram o desenvolvimento do algoritmo proposto. Na terceira parte mostram-se os resultados obtidos e uma discussão sobre eles.

2 REFERENCIAL TEÓRICO

Nesta seção são apresentados os conceitos, as tecnologias, as recomendações técnicas e o estado da arte dos algoritmos de avaliação da qualidade de voz. Os tópicos principais estão divididos em qualidade de experiência, serviços de comunicação, algoritmos de avaliação de qualidade de voz, algoritmo de detecção de atividade de voz, *Web Service*, *Android* e, finalmente, o estado da arte, nos quais são apresentados outros trabalhos relacionados com a solução proposta neste documento.

2.1 Qualidade de experiência

De acordo com Laghari e Connelly (2012), a noção de QoE tem sido, nos tempos atuais, o principal tema de pesquisa da comunidade científica relacionada a telecomunicações. Deve-se destacar que, apesar dessa constatação, 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 produto suficientemente bom para o valor que está pagando, ele pode abandonar o provedor do serviço em questão (FIEDLER et al., 2010).

2.1.1 Definição

De acordo com Tsiaras et al. (2014), as métricas de qualidade de serviço (*quality of service*, QoS) são utilizadas para avaliar a qualidade percebida dos serviços entregues, sob a perspectiva dos operadores de rede. Entretanto, essas métricas não são adequadas para avaliar a experiência dos usuários finais, pois ela deve ser quantificada com base em atividades tais como velocidade de carregamento de uma página *web*, qualidade de vídeo ou voz em um serviço de VoIP, dentre outros. E, por isso, é importante distinguir entre o conceito de QoE e QoS.

Em concordância com essa afirmação, Mellouk, Hoceini e Tran (2013) citam que a QoE tem sido, nos dias atuais, um critério decisivo para o sucesso do gerenciamento de redes de telecomunicações, e que a percepção do usuário deve ser a base para o desenvolvimento dos serviços futuros neste setor.

Antons et al. (2015) afirmam também que a definição de QoE tem provocado um interesse crescente pela relação entre emoções, características de qualidade e percepção de qualidade.

Dessa forma, os resultados da avaliação de QoE são apresentados como um valor escalar simples, tipicamente usando o índice MOS da ITU-T P.800 (ITU-T, 1996), que, embora útil, suas limitações são evidentes para várias aplicações por utilizar somente a média aritmética (HOBFELD; HEEGAARD; VARELA, 2015). Este fato também é criticado em Antons et al. (2015), pois a média como é dada impede um provedor de serviço de ter um número real de quantos usuários estão insatisfeitos. Mas, apesar das limitações indicadas pelos autores, esta é a forma estabelecida pela ITU-T de se avaliar QoE em cenários de chamadas de voz.

2.1.2 Fatores que influenciam a QoE e as suas áreas de aplicação

O que faz com que o conceito de QoE seja um pouco evasivo é que os parâmetros de avaliação que o definem podem diferir de serviço para serviço. Por exemplo, se determinada operadora telefônica oferece um bom serviço de ligações nacionais, mas seu serviço de ligações internacionais for ruim, os usuários que não efetuam ligações para outros países terão uma qualidade de experiência melhor, e, por isso, mais satisfeitos que os demais (KUIPERS et al., 2010).

Os parâmetros de avaliação podem ser classificados em três grupos que são:

1. a qualidade do conteúdo do vídeo/áudio na origem;
2. QoS, a qual se refere à entrega do conteúdo na rede;
3. percepção humana, que inclui expectativas, ambiente, estado emocional, etc.

Deste modo, a qualidade do conteúdo refere-se ao tipo de Codificador/Decodificador (*CODEC*) utilizado, à taxa de *bits* e de amostragem, etc. Os parâmetros da QoS que afetam o desempenho dos serviços de *streaming* são, na maioria das vezes, a largura de banda, os atrasos, o *jitter* e a perda de pacotes (KUIPERS et al., 2010). Adicionalmente, o autor cita que os dois primeiros parâmetros de avaliação são relativamente fáceis de se quantificar, enquanto a última não é, pois trata-se da percepção humana.

De forma geral, e em uma visão mais ampla, há três possíveis tipos de metodologia para a avaliação da qualidade do sinal avaliado. O primeiro tipo consiste num modelo sem referência

que não conhece o arquivo original; o segundo é um modelo de referência reduzida que conhece de forma limitada o arquivo original e tenta combinar isto com avaliações em tempo real para conseguir predizer o MOS, e o terceiro é o modelo de referência completa que tem acesso ao arquivo original (KUIPERS et al., 2010).

Dessa forma, classifica-se a Rec. ITU-T P.563 como sendo um exemplo do primeiro tipo de metodologia e as recomendações ITU-T P.862 e P.863 do terceiro tipo de metodologia.

Apesar das soluções existentes para se predizer o índice de satisfação do usuário, Antons et al. (2015) afirmam que, até a presente data, não está claro como, e sob quais circunstâncias, os aspectos técnicos de qualidade, tais como artefatos de compresão de vídeo, interrupções, etc., podem influenciar o estado emocional dos usuários e, assim, suas avaliações de QoE. Além disso, também não é sabido como a conotação emocional do próprio conteúdo experimentado, em conjunto com outros fatores, pode influenciar a avaliação e o processo de percepção da qualidade.

2.1.3 Diferença entre qualidade de serviço e acordo de nível de serviço

A Rec. ITU-T E.860 (ITU-T, 2002) descreve que a crescente competição, em parte favorecida pelos requisitos de performance impostos pelos usuários, tem produzido uma grande pressão sobre os provedores de redes e serviços. Estes últimos, após terem enfrentado redução de custos por vários anos, têm, nos dias atuais, tentado melhorar sua QoS, na tentativa de se diferenciar no mercado.

Além disso, a situação complicou-se com o crescimento da demanda global de serviços desta natureza e, por isso, o papel de todas as entidades que fazem parte da provisão destes serviços e seus parceiros deve ser formalmente descrito. O escopo deve estabelecer responsabilidades para cada provedor, além de garantir a QoS requerida pelo usuário (ITU-T, 2002).

Esta recomendação relata também que, para formalizar as inter-relações entre as entidades, foi criada uma importante ferramenta, o acordo de nível de serviço (*Service Level Agreement*, SLA), que é um acordo formal entre duas ou mais partes, elaborado após negociação, a fim de estabelecer as regras para se avaliar as características do serviço, as responsabilidades e as prioridades de cada parte envolvida.

Por fim, a Rec. ITU-T E.860 estabelece que, dentro da estrutura da SLA, deve haver um elemento denominado **"Acordo de QoS"**, no qual devem estar incluídas regras formais

entre duas partes para monitorar, medir e decidir os parâmetros de QoS para que o objetivo seja atingido e, no final, o usuário esteja satisfeito com o serviço prestado.

2.2 Serviços de comunicação

A seguir faz-se uma abordagem sobre os tipos de serviços de comunicação relevantes para este trabalho e outros aspectos que os afetam diretamente, tais como a codificação/decodificação de voz e as formas de comutação.

2.2.1 Codificação e decodificação de voz

Segundo Hartpence (2013), atualmente, conversações de voz e vídeo são capturadas do meio analógico, digitalizadas, transmitidas e convertidas de volta na outra extremidade, para que o receptor possa entender o conteúdo transmitido. Esse processo é feito por um *CODEC*, o qual é utilizado tanto nos meios de telefonia convencionais quanto em implementações de VoIP.

Há várias técnicas diferentes utilizadas para tratar estes fluxos de áudio e vídeo. A maioria dos *CODECs* utilizados atualmente são padronizados nas recomendações ITU-T, embora haja vários outros. A maior parte do trabalho realizado pelos *CODECs* é um esforço para reduzir a quantidade de largura de banda consumida pelos fluxos de voz por meio do uso de compressão. Aplicações de VoIP requerem o mesmo processo de conversão, embora nem sempre exista a preocupação com largura de banda das topologias de rede tradicionais (HARTPENCE, 2013).

As amostras de áudio das bases de dados utilizadas neste trabalho usam o *CODEC* definido pela Rec. ITU-T G.711 (ITU-T, 1993) que também é conhecido como Modulação por Código de Pulso (*Pulse Code Modulation*, PCM). Este, por sua vez, codifica um único canal de voz realizando a amostragem 8.000 vezes por segundo com amostras de 8 *bits*, a fim de fornecer voz descompactada a 64 kbps (TANENBAUM; WETHERALL, 2010).

2.2.1.1 Sinal de voz

O sistema auditivo humano consegue ouvir frequências entre 20 Hz e 20 KHz. Entretanto, segundo Hartpence (2013), o mais baixo som produzido pelo sistema vocal humano já registrado foi um pouco acima de 0 Hz e o mais alto, acima 4.000 Hz. De qualquer forma, esses limites não são alcançados em situações normais de fala. Por essa razão os sistemas telefônicos operam com frequências que vão de 300 Hz a 3.400 Hz para transmissão de voz, como documentado pela Rec. ITU-T G.107. Adicionalmente, o autor cita que o sinal de voz tem o

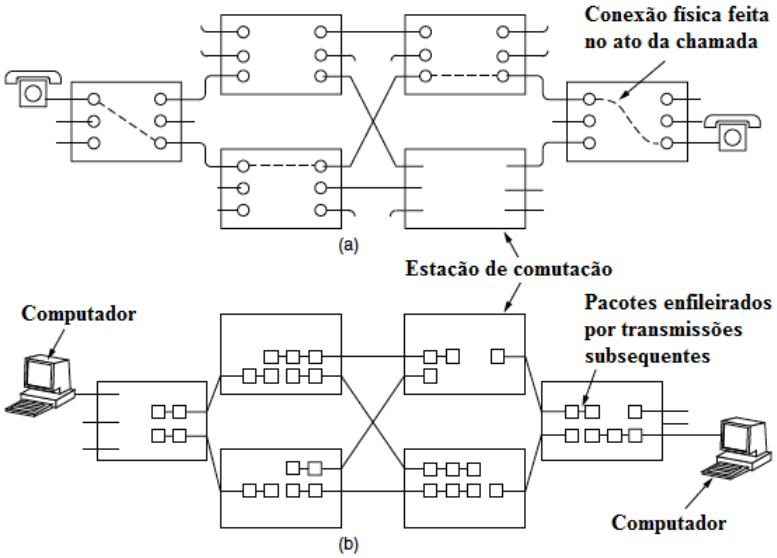
formato de uma onda que varia em função do tempo, da frequência e da amplitude. E, como o sinal produzido pela fala humana é analógico, se uma pessoa repetir a mesma palavra N vezes, o sinal de voz capturado terá uma variação de intensidade e tom, mesmo que se tente manter estas propriedades em mesmo nível.

Outro fato importante que implica sobre o sinal de voz é que o sistema telefônico tende a diminuir a amplitude do sinal. Deste modo, o desafio de um *CODEC* é capturar amostras suficientes para codificar e decodificar este sinal de modo a não deformá-lo no ponto de recepção, e garantir certo nível de qualidade.

2.2.2 Comutação

Tão importante quanto o processo de codificação e decodificação é a forma como esta informação trafega pela rede. Assim, a comutação consiste na forma como o sistema telefônico se conecta para transmitir fim-a-fim dados ou voz. Segundo Tanenbaum e Wetherall (2010), atualmente são utilizadas as técnicas de comutação de circuitos e de comutação de pacotes, sendo essa última a utilizada em chamadas VoIP.

Figura 2.1 – (a) Comutação de circuitos. (b) Comutação de pacotes.



Fonte: Tanenbaum e Wetherall (2010)

Conforme pode ser observado na Fig.2.1(a), na comutação de circuitos, quando um usuário efetua uma ligação, a central de comutação telefônica, equipamento representado por um retângulo, procura um caminho físico do telefone de origem até o de destino. Este processo na busca por um caminho é responsável pelo intervalo de tempo gasto entre a discagem do último número até o primeiro toque do telefone. Nesse exemplo, cada estação tem três linhas

de entrada e três linhas de saída. Quando uma chamada passa por uma estação de comutação, é conceitualmente estabelecida uma conexão física entre a linha que transportou a chamada e uma das linhas de saída, como mostram as linhas pontilhadas.

Na comutação de pacotes, representada pela Fig. 2.1(b), pacotes individuais são enviados conforme necessário, sem a configuração com antecedência de qualquer caminho dedicado. Cabe a cada pacote descobrir de forma independente seu caminho até o destino (TANENBAUM; WETHERALL, 2010).

Existem várias diferenças entre estas duas técnicas de comutação e elas se equilibram entre vantagens e desvantagens de lado a lado. No entanto, segundo Tanenbaum e Wetherall (2010), a comutação de pacotes é mais tolerante a falhas que a comutação de circuitos, pois, entre outros fatores, o caminho físico percorrido não é dedicado; cada pacote pode seguir uma rota diferente e a ocorrência de falhas em uma central de comutação entre a origem e o destino não interrompe a ligação.

2.2.3 Serviço de VoIP

O serviço de chamadas de VoIP é, segundo Hartpence (2013), um serviço de voz que funciona sobre uma rede IP e que se baseia em comutação de pacotes, diferentemente do Serviço Telefônico Fixo Comutado (STFC), que emprega comutação de circuitos. Nesse cenário deve-se considerar também o conceito de *Internet Protocol Multimedia Subsystem* (IMS), que é uma arquitetura de controle de serviço global, de acesso independente e de conectividade baseada no padrão IP, que permite vários tipos de serviços multimídia para os usuários finais utilizando protocolos de Internet comuns (POIKSELKA; MAYER, 2013).

Assim, a verdadeira integração de serviços de voz e dados aumenta a produtividade e a efetividade global, enquanto o desenvolvimento de aplicações inovadoras integrando voz, dados e multimídia irá criar demandas por novos serviços. A habilidade de combinar a mobilidade e a rede IP será crucial para o sucesso destes serviços no futuro (POIKSELKA; MAYER, 2013).

Adicionalmente, com a adoção da LTE e do IMS por parte das operadoras, as comunicações realizadas por telefonia móvel serão cada vez mais comutadas por pacotes através de chamadas VoIP, conforme dados publicados pelo Relatório Anual de 2015 da ANATEL em Brasil (2016).

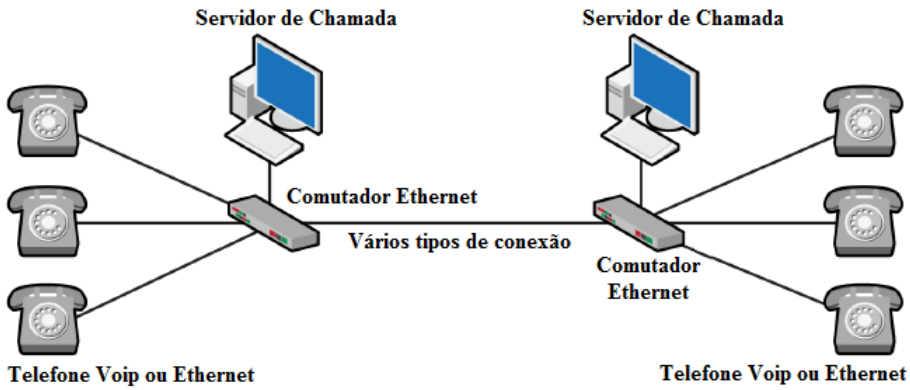
2.2.3.1 Arquitetura

Do ponto de vista da arquitetura, os sistemas de VoIP nativos têm um formato diverso do formato utilizado na telefonia tradicional, em que a comutação de circuitos aloca recursos para cada chamada individual. Além disso, este sistema vem sendo utilizado por décadas para entregar chamadas telefônicas confiáveis e de baixo consumo de banda, com alto nível de qualidade (HARTPENCE, 2013).

Do outro lado, as redes IP utilizam comutação de pacotes, e cada pacote enviado é semiautônomo, tem seu próprio cabeçalho IP, e é transmitido separadamente pelos roteadores ao longo da rede (HARTPENCE, 2013).

Segundo Hartpence (2013), os aparelhos telefônicos utilizados no sistema VoIP são denominados telefones VoIP, ou *Ethernet*. Adicionalmente, para que o serviço funcione são necessários dois tipos de protocolos: os de sinalização, responsáveis pelo processo inicial e final da transmissão, e os protocolos de transporte, que se encarregam de transmitir os sinais de voz e/ou vídeo e dados. Esses protocolos são descritos a seguir. Na Fig. 2.2 apresenta-se a arquitetura básica do VoIP.

Figura 2.2 – A arquitetura básica do VoIP.



Fonte: Hartpence (2013)

2.2.3.2 Protocolos de sinalização

Segundo Hartpence (2013), os protocolos de sinalização mais utilizados são o *Skinny*, protocolo proprietário da Cisco, o H.323 da família de recomendações da ITU-T, e o *Session Initiation Protocol* (SIP), que é o mais recomendado, por não ser proprietário, ser mais fácil de ler e utilizar e por ser suportado por diversos fornecedores da indústria VoIP.

Hartpence (2013) cita também que todos desempenham praticamente as mesmas funções, sendo o H.323 o mais completo e, por isso, o mais pesado e complexo. O *Skinny*, como dito, tem a desvantagem de ser proprietário e, então, o SIP seria o de maior relação custo/benefício. Entretanto, Tanenbaum e Wetherall (2010) afirmam que a recomendação H.323 é mais uma avaliação da arquitetura de telefonia da Internet que um protocolo específico, e faz referência a um grande número de protocolos específicos para codificação de voz, configuração de chamadas, sinalização, transporte de dados e outras áreas, em vez de especificar propriamente cada um desses elementos.

O SIP opera na camada de aplicação e sua função é iniciar sessões de usuário para transmissão de voz, vídeo, jogos e realidade virtual. Além disso, também foi projetado para configurar e encerrar sessões de mídia, localização do usuário e capacidades, disponibilidade e informação de manipulação de sessão (HARTPENCE, 2013).

Diferente do H.323, um conjunto de protocolos completo, o SIP é um único módulo, mas foi projetado para interoperar bem com aplicações da Internet existentes (TANENBAUM; WETHERALL, 2010).

Considerando, então, seu modo de operação, quando dois usuários de VoIP se conectam, tem de haver um mecanismo para a criação da comunicação e o estabelecimento de algumas regras. Este trabalho é feito pelo SIP e pelo Protocolo de Descrição de Sessão (*Session Description Protocol*, SDP), que também é responsável pela negociação de parâmetros da conexão multimídia (HARTPENCE, 2013).

Mas, antes de os terminais se conectarem, eles precisam se encontrar. Isso é possibilitado pelo endereçamento do SIP, que é similar ao *e-mail*, mas com adição da porta que, caso não seja indicada, o sistema utiliza a porta 5060 por padrão. A seguir os exemplos.

- sip:usuario@dominio:porta

- sip:usuario@host:porta

- sip:<numero do telefone>@dominio

Este formato é conhecido como Identificador Uniforme de Recursos (*Uniform Resource Identifier*, URI) e é comumente integrado com Sistemas de Nomes de Domínio (*Domain Name System*, DNS) (HARTPENCE, 2013).

2.2.3.3 Protocolos de transporte

O principal protocolo de transporte utilizado neste processo é o *User Datagram Protocol* (UDP). Segundo Tanenbaum e Wetherall (2010), o UDP e o *Transmission Control Protocol* (TCP) são os dois principais protocolos da Internet para a camada de transporte.

O UDP é um protocolo de transporte sem conexão e oferece um meio para as aplicações enviarem datagramas IP encapsulados sem que seja necessário estabelecer uma conexão (TANENBAUM; WETHERALL, 2010).

O UDP transmite segmentos que consistem em um cabeçalho de 8 *bytes*, seguido pela carga útil. Assim, o principal valor de se ter o UDP em relação ao uso do IP bruto é a adição das portas de origem e destino. Sem os campos de portas, a camada de transporte não saberia o que fazer com o pacote. Com eles, a camada entrega segmentos corretamente (TANENBAUM; WETHERALL, 2010).

Uma área na qual o UDP é, segundo Tanenbaum e Wetherall (2010), especialmente útil é a de situações cliente/servidor, o que tecnicamente, também abrange aplicações de tempo real. Algumas vezes, o cliente envia uma pequena solicitação ao servidor e espera uma pequena resposta de volta. Se a solicitação ou a resposta se perder, o cliente simplesmente chegará ao *timeout* e tentará de novo. Não só o código é simples, mas é necessário um número menor de mensagens do que no caso de um protocolo que exige uma configuração inicial.

Do outro lado, o TCP é um protocolo confiável e orientado a conexões, que permite a entrega sem erros de um fluxo de *bytes* originário de uma determinada máquina em qualquer computador da inter-rede (TANENBAUM; WETHERALL, 2010).

Assim, esse protocolo fragmenta o fluxo de *bytes* de entrada em mensagens discretas e passa cada uma delas para a camada inter-redes. No destino, o receptor volta a montar as mensagens recebidas no fluxo de saída. O TCP também cuida do controle de fluxo, impedindo que um transmissor rápido sobrecarregue um receptor lento com um volume de mensagens maior do que ele pode manipular (TANENBAUM; WETHERALL, 2010).

De acordo com Hartpence (2013), o registro dos terminais VoIP, a configuração, os números de discagem, as sessões de mídia e outras características são responsabilidades dos protocolos de sinalização.

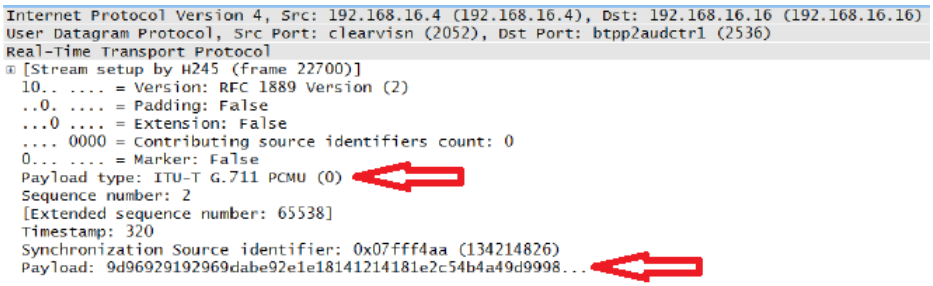
Além disso, independente de qual deles será utilizado, todos necessitam transportar dados de voz de um telefone para outro, e todos utilizam o Protocolo de Transporte de Tempo Real (*Real-Time Transport Protocol*, RTP), conforme afirma Hartpence (2013).

O RTP utiliza também outro protocolo, conhecido como Protocolo de Controle de Tempo Real (*Real-Time Control Protocol*, RTCP), que fornece retorno sobre a qualidade ou o desempenho do fluxo RTP. O objetivo é permitir que, por meio do monitoramento provido pelo RTCP, recursos de rede possam ser alocados sob demanda, melhorando, assim, a qualidade da transmissão (HARTPENCE, 2013).

Neste contexto, o RTP encapsula os dados de voz e/ou vídeo processados pelo *CODEC* e, então, os coloca dentro do pacote RTP que, por sua vez, é colocado dentro de um pacote do UDP, a fim de diminuir o tempo de entrega (HARTPENCE, 2013).

Na Fig. 2.3 mostra-se um exemplo de pacote RTP e a sua carga útil (*payload*) devidamente codificada, dentro de um pacote UDP.

Figura 2.3 – Exemplo de um pacote RTP.



Fonte: Hartpence (2013)

O receptor deve conhecer o tipo de *CODEC* utilizado e ser capaz de decodificar o *payload* contido no pacote RTP recebido.

Hartpence (2013) ressalta também que os campos principais no RTP são o *payload*, e o identificador de sequência, utilizado para reordenar os pacotes no receptor e, assim, reconstruir corretamente o sinal transmitido. O uso deste identificador de sequência se faz necessário, uma vez que o UDP não dispõe de função específica para este fim.

2.3 Algoritmos de avaliação de qualidade de voz

Os testes de avaliação de qualidade de voz podem ser realizados seguindo métodos objetivos, em que o índice MOS é dado por *software*, ou subjetivos, nos quais há intervenção de indivíduos no resultado da avaliação.

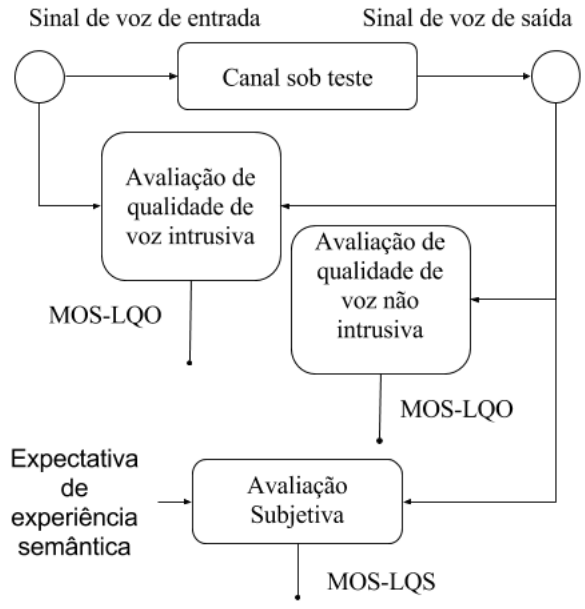
Os métodos subjetivos são baseados em testes de audição conduzidos em um ambiente controlado de laboratório onde voluntários seguem os procedimentos que o supervisor do teste

estabelece. E os métodos objetivos são baseados em algoritmos que tentam predizer a avaliação humana sobre um sinal de voz (ITU-T, 1996).

Os métodos objetivos são divididos em dois tipos: não intrusivos e intrusivos. Os métodos intrusivos precisam de um sinal de referência para comparar com o sinal no ponto final e garantir a qualidade da pontuação (ITU-T, 1996). Por esta razão os métodos intrusivos são mais confiáveis e considerados referência para a avaliação objetiva. Em contrapartida, métodos não intrusivos são os métodos que precisam apenas do sinal no ponto final, ou em um dado ponto onde o sinal deve ser avaliado e, assim, são mais baratos e rápidos, o que possibilita seu uso em serviços de tempo real (ITU-T, 2001).

Na Fig. 2.4, adaptada da ITU-T P.563 (ITU-T, 2004), apresenta-se a diferença entre os tipos de métodos utilizados na avaliação da qualidade do sinal de voz.

Figura 2.4 – Diferença entre os métodos subjetivos, objetivos, não intrusivos e intrusivos.



Fonte: ITU-T P.563 (ITU-T, 2004)

As recomendações ITU-T: P.800, P.862, P.863 e P.563 são descritas a seguir.

2.3.1 Recomendação ITU-T P.800

A Rec. ITU-T P.800 descreve métodos e procedimentos para condução e avaliação subjetiva da qualidade de transmissão. As avaliações subjetivas de equipamentos e sistemas de telecomunicações devem, em princípio, ser conduzidas utilizando-se apenas ouvintes ou métodos conversacionais de testes subjetivos (ITU-T, 1996).

Estes métodos compreendem os testes de opinião em conversação, a opinião em audição, os testes baseados em entrevistas e inspeção, entre outros. A semelhança entres eles é que todos se baseiam na opinião de vários indivíduos que, usando a escala de opinião da Tabela 2.1, fazem sua avaliação.

2.3.1.1 Escalas de opinião recomendadas pela ITU-T

Várias escalas de julgamento por categoria de cinco pontos são utilizadas para diferentes propósitos. O leiaute e a formulação das escalas de opinião, como visto pelos indivíduos nos experimentos, são muito importantes e devem seguir o padrão estabelecido pela ITU-T.

Nestas escalas são avaliados diversos critérios e cada critério é avaliado recebendo uma pontuação de 1 a 5, em que o valor 1 representa o pior caso e o valor 5 representa o melhor caso.

Assim, os valores dados pelos indivíduos participantes são, ao final dos testes, computados e, então, é calculada a média aritmética dos valores. A esta média é dado o nome de Pontuação Média de Opinião (*Mean Opinion Score*, MOS). As seguintes escalas de opinião são as mais frequentemente utilizadas pelas aplicações ITU-T e devem ser adaptadas ao idioma desses indivíduos durante os testes (ITU-T, 1996).

Na Tabela 2.1 apresenta-se a *Listening-Quality Scale*, que é utilizada para avaliação da qualidade do sinal, utilizada nos testes conversacionais e de audição (ITU-T, 1996).

Tabela 2.1 – Pontuação de MOS conversacional e testes de audição.

Qualidade do sinal	**Pontuação**
Excelente	5
Bom	4
Regular	3
Ruim	2
Péssimo	1

Fonte: ITU-T P.800 (ITU-T, 1996)

Na Tabela 2.2 apresenta-se a *Listening-Effort Scale*, ou Escala de Esforço de Audição. A quantidade avaliada a partir das pontuações é representada pelo símbolo MOS_{LE}.

Na Tabela 2.3 está demonstrada a *Loudness-Preference Scale*. A quantidade avaliada a partir das pontuações é representada pelo símbolo MOS_{LP}.

Tabela 2.2 – Pontuação MOS para testes de esforço de audição

Esforço necessário para compreender o significado das frases	Pontuação
Audição perfeita, nenhum esforço	5
Necessário atenção; nenhum esforço considerável foi requerido	4
Requerido esforço moderado	3
Requerido esforço considerável	2
Nenhum significado compreendido com qualquer esforço viável	1

Fonte: ITU-T P.800 (ITU-T, 1996)

Tabela 2.3 – Pontuação MOS para a escala de preferência de audibilidade

Preferência de audibilidade	Pontuação
Muito mais alto do que o preferido	5
Mais alto do que o preferido	4
Preferido	3
Mais baixo do que o preferido	2
Muito mais baixo do que o preferido	1

Fonte: ITU-T P.800 (ITU-T, 1996)

2.3.2 Recomendação ITU-T P.862

A Rec. ITU-T P.862, intitulada Avaliação Perceptiva da Qualidade de Voz (*Perceptual Evaluation of Speech Quality*, PESQ) trata-se de um método objetivo e intrusivo e, por isso, é considerada uma referência para a validação de outros métodos objetivos não intrusivos.

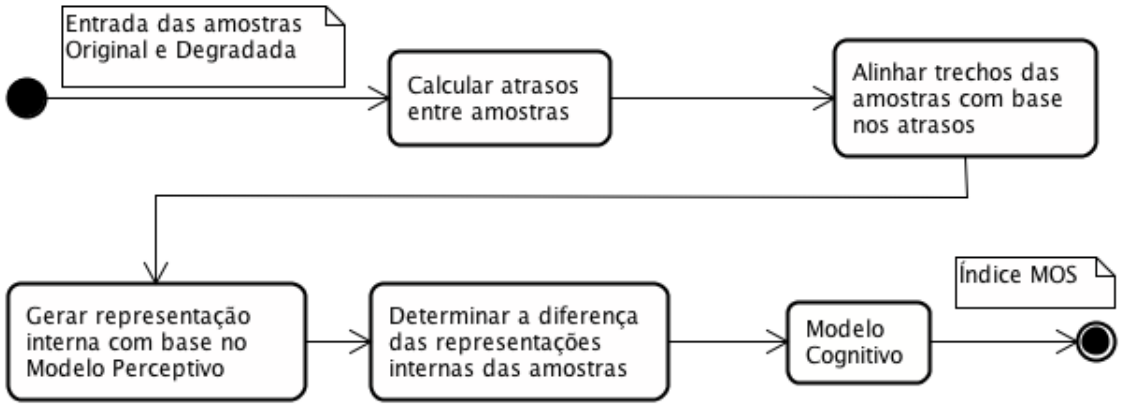
2.3.2.1 Visão geral

De acordo com a ITU-T P.862 (ITU-T, 2001), o PESQ compara um sinal original de áudio $X(t)$ com um sinal de áudio degradado $Y(t)$, que é resultado da passagem de $X(t)$ pelos sistemas de telecomunicação. A saída do PESQ é uma predição da QoE percebida de $Y(t)$ por um ouvinte em um teste.

Conforme demonstrado na Fig. 2.5, na primeira etapa do processamento os atrasos entre o arquivo original e o arquivo degradado são computados. Isso é feito para cada intervalo de tempo para o qual o atraso é significativamente direfente do intervalo de tempo anterior. Depois, para cada um destes intervalos são calculados os pontos de início e fim. O algoritmo pode também tratar mudanças de atraso durante os silêncios e durante a atividade de voz dos indivíduos (ITU-T, 2001).

Assim, baseando-se no conjunto destes atrasos, o PESQ compara o sinal original de entrada com o degradado de saída, já alinhados. O ponto chave deste processo é a transformação

Figura 2.5 – Visão geral da filosofia básica usada no PESQ.



Fonte: Do autor (2017)

destes áudios para uma representação interna análoga à representação psicofísica do sistema auditivo humano, tendo em conta a frequência perceptual e a sonoridade (ITU-T, 2001).

Segundo a ITU-T P.862 (ITU-T, 2001), um modelo computacional do sujeito, que consiste de um modelo perceptivo e cognitivo, é utilizado para comparar a saída do dispositivo sob teste com a entrada, empregando informações de alinhamento derivadas dos sinais de voz no domínio do tempo no módulo de alinhamento de tempo.

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. Tal processo é realizado limitando-se a quantidade de compensação e fazendo a compensação do atraso por trás do efeito. Assim, pequenas diferenças de estado de equilíbrio entre o original e o degradado são compensadas (ITU-T, 2001).

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. Em PESQ, dois parâmetros de erro são computados no modelo cognitivo, os quais são combinados para dar qualidade objetiva de audição MOS (ITU-T, 2001).

2.3.3 Recomendação ITU-T P.863

A ITU-T P.863 (ITU-T, 2014) descreve um método objetivo para predição global da qualidade de voz a partir da banda estreita (300 até 3.400 Hz) até a super bandalarga (50 até 14.000 Hz) em cenários de telecomunicações como percebido pelo usuário em um teste da Rec. ITU-T P.800.

Esta recomendação, por se tratar de uma sucessão da Rec. ITU-T P.862, tem a mesma filosofia de avaliação e formato de trabalho. As únicas diferenças, neste caso, são que, nesta recomendação, o algoritmo leva em consideração os aspectos inerentes à superbanda larga no seu processo de avaliação e também faz um tratamento nos sinais de áudio tratando ruídos de baixo nível.

2.3.3.1 Visão geral

O algoritmo desta recomendação segue o modelo de referência completa e opera fazendo uma comparação entre um sinal de referência conhecido e um sinal degradado, capturado após passar pelo meio de comunicação que está sendo testado. A ITU-T P.863 (ITU-T, 2014) cita que este algoritmo é consistente com os algoritmos descritos nas recomendações ITU-T P.861 e ITU-T P.862 e pode ser visto como sendo um sucessor destes.

A Rec. ITU-T P.861, publicada em 1996, foi inicialmente focada na identificação do impacto da qualidade dos *CODECs*. Logo após, a ITU-T começou a trabalhar na criação de um algoritmo capaz de avaliar o impacto adicional dos problemas relacionados à rede. Este trabalho resultou, então, na Rec. ITU-T P.862, lançada em 2001.

Agora, a Rec. ITU-T P.863 incorpora requisitos atuais da indústria e, em particular, permite a avaliação de voz na super bandalarga, bem como em redes e *CODECs* que introduzem a sincronização temporal (ITU-T, 2014).

2.3.4 Recomendação ITU-T P.563

A ITU-T P.563 (ITU-T, 2004) descreve um método objetivo e não intrusivo para a predição da qualidade subjetiva de voz, oriunda de aplicações telefônicas de banda estreita (3.1 kHz).

O algoritmo apresentado por esta recomendação não necessita de um sinal de referência para sua avaliação e, por esta razão, é recomendado para o monitoramento de redes em tempo real e a avaliação de fontes de voz desconhecidas em um dado ponto da conexão telefônica. Além disso, esta recomendação leva em conta todas as classes de distorção que ocorrem nas redes do STFC.

O sistema de pontuação utilizado é baseado na escala MOS-LQO, que está de acordo com a Rec. 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, conforme a escala citada.

A validação da Rec. ITU-T P.563 inclui todos os experimentos disponíveis no processo de validação da Rec. 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.

2.3.4.1 Escopo

O escopo da Rec. 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 que traz fatores de testes, tecnologias de codificação e aplicações às quais esta recomendação se aplica.

A seguir é demonstrado o conjunto de fatores de teste com os quais o algoritmo da Rec. ITU-T P.563 produz resultados com aceitável precisão. São eles:

- características do ambiente acústico como utilizado na fase de validação;
- ruído do ambiente no lado de envio do sinal;
- características da interface acústica do terminal que enviou o sinal;
- níveis de entrada de voz em um codificador;
- erros do canal de transmissão;
- perda de pacotes e dissimulação da perda de pacotes com codificadores CELP;
- taxas de *bits* se um *CODEC* tem mais de um modo de taxa de *bits*;
- transcodificadores;
- efeito da variação de atrasos na qualidade ouvida em testes ACR;
- distorção em longo e curto prazo de tempo do sinal de voz;
- sistemas de transmissão, incluindo cancelamento de *echo* e sistemas de redução de ruído em condições de fala simples e como eles serão pontuados em uma escala ACR.

As tecnologias de codificação que produziram resultados com aceitável precisão são:

- codificadores em forma de onda, por exemplo, G.711, G.726 e G.727;
- CELP e codificadores híbridos com taxa de *bits* superior ou igual a 4 kbit/segundo, tais como G.728, G.729 e G.723.1;
- outros codificadores: GSM-FR, GSM-HR, GSM-EFR, GSM-AMR, CDMA-EVRC, TDMA-ACELP, TDMA-VSELP, TETRA.

Assim, os cenários de aplicação recomendados para Rec. ITU-T P.563 são:

- monitoramento de rede em tempo real, utilizando conexão digital ou analógica para a rede;
- testes fim-a-fim de redes em tempo real, utilizando conexão digital ou analógica para a rede;
- testes fim-a-fim de redes em tempo real, com fontes de voz desconhecidas na outra extremidade.

Na sequência é apresentado um conjunto de condições para as quais esta recomendação fornece previsões imprecisas. São elas:

- nível de audição, perda de sonoridade;
- retorno da própria voz;
- efeito de atrasos em testes conversacionais;
- eco do interlocutor;
- música ou tom da rede como sinal de entrada.

As tecnologias de codificação que produziram resultados imprecisos são as *vocoder* LPC com taxas de *bits* menores que 4 kbit/segundo, como, por exemplo, IMBE, AMBE e LPC10e. Adicionalmente, a Rec. ITU-T P.563 indica que os cenários de aplicação imprecisos com os fatores citados são a predição da qualidade do falante e o desempenho de comunicação de duas vias.

Finalmente, um conjunto de condições para as quais esta recomendação não foi total ou parcialmente validada é listado. Na sequência, os fatores de teste são listados.

- Recorte da amplitude da voz.
- Múltiplos locutores simultaneamente e dependências do locutor.
- Voz cantada ou voz de criança como entrada do *CODEC*.
- Desencontro da taxa de *bits* entre um codificador e um decodificador, se um *CODEC* tem mais do que um modo de taxa de *bits*.
- Sinal de voz artificial para um *CODEC*.
- Eco do ouvinte.
- Efeitos/artefatos de inibidores de eco isolados.
- Efeitos/artefatos de algoritmos de redução de ruídos isolados.
- Avaliação de voz sintética e/ou utilização destes como entrada para um *CODEC* de voz.

As tecnologias de codificação para as quais esta recomendação não foi total ou parcialmente validada são MPEG-4 HVXC, CELP e codificadores híbridos com taxa de *bits* menores que 4 kbit/segundo. O cenário de aplicação condizente com estas condições é o cenário de medidas na interface acústica do terminal de recepção/aparelho.

A Rec. ITU-T P.563 cita também 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. Entretanto, pode-se aplicar estas medições onde os testes subjetivos seriam muito caros.

Deve-se observar também que este algoritmo não fornece uma avaliação global da qualidade da transmissão, medindo apenas alguns efeitos da distorção da fala unidirecional e do ruído sobre a qualidade percebida da voz, do mesmo jeito que isso pode ser investigado por um teste auditivo avaliando a qualidade percebida em uma escala ACR. 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 (*Sound Pressure Level*, SPL) de 79 dB no ponto de referência da orelha (*Ear Reference Point*, ERP).

Dessa forma, apenas os efeitos da perda de sonoridade, atrasos, eco do orador e outras deficiências relacionadas à qualidade da fala ou da interação bidirecional não afetam a pontuação dada pelo algoritmo da Rec. ITU-T P.563.

É importante ressaltar que o algoritmo da Rec. 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.

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 fala;
- nível de fala entre –36 e –16 *Decibel to Overload Point* (dBov).

Apesar do último requisito, esta recomendação realiza um ajuste do sinal para -26 dBov, a fim de evitar artefatos adicionais em função da baixa relação entre o sinal e o ruído (Signal-to-Noise Ratio, SNR) ou recorte de amplitude.

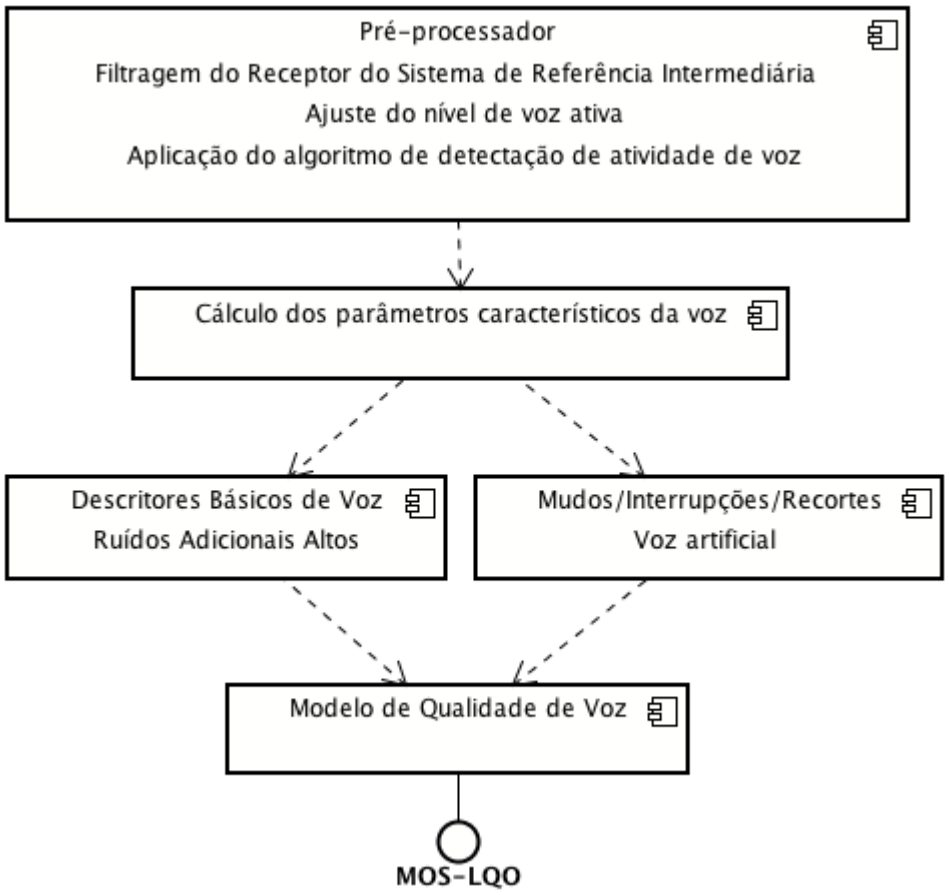
2.3.4.2 Visão geral do algoritmo da Rec. ITU-T P.563

Segundo a Rec. ITU-T P.563 (ITU-T, 2004), a abordagem dada por seu algoritmo deve ser vista pela perspectiva de um especialista que está ouvindo uma chamada real com um dispositivo de teste. Entretanto, para avaliar este mesmo sinal, o algoritmo primeiro faz um pré-processamento. Este por sua vez, começa com o modelo do dispositivo receptor e, em seguida, um algoritmo de detecção de atividade de voz (*Voice Activity Detector*, VAD) é usado para identificar a voz que será avaliada. Por fim, um ajuste no nível de voz é aplicado.

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 do algoritmo. Na Fig. 2.6 observa-se o diagrama de componentes do algoritmo da Rec. ITU-T P.563 em blocos.

Basicamente, o processo de parametrização do sinal a ser submetido ao algoritmo da Rec. ITU-T P.563 pode ser dividido em três blocos funcionais independentes que correspondem às principais classes de distorção (ITU-T, 2004). São eles:

Figura 2.6 – Diagrama de componentes do algoritmo da Rec. ITU-T P.563.



Fonte: Do autor (2017)

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

Além disso, um conjunto básico de descritores de voz é utilizado, principalmente para ajustar o pré-processamento e o VAD. Esses três tópicos são explicados em detalhes a seguir.

2.3.4.3 Análise do trato vocal e artificialidade da voz

O bloco principal procura por artificialidades no sinal de voz, usando um modelo de produção de voz para extrair partes do sinal que podem ser interpretados como voz e separando-os das outras partes. Além disso, uma análise estastística de alta ordem dá informações adicionais sobre o quão humana a voz é.

Estes sinais classificados como artificiais são separados em vozes masculinas e femininas, e, em caso de forte robotização, outra classificação será realizada, a qual é independente de gênero.

Outro distúrbio impactante na avaliação do algoritmo da Rec. ITU-T P.563 é a repetição de quadros de voz, ocorrida em sistemas de transmissão baseados em pacotes. Este distúrbio é provocado por alguns *CODECs* que enviam seus pacotes repetidos que seriam utilizados para substituir os pacotes perdidos. No entanto, esta técnica tende a diminuir a qualidade do sinal, mais do que aumentar (ITU-T, 2004).

2.3.4.4 Análise de fortes ruídos adicionais

Segundo Tanenbaum e Wetherall (2010), o ruído consiste em energia indesejável proveniente de outras fontes diversas do transmissor. Este ruído pode ser térmico, quando é causado pelo movimento aleatório dos elétrons em um fio, ou de impulso, quando provocado, dentre outros fatores, por picos de voltagem na linha de energia.

Assim, é inevitável que a análise de ruídos seja feita a fim de calcular diferentes características provenientes deste problema. Esta análise se baseia em dois parâmetros-chave, decidindo se o ruído adicional é a principal classe degradante. Caso seja, o tipo de ruído deve ser identificado, podendo ser do tipo estático e presente sobre todo o sinal, de tal forma que a potência do ruído não está correlacionada com o sinal, ou a potência do ruído mostra dependências do envelope de energia do sinal.

Conforme a Rec. ITU-T P.563 (ITU-T, 2004), uma vez detectada a presença de ruído estático, vários detectores tentam quantificá-lo localmente e globalmente, sendo o primeiro ocorrido especialmente entre os fonemas, enquanto o segundo ocorre entre as sentenças. A distinção destes dois tipos é importante, por exemplo, nas comunicações móveis que, geralmente, usam diferentes configurações para partes de atividade de voz e para partes sem atividade de voz.

2.3.4.5 Interrupções, silêncios e recortes de tempo

Silêncios e interrupções também formam uma classe de distorção. Tais distorções são descritas apenas parcialmente pela análise do trato vocal. Assim, uma análise separada é realizada para detectar e avaliar os recortes de tempo e silêncios não naturais no sinal.

Já a interrupção do sinal de voz pode ocorrer como recorte de voz temporal, ou como interrupção de voz; ambos levam à perda de informação.

Segundo a Rec. ITU-T P.563 (ITU-T, 2004), o recorte temporal pode ocorrer quando o VAD ou o equipamento de multiplicação de circuito digital (*Digital Circuit Multiplication Equipment*, DCME) são utilizados, ou o sinal é interrompido. É possível detectar as interrupções do sinal de voz, que ocorrem durante os intervalos ativos de voz. Os algoritmos utilizados na Rec. ITU-T P.563 são capazes de distinguir entre fins de palavras normais e interrupções de sinal anormal, bem como intervalos de silêncio não naturais em um trecho de fala.

2.4 Algoritmo de detecção de atividade de voz

Sadjadi e Hansen (2013) citam que a detecção de atividade de voz tem aplicações em uma grande variedade de contextos, tais como codificação de voz, reconhecimento automático de voz, identificação de idioma, melhoramento do sinal de voz e aplicações de monitoramento e vigilância que envolvem análise de longos trechos de fala.

Segundo Ramirez, Górriz e Segura (2007), a detecção de trechos com e sem fala em um sinal de voz é um problema relevante no processamento de voz e afeta diversas aplicações incluindo o reconhecimento de voz. No entanto, realizar esta classificação não é uma tarefa tão trivial quanto parece e a maioria dos algoritmos detectores de atividade de voz (*Voice Activity Detector*, VAD) falha quando os níveis de ruído do ambiente aumentam.

2.4.1 Definição

Como o próprio nome sugere, um algoritmo de VAD consiste em um algoritmo capaz de detectar os trechos de voz em um dado sinal de voz.

Este processo pode ser realizado utilizando-se diferentes métodos, como citam Ramirez, Górriz e Segura (2007), sendo eles baseados na separação por nível de energia, ou *power threshold* (SADJADI; HANSEN, 2013), detecção de *pitch* (ABDULLAH-AL-MAMUN; SARKER; MUHAMMAD, 2009), análise de espectro (MARZINZIK; KOLLMEIER, 2002), *Zero-Crossing Rate* (BENYASSINE et al., 1997), estatísticas de ordem superior (NEMER; GOUBRAN; MAHMOUD, 2001) ou combinações de diferentes características.

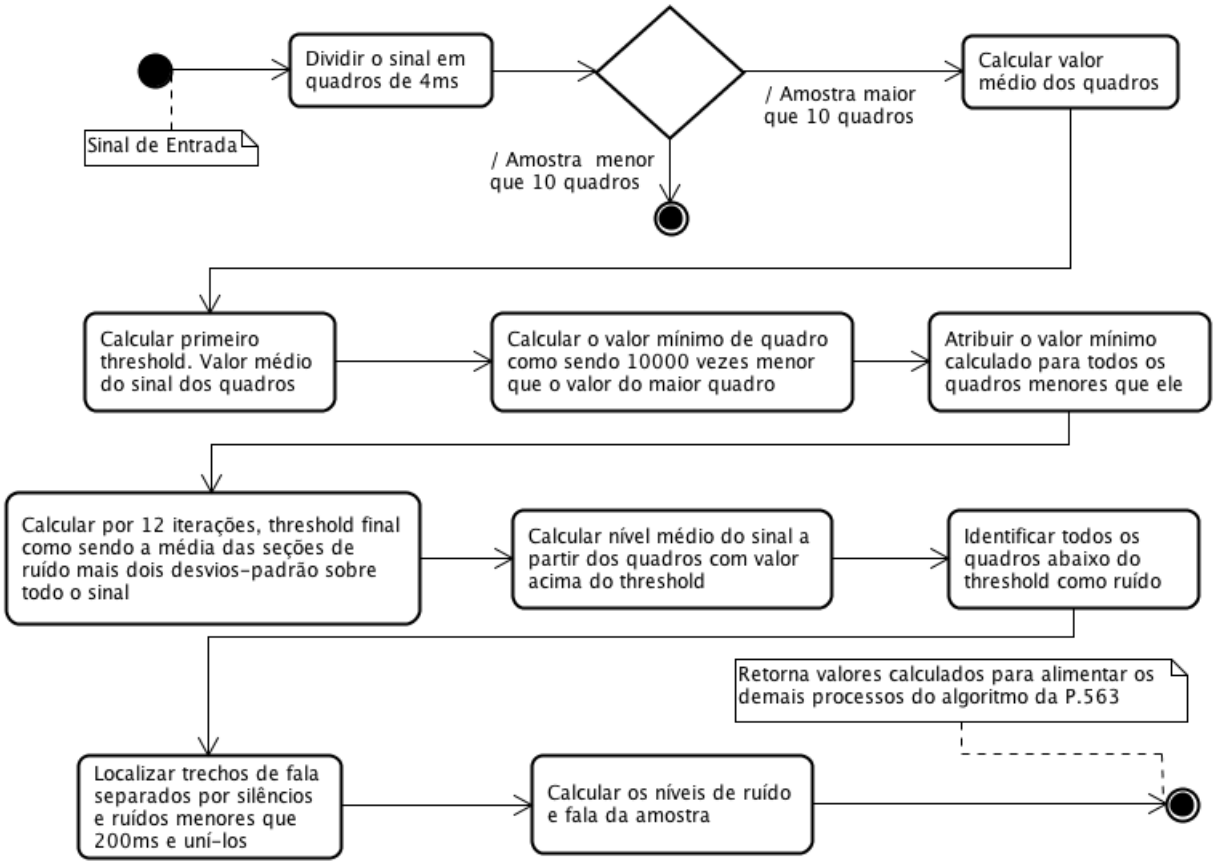
Segundo Ramirez, Górriz e Segura (2007), VAD é uma técnica muito útil para o melhoramento do desempenho de sistemas de reconhecimento de voz.

No entanto, módulos com essa função podem ser utilizados em qualquer tipo de sistema que necessite separar trechos de voz em um dado sinal, como é o caso da Rec. ITU-T P.563.

2.4.2 Utilização do VAD na Rec. ITU-T P.563

O método de VAD utilizado pela Rec. ITU-T P.563 é o *power threshold* usando uma abordagem iterativa. Segundo a Rec. ITU-T P.563 (ITU-T, 2004), o sinal é dividido em quadros e os que estão acima de um dado limite, tecnicamente chamado de *threshold*, são classificados como voz, e abaixo dele, como ruído.

Figura 2.7 – Diagrama de atividade do algoritmo VAD usado pela Rec. ITU-T P.563.



Fonte: Do autor (2017)

Na Fig. 2.7 observa-se um diagrama de atividade do algoritmo VAD utilizado pela Rec. ITU-T P.563. Em síntese, o valor limite é iterativamente reestimado para ser a média mais dois desvios padrões da energia das seções de ruído sobre todo o sinal. Então, após doze iterações serem concluídas, o valor limite final é determinado e empregado para tomar a decisão final entre sinal ou ruído.

Além disso, é feito um processamento adicional para rotular eventos curtos. Os trechos que estão separados por menos de 200 milésimos de segundo são unidos.

Assim é possível perceber como a classificação da amostra é feita, o nível de sinal e ruído são calculados e também como este algoritmo gera valores de saída para guiar o restante do cálculo do índice MOS pelo algoritmo da Rec. ITU-T P.563.

Nota-se, com isso, que o VAD interno ao algoritmo da Rec. ITU-T P.563 tem um papel central porque gera parâmetros de saída que são utilizados no restante do processo que calcula o MOS da amostra. Este fato inviabiliza, a princípio, a manipulação deste VAD, a fim de resolver a deficiência na avaliação de amostras com silêncios naturais da fala. Mesmo a possibilidade de anular esse VAD, e, portanto, transferir seu processo a um algoritmo de VAD externo, não seria recomendável.

Por esta razão, optou-se por preservar integralmente o algoritmo da Rec. ITU-T P.563, adicionando ao modelo proposto um algoritmo de VAD externo que se limita, exclusivamente, a tratar a deficiência citada ao gerar o índice MOS das amostras.

2.4.3 VAD externo ao algoritmo da Rec. ITU-T P.563

O algoritmo VAD escolhido para eliminar os silêncios naturais da fala humana também emprega o método de *power threshold*, e foi desenvolvido por Sadjadi e Hansen (2013). Entretanto, neste algoritmo o valor do *threshold* é fixado em 0,0001.

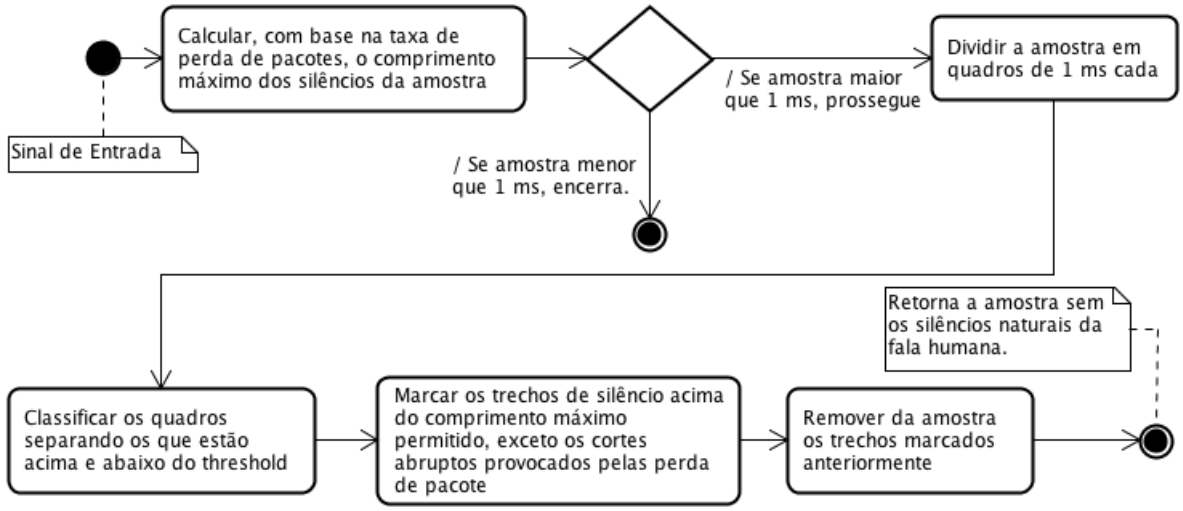
Segundo os autores, seu algoritmo trabalha bem em cenários de comunicação com muito ruído, pois utiliza quatro medidas diferentes de expressão vocal, combinadas com uma característica perceptual de fluxo espectral.

Na Fig. 2.8 apresenta-se o diagrama de atividade do VAD externo. Nele constam os principais passos percorridos para classificar e remover os silêncios já mencionados. É importante citar que o algoritmo original foi adaptado de forma que o comprimento máximo de silêncio permitido fosse variável em função da taxa de perda de pacotes (*Packet Loss Rate*, PLR).

2.5 *Web service*

A *Web* nos dias atuais, é utilizada cada vez mais para a comunicação entre aplicações. A interface programática desenvolvida para este fim é chamada de *Web Service* (WORLD WIDE WEB CONSORTIUM - W3C, 2015).

Figura 2.8 – Diagrama de atividade do algoritmo VAD externo para tratar os silêncios naturais da fala.



Fonte: Do autor (2017)

O conjunto de tecnologias e padrões desenvolvidos para regular e explorar ao máximo os *Web Services* são mantidos pela W3C e disponíveis em *https://www.w3.org/2002/ws/Activity*.

Assim, nas subseções seguintes apresentam-se as tecnologias utilizadas no seu desenvolvimento, podendo ele ser consumido por qualquer outro *software* desenvolvido em qualquer linguagem.

2.5.1 Java e a plataforma *Java Enterprise Edition*

Segundo Deitel e Deitel (2012), a linguagem Java é a linguagem de programação mais utilizada do mundo e a linguagem preferida por grandes organizações para atender às suas necessidades, principalmente aquelas ligadas a aplicações baseadas em Internet e dispositivos que se comunicam na rede.

Adicionalmente, os autores afirmam que pelo fato de Java abranger uma gama muito grande de aplicações, ela é dividida em três edições que são: (i) a *Java Micro Edition*, para dispositivos menores ou de memória limitada, (ii) a *Java Standard Edition* voltada para aplicações básicas, geralmente desenvolvidas para executarem diretamente nos computadores (iii) e a *Java Enterprise Edition*, que é voltada para o desenvolvimento em grande escala, aplicações de redes distribuídas e aplicações baseadas na web. Destes, o último é o tipo de aplicação em que o *Web Service* proposto se enquadra e, para seu pleno funcionamento, foi também necessária a utilização de outros recursos, o quais serão descritos a seguir.

2.5.2 *Representational State Transfer*

Segundo Fielding (2000), a fim de identificar os aspectos da *Web* que necessitam ser melhorados e evitar modificações indesejáveis, era necessário um modelo para uma arquitetura *Web* moderna. Assim, Roy Thomas Fielding, em sua tese de doutorado (FIELDING, 2000), criou o estilo de arquitetura de transferência do estado representacional (*Representational State Transfer*, REST). Nele, o autor também descreve como o REST foi utilizado para orientar o projeto e o desenvolvimento dessa arquitetura.

REST enfatiza a escalabilidade das interações de componentes, generalidade das interfaces, implantação independente de componentes e componentes intermediários para reduzir a latência de interação, reforçar a segurança e encapsular sistemas legados.

Na prática, em uma aplicação *Web* usando a arquitetura REST, a ênfase é dada nos recursos, nos nomes. Assim, uma aplicação com esta arquitetura definiria os recursos com os nomes constantes na Tabela 2.4.

Tabela 2.4 – Exemplo de recursos em uma arquitetura REST.

Nomes de recursos
AudioEvaluation{}
AudioFile{}

Fonte: Do autor (2017)

Em contrapartida, uma aplicação com a arquitetura convencional, conhecida como Chamada de Procedimento Remoto (*Remote Procedure Call*, RPC), daria ênfase à diversidade de operações do protocolo. Então, em um exemplo, uma aplicação RPC definiria os nomes das operações como na Tabela 2.5.

Tabela 2.5 – Exemplo das operações disponíveis em uma aplicação com RPC.

Nomes das operações
getAudioFile()
removeAudioFile()
addAudioFile()
getAudioEvaluation()
addAudioEvaluation()
...
removeAudioEvaluation()

Fonte: Do autor (2017)

Além disso, usando REST, cada recurso teria seu próprio identificador, como, por exemplo, http://www.ufla.br/deg/AudioEvaluation/evaluate, em que *evaluate* seria um método de *Au-*

dioEvaluation. Os clientes trabalhariam com esses recursos por meio das operações padrão do Protocolo de Transferência de Hipertexto (*Hypertext Transfer Protocol*, HTTP), como *GET, POST, DELETE e PUT*, para chamar uma cópia do recurso.

2.5.3 *JavaScript Object Notation*

A Notação de Objetos *JavaScript* (*JavaScript Object Notation*, JSON) foi criada inicialmente por Douglas Crokford e, atualmente, é mantida pela *Internet Engineering Task Force* (IETF, 2014), onde é descrita como sendo um formato de troca de dados independente de linguagem, leve e baseada em texto. Assim, o JSON consiste em um pequeno conjunto de regras de formatação para a representação portável da estrutura de dados.

Na prática, o JSON desempenha o mesmo papel da Linguagem de Marcação Extensível (*Extensible Markup Language*, XML) como sendo um formato para a troca de dados. Na Fig. 2.9 observa-se um exemplo de uma lista de objetos representada na notação do JSON. Esta lista trás vários objetos da classe **AudioFile**, em que cada objeto apresenta um valor para seus atributos: nome, tipo e MOS.

Figura 2.9 – Exemplo de uma lista de objetos representados pela notação JSON

```
{ "AudioFile" : [
        { "nome": "f_33_en_c_se01", "tipo": ".wav", "MOS": 5.0 },
        { "nome": "f_25_en_c_se01", "tipo": ".wav", "MOS": 3.893 },
        { "nome": "m_20_pt_c_se01", "tipo": ".wav", "MOS": 4.025 }
    ]
}
```

Fonte: Do autor (2017)

2.5.4 *Web Service*, REST e JSON

Considerando separadamente os devidos conceitos de *Web Service*, REST e JSON, pode não ser muito claro o papel de cada um deles em um contexto global de aplicação.

De forma simplista, pode-se dizer que o *Web Service* é um *software* que executa como sendo um serviço e segue dado padrão de interface para que outros *softwares* possam solicitar a execução de um ou mais métodos nele implementados.

O REST, por sua vez, determina a arquitetura e a forma como este *Web Service* pode ser acessado, como, por exemplo, pelos métodos *GET, POST, DELETE e PUT* da HTTP, usando URLs amigáveis.

Por fim, o JSON estabelece o formato e a notação dos dados que um determinado *software* troca com o *Web Service* para enviar parâmetros e receber o retorno da execução realizada.

2.6 *Android*

De acordo com Google (2015), o *Android* é uma plataforma composta por um Sistema Operacional (SO), *middlewares* e um conjunto de aplicativos principais. Ele é voltado para dispositivos móveis e já abrange mais de um bilhão de *smartphones* e *tablets* em todo o mundo.

Monteiro (2012) destaca que o *Android* é baseado no SO Linux e teve seu desenvolvimento iniciado em 2003, pela empresa *Android Inc.*, que foi adquirida pela *Google* em 2005.

Neste projeto, a fim de auxiliar na validação do algoritmo proposto, utilizou-se um aplicativo móvel desenvolvido para *Android*, em que o referido algoritmo será implementado. Todo este processo está descrito em detalhes na seção 3.

Como o *Android* tem um papel secundário neste trabalho, não foi necessária uma abordagem detalhada descrevendo suas características.

2.7 Estado da arte

Até a presente data, diversos trabalhos já foram publicados com o objetivo de propor melhorias ao algoritmo da Rec. ITU-T P.563. Assim, a seguir, serão apresentados os mais relevantes para o contexto deste trabalho.

Grancharov et al. (2006) indicaram que os algoritmos de avaliação de qualidade ou têm alta complexidade, suportando muitas aplicações ou baixa complexidade e utilidade limitada. Dessa forma, a proposta dos autores foi criar uma solução não intrusiva de baixa complexidade para a avaliação da qualidade da voz (*Low-Complexity, Nonintrusive Speech Quality Assessment*, LCQA) que pudesse funcionar com uma ampla gama de aplicações, nas quais um modelo de distorção explícito não é utilizado. A previsão de qualidade é baseada em propriedades estatísticas globais de recursos por quadro.

Os autores utilizaram duas abordagens para avaliar o desempenho da sua proposta, a complexidade computacional e a precisão da predição, em relação à Rec. ITU-T P.563. Suas soluções avaliaram amostras de áudio de 8 segundos de comprimento em 1,24 segundos contra 4,63 segundos gastos pelo algoritmo da Rec. ITU-T P.563. Entretanto, os autores não espe-

cificaram que configuração de *hardware* e *software* foi utilizada. Em relação à precisão para predizer a pontuação MOS, também foi utilizada a *ITU-T Supplement 23*. O coeficiente de correlação por condição (*Per-Condition Correlation Coefficient*, PCCC) atingiu valores de 0,93 a 0,95 que, segundo os autores, são mais altos do que os atingidos pela Rec. ITU-T P.563.

Falk e Chan (2006) propõem um algoritmo de medição de qualidade de voz mais robusto e mais preciso que o utilizado pela Rec. ITU-T P.563, baseando-se em modelos de probabilidade de Mistura Gaussiana (*Gaussian Mixture Model*, GMM). Segundo os autores, o resultado foi possível por meio do uso de informações relacionadas ao comportamento do sinal de fala degradada de diferentes transmissões e/ou esquemas de codificações.

Com o foco na limitação da Rec. ITU-T P.563 em redes IP, Raja et al. (2007) apresentam em seu trabalho um método não intrusivo de estimativa da qualidade de voz em tempo real que emula a medição da qualidade de escuta subjetiva baseada em MOS sobre VoIP, empregando uma abordagem de regressão simbólica baseada em Programação Genética (PG) para derivar um modelo de estimativa de qualidade de voz. Os autores citam que os resultados se comparam favoravelmente com o da Rec. ITU-T P.862 e também com outras pesquisas baseadas em Redes Neurais Artificiais (RNA), cuja maior limitação é a interpretação dos resultados.

Embora esses dois últimos trabalhos tenham uma abordagem diferente, os autores não citam ou referenciam como avaliariam os períodos de silêncio naturais da fala humana, os quais estão cobertos pelo trabalho aqui proposto.

Outro trabalho desenvolvido para melhorar a referida recomendação em ambientes VoIP foi o trabalho de Abareghi et al. (2008), no qual sugerem a criação de uma nova classe de distorção, a ser acrescentada às já existentes na Rec. ITU-T P.563, baseada nos parâmetros Estimativa de Comprimento de Silêncios, Queda Acentuada e Detecção de Interrupção de Sinal. Segundo descrito no trabalho citado, a perda de pacotes na rede interfere diretamente nestes parâmetros, de forma que a medida que a perda aumenta, os valores destes parâmetros também aumentam. E, então, no projeto aqui proposto, espera-se definir uma função de ajuste, a fim de compensar tal fator, melhorando a avaliação dada pelo algoritmo da Rec. ITU-T P.563.

Mossavat, Petkov e Klejin (2011) declararam que a heterogeneidade dos dados é um problema para a avaliação não intrusiva da qualidade de voz porque é prejudicial à predição de desempenho. Assim, seu trabalho se concentrou em resolver esse problema, utilizando um *Bayesian Framework* (BF) para criar um algoritmo de aprendizado multitarefas chamado *Hierarchical Bayes* (HB). A avaliação de desempenho foi realizada utilizando-se a base de dados

de áudio *ITU-T P.Supplement 23*, devido à sua heterogeneidade. O HB proposto, atingiu um coeficiente de correlação de Pearson (*Pearson Correlation Coefficient*, PCC) e erro quadrático médio da raiz (*Root Mean Square Error*, RMSE) de 0,91 e 0,29, respectivamente, superando, assim, o algoritmo da Rec. ITU-T P.563.

Já Cherif et al. (2012) apresentam uma nova ferramenta para avaliação da qualidade de voz em serviços VoIP para os *CODECs* iLBC e Speex. Seu método é objetivo, não intrusivo e baseado em RNAs que aprendem a relação não linear entre os parâmetros da rede e a qualidade de voz percebida pelo usuário para realizar a avaliação que, segundo os autores, resultaram em valores com alta correlação com o algoritmo da Rec. ITU-T P.862.

Li et al. (2014) apresentam um trabalho também voltado para a predição da avaliação da qualidade de voz, utilizando uma métrica não intrusiva, mas com o foco voltado para sinais de voz gravados em ambientes com grande ruído, o que, conforme citam, é uma deficiência do algoritmo da Rec. ITU-T P.563. Assim, por meio de um processo baseado em extração de características, os autores realizaram análises estatísticas para validar seu modelo e obtiveram resultados mais satisfatórios que os dados pelos métodos convencionais para o tipo de sinal avaliado.

Polacky e Pocta (2014), em um trabalho de pesquisa voltado para identificar o impacto da perda de pacotes, dos *CODECs* e do tipo de voz nos parâmetros internos da Rec. ITU-T P.563, afirmam que, até certo nível, o algoritmo da Rec. ITU-T P.563 consegue predizer a qualidade de voz transmitida em uma rede IP. Assim, em sua obra foram identificados os parâmetros internos dominantes da Rec. ITU-T P.563 para todos os *CODECs* investigados por meio da realização de análises de duas vias dos testes de variância em todos os parâmetros internos da Rec. ITU-T P.563, cujos resultados tiveram, estatisticamente, alta relevância em todos os casos investigados.

Em outro trabalho, Han e Muntean (2015) apresentam um modelo híbrido de avaliação da qualidade de chamada, em que eles combinam métodos objetivos do tipo intrusivo e não intrusivo, no intuito de extrair o melhor de ambos. Dessa forma, o modelo proposto estima a qualidade da chamada de voz usando um método não intrusivo no início da chamada e, ao final, tem sua avaliação corrigida utilizando-se um método intrusivo. Assim, utilizam-se pesos que combinam os resultados dos dois métodos para chegar a um valor final. Conforme descrito no trabalho citado, os pesos foram determinados baseando-se em testes subjetivos e considerando diferentes taxas de perda de pacote da rede. Entretanto, apesar de terem conseguido uma alta

correlação, não trataram ou realizaram testes sobre o resultado da avaliação de qualidade em cenários nos quais há silêncios maiores durante a fala. Outro problema dessa abordagem que não foi citado pelos autores consiste no consumo redundante da banda de rede, uma vez que todo o áudio original terá que ser transmitido via pacotes TCP para que se tenha condições de realizar a avaliação com uma métrica intrusiva.

Dubey e Kumar (2015) apresentam uma métrica não intrusiva de avaliação da qualidade de voz baseada em um modelo auditivo de múltipla resolução que provê uma modelagem de tempo/frequência detalhada do sistema auditivo humano. Os autores também utilizaram modelos de probabilidade de GMM para estimar o MOS dos arquivos de áudio avaliados e citaram que os resultados obtidos tiveram um desempenho melhor que outros métodos semelhantes. Entretanto, não citaram ou se referiram ao tratamento de silêncios naturais da voz humana.

Os trabalhos citados demonstram formas diferentes de alcançar melhores resultados que os obtidos pelos algoritmos das recomendações ITU-T P.563 e P.862, mas nenhum deles levou em consideração o fato de que, se um usuário faz uma pausa maior entre uma palavra e outra durante sua fala, isso não deve ser tido como um fator degradante da qualidade de voz. Além disso, nenhuma das soluções apresentadas visou disponibilizar aos usuários, operadoras e agências reguladoras um mecanismo que permitisse monitorar a QoE dos serviços contratados.

3 METODOLOGIA

Para implementar a solução proposta, várias etapas foram necessárias, as quais serão tratadas nessa seção. Inicialmente, aborda-se o modelo proposto, a fim de melhorar o desempenho do algoritmo da Rec. ITU-T P.563 e, posteriormente, apresentam-se os passos para criação da solução completa, que é composta pelo algoritmo de melhoria da Rec. ITU-T P.563, pelo Servidor de Qualidade de Voz (SQV) e por sua versão móvel, o SQV$_{mobile}$.

3.1 Modelo de melhoria proposto para o algoritmo da Rec. ITU-T P.563

A proposta de melhoria do algoritmo da Rec. ITU-T P.563 consiste em dois blocos fundamentais que são (i) a determinação da função de ajuste baseada nos parâmetros de rede para compensar a perda na pontuação MOS em função da PLR, conforme publicado em Pereira et al. (2015) e (ii) identificação e eliminação de silêncios naturais da fala humana, uma vez que esta propriedade diminui a pontuação MOS do áudio avaliado, como demonstrado por Nunes et al. (2016).

3.1.1 Perda de pacotes na rede IP

Conforme citado na seção 2.7, diversos estudos apontam o tratamento da perda de pacotes como um problema da Rec. ITU-T P.563. Em outra pesquisa, Malfait, Berger e Kastner (2006) concluíram que a situação mais problemática para os algoritmos de métodos não intrusivos, incluindo o da Rec. ITU-T P.563, é que, quando uma parte do sinal é perdida, se torna impossível saber qual informação estava contida no trecho perdido, fato esse que justifica a interferência ocasionada pela perda de pacotes na pontuação MOS.

Adicionalmente, a fim de simular e, então, propor uma solução para esse problema, foram criadas amostras degradadas, conforme detalhado na seção 3.3, utilizando as bases de dados apresentadas na seção 3.2, pois não se dispõe de amostras com essas características em nenhuma base conhecida.

3.1.2 Detecção e eliminação dos silêncios

O VAD interno ao algoritmo da Rec. ITU-T P.563 é responsável por classificar os trechos da amostra sendo processada como voz ou ruído e também gerar informação adicional utilizada na geração do índice MOS correspondente. No entanto, sua eficiência é questionável

e, por isso, entende-se que a adoção de um algoritmo de VAD externo possa melhorar seus resultados.

Assim, para provar que o algoritmo da Rec. ITU-T P.563 é impreciso ao avaliar trechos de silêncios naturais da fala humana, 20 amostras de áudio da BD constante na Rec. ITU-T P.862 foram utilizadas. Nenhuma das amostras apresentava qualquer tipo de degradação, contendo apenas silêncios normais da conversação humana.

Para cada amostra foram gerados os índices MOS dos algoritmos das recomendações ITU-T P.862 e P.563. Na Tabela 3.1 mostra-se a média dos valores obtidos.

Tabela 3.1 – Média dos índices MOS dos áudios originais.

	P.563	P.862
Média MOS	2,7	4,5

Fonte: Nunes et al. (2016)

Sendo a Rec. ITU-T P.862 um método de referência e considerando que sua pontuação máxima é 4,5, nota-se, pelo resultado da Tabela 3.1, que o valor obtido com o algoritmo da Rec. ITU-T P.563 não tem correlação. Permite-se, assim, concluir que é necessário considerar o fator dos silêncios na avaliação de voz com o algoritmo da Rec. ITU-T P.563, a fim de corrigir a pontuação dada.

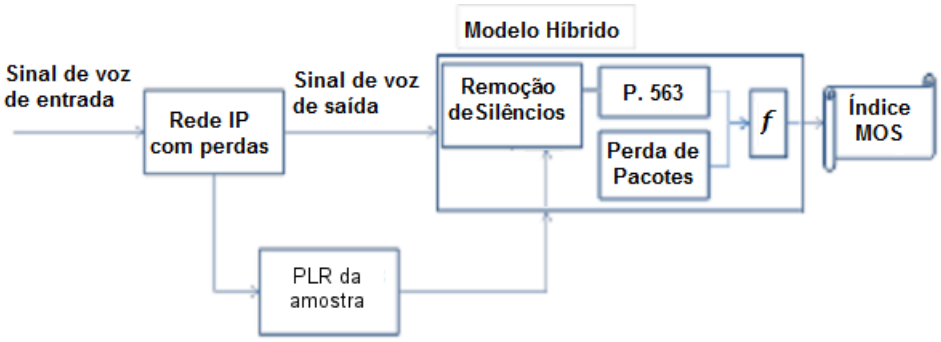
No próximo item será apresentado o modelo proposto completo em que é contemplado o devido tratamento dos silêncios naturais da fala humana, bem como a perda de pacotes.

3.1.3 Arquitetura do modelo proposto

Na Fig. 3.1 apresenta-se uma solução que utiliza um modelo híbrido que considera a remoção de silêncios do sinal de voz e as perdas de pacotes na rede, conseguindo, assim, um índice MOS mais correlacionado com os valores obtidos pelo algoritmo da Rec. ITU-T P.862.

O modelo apresentado está baseado no índice MOS, obtido pelo algoritmo da Rec. ITU-T P.563, que analisa o sinal de voz e o parâmetro de porcentagem de perda de pacotes após a remoção dos silêncios. Assim, o índice MOS obtido pelo algoritmo da Rec. ITU-T P.563 é ajustado por uma função f, descrita na seção 3.4, obtendo-se, finalmente, um índice MOS que se aproxima dos resultados obtidos pelo algoritmo da Rec. ITU-T P.862. O algoritmo apresentado na Fig. 3.2 apresenta uma visão geral da implementação da solução representada na Fig. 3.1.

Figura 3.1 – Modelo com base na Rec. ITU-T P.563.



Fonte: Pereira et al. (2015)

Figura 3.2 – Algoritmo para melhorar o desempenho do algoritmo da Rec. ITU-T P.563.

```
public float calcular(File audioFile) {
   //Vai processar, identificar e remover os trechos de silencios
   audioFile = executarVADExterno(audioFile);
   //Vai gerar o MOS utilizando o algoritmo original da Rec. ITU-T P.563
   float mos = calcularP563(audioFile);
   //Vai realizar o calculo conforme funcao de ajuste de PLR modelada
   mos = executeAdjustmentFunction(mos);
   return mos;
}
```

Fonte: Do autor (2017)

Observa-se, assim, que será utilizado um algoritmo de VAD externo ao algoritmo da Rec. ITU-T P.563 para detectar os segmentos de voz e concatená-los. Após isso, o índice MOS do áudio sem silêncios será gerado, seja pelo *Web Service* ou direto no próprio dispositivo móvel, levando-se em conta os parâmetros de rede e a função de ajuste apresentada anteriormente para melhorar o MOS final. Assim, o valor de MOS retornado será mais correlacionado com o valor dado pelo algoritmo da Rec. ITU-T P.862.

3.2 Bases de dados de amostras de voz

No desenvolvimento deste trabalho foram utilizadas três bases de dados de voz. Todas elas são amplamente utilizadas pela comunidade científica internacional em trabalhos relacionados à avaliação da qualidade de voz e são descritas na sequência.

3.2.1 *Audio eNhancement in Secured Telecom Applications*

A primeira base de dados (BD), denominada *Audio eNhancement In Secured Telecom Applications* (ANITA), foi utilizada para determinar a função de ajuste referente à PLR, que é

apresentada na Eq. 3.2 da seção 3.4. Essa BD contém arquivos de voz em diferentes idiomas, com locutores nativos e estrangeiros, e inclui gravações de homens e mulheres em condições normais, de estresse e de pânico. Além disso, estas amostras têm, em média, 28% de silêncio.

Segundo EADS Telecom (2003), estas gravações foram realizadas em ambiente de laboratório com um ruído ambiente de 26 a 54 dB. Adicionalmente, ANITA conta com arquivos gravados no interior de veículos para captação de ruídos externos, utilizando cenários alternados com as janelas abertas e fechadas e com os veículos em diferentes velocidades.

Dessa forma, as gravações têm como fontes de ruído o vento, o tráfego e o barulho do motor. Estas gravações foram armazenadas em arquivos .wav com 16 kHz, mono, e com taxa de 16 *bits* (EADS TELECOM, 2003).

As locuções dos arquivos de voz consistem em conjuntos de letras e números, sentenças foneticamente balanceadas e um texto longo com 10 minutos de duração.

3.2.2 Base de dados da Rec. ITU-T P.862

A Rec. ITU-T P.862 tem uma base de dados de voz própria com 20 arquivos sem degradação e suas respectivas cópias degradadas. Esta BD foi utilizada para modelar a Eq. 3.3, que será apresentada na seção 3.5, e suas amostras têm 8 segundos de duração e, em média, 43% de silêncio. Assim como na BD ANITA, as amostras que compõem esta BD são armazenadas em arquivos .wav, mono, com 16 *bits*, mas em uma taxa de 8 kHz de frequência.

3.2.3 Base de dados *ITU-T P.Supplement 23*

A terceira BD foi utilizada na validação da Eq. 3.3. Esta BD foi criada para auxiliar no processo de padronização da Rec. ITU-T P.861, provendo, assim, material de voz que pudesse, mais tarde, ser utilizado para validar também outras recomendações e *CODECs*.

Dessa BD foram retiradas 15 amostras com 47% de silêncios, em média, e também com 8 kHz e duração de 8 segundos.

3.3 Geração de amostras de voz degradadas por perda de pacotes

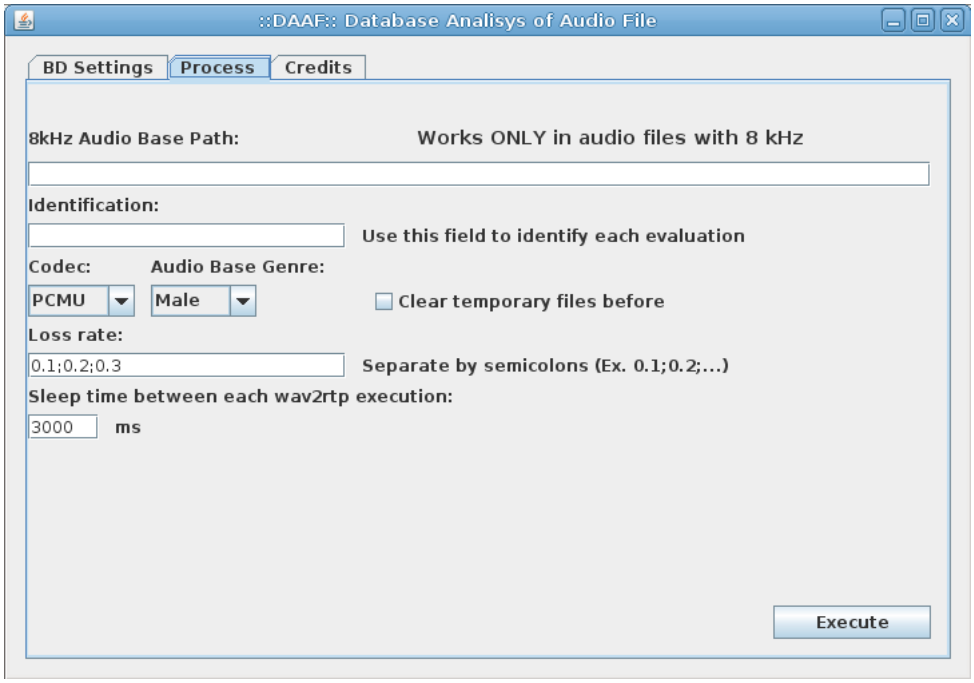
Para gerar os dados utilizados nas análises que guiaram a modelagem e a validação das funções, os arquivos de voz das três BDs foram degradados em diferentes PLRs por meio do *software* Wav2RTP (IMANKULOV, 2008). Este, por sua vez, é um emulador de redes que utiliza o protocolo RTP e processa apenas arquivos de 8 kHz de frequência.

Dessa forma, a partir de um arquivo de voz ora denominado original, foi possível obter suas versões degradadas com diversas taxas de perda e, assim, submetê-los aos *softwares* das recomendações ITU-T P.862 e P.563 para a obtenção dos índices MOS.

Contudo, degradar arquivo por arquivo de cada BD para cada PLR representaria um enorme trabalho mecânico de repetição de todos esses comandos. Assim, percebeu-se a necessidade de desenvolver um *software*, cujo propósito foi automatizar a execução do Wav2Rtp, e dos algoritmos das recomendações ITU-T P.862 e P.563.

Este *software* foi desenvolvido utilizando-se a linguagem Java com a API *Swing*, em conjunto com o Sistema de Gerenciamento de Banco de Dados (SGBD) MySql e a ferramenta para geração de relatórios iReport. Na Fig. 3.3 observa-se a tela principal do sistema, o qual está disponível para *download* gratuitamente em <https://github.com/rdantasnunes/AudioBaseReport>.

Figura 3.3 – *Software* desenvolvido e utilizado para automatizar geração dos valores MOS.



Fonte: Pereira et al. (2015)

Em síntese, esse *software* recebe o diretório de origem da BD a ser degradada, uma identificação textual a ser atribuída à presente execução, o gênero do locutor, o *CODEC*, uma lista das taxas de perda e um valor, em milésimos de segundo, que determina o intervalo entre as execuções do Wav2Rtp. Terminada a execução do *software*, são armazenados na base de dados do mesmo os dados gerais relativos à configuração utilizada e dados específicos de cada arquivo.

Além disso, o sistema permite a emissão de relatório com as taxas de perda e os valores MOS gerados para cada arquivo, possibilitando, assim, a análise dos resultados. Dessa forma, além de automatizar o processo de execução dos softwares em questão, foi possível a criação de cenários com diferentes taxas de perda de forma rápida e fácil.

Com todos esses dados disponíveis no SGBD, o processo de análise e geração de resultados ganhou em dinamicidade, pois, por meio da linguagem *Structured Query Language* (SQL) foi possível realizar cálculos e processamentos sem a necessidade de executar novamente os *softwares* para avaliação e geração dos índices em momentos posteriores aos da execução dos *softwares* citados.

3.4 Função de ajuste para perda de pacotes

A função de ajuste de perda de pacotes, denominada f, foi modelada por meio de testes experimentais, nos quais foram utilizados 27 arquivos de áudio da base de dados ANITA. Cada arquivo de áudio foi degradado inserindo-se 30 diferentes probabilidades de perda de pacotes, as quais são de 0,5% até 10% com intervalos de 0,5% e de 11% até 20% com intervalos de 1%.

Assim, para obter um arquivo degradado, utilizou-se a ferramenta Wav2Rtp, a qual foi executada 100 vezes para cada arquivo, considerando um desvio padrão de 2,5%. Dessa forma, um total de 810 arquivos degradados foi analisado pelos algoritmos das recomendações ITU-T P.563 e P.862. Com os resultados obtidos foi criada a função de ajuste f, conforme mostrado na Equação 3.1. Considere, então, que, por exemplo, para o cenário de perda de pacotes de $n\%$, considerando também o arquivo de áudio 1 (Arq-1), tem-se

$$f_{Arq-1}^{n\%} = \frac{MOS\ P.862_{Arq-1}^{n\%}}{MOS\ P.563_{Arq-1}^{n\%}} \tag{3.1}$$

em que MOS P.862 e MOS P.563 representam os valores de índice MOS obtidos pelos algoritmos das recomendações ITU-T P.862 e P.563, respectivamente. Assim, o valor de $f_{medio}^{n\%}$ é a média aritmética dos valores f obtidos para cada um dos 27 arquivos degradados com $n\%$ de perda de pacotes. Como foram considerados 30 valores possíveis para n, a função f' discreta é definida como: $f' = [f_{medio}^{0,5\%}, f_{medio}^{1\%}, ..., f_{medio}^{30\%}]$.

No intuito de que f' não esteja restrita somente a 30 casos de perda de pacotes, f'' foi utilizada para modelar a função $f(n)$, utilizando regressão linear, com a seguinte função polinomial:

$$f(n) = \alpha * n^3 + \beta * n^2 + \gamma * n + D \tag{3.2}$$

em que α, β, γ e D são fatores constantes determinados conforme demonstrado na seção 4.1.1 e n é a taxa de perda de pacotes na rede IP correspondente à amostra sendo avaliada.

3.5 Detecção e supressão de segmentos silenciosos usando um algoritmo de VAD

Como apontado anteriormente, se fez necessário utilizar um VAD externo ao algoritmo da Rec. ITU-T P.563. Assim, foi selecionado o algoritmo de Sadjadi e Hansen (2013) porque, assim como o VAD usado na Rec. ITU-T P.563, este utiliza o método de *power threshold*.

O VAD externo é utilizado na primeira fase do modelo proposto para suprimir os silêncios. Contudo, havendo perda de pacotes, algumas informações do sinal são perdidas e criam-se também mais trechos de silêncios. Desse modo, foi necessário regular o comprimento máximo de silêncios em milissegundos, em função da PLR, para, assim, equilibrar o nível de energia no sinal.

A fim de estabelecer um limite para o comprimento dos segmentos de silêncios, vários testes foram realizados. Então, 20 amostras de áudio da BD da Rec. ITU-T P.862 foram degradadas utilizando-se 31 diferentes PLRs. Estas taxas vão de 0% até 10%, com intervalos de 0,5%, e de 11% até 20% com intervalos 1%.

Este processo resultou em 620 amostras de voz que foram avaliadas pelos algoritmos das recomendações ITU-T P.563, P.862 e da solução proposta. A execução do SQV$_{mobile}$, realizada para modelar e, posteriormente, validar o modelo se deu em um *smartphone* convencional, o qual é facilmente adquirido no comércio deste tipo de dispositivo.

Com os resultados, a função nomeada SS_{PLR} foi determinada, a qual é uma Função de Heaviside, conhecida como um exemplo da classe geral de funções de degrau. É importante observar que outras funções, lineares e exponenciais, foram testadas, mas apresentaram resultados menos eficientes.

$$SS_{PLR} = A_0(\frac{1}{2}(1+sgn(x-a_0))) + A_1(\frac{1}{2}(1+sgn(x-a_1))) + ...A_n(\frac{1}{2}(1+sgn(x-a_n))) \quad (3.3)$$

em que $SS_{PLR}(x)$ é o comprimento em milessegundos do segmento de silêncio máximo permitido na amostra, x é o parâmetro da PLR e n é o número ideal de taxas de perda a serem determinadas. $A_0, A_1, ...A_n$ e $a_0, a_1, ...a_n$ são fatores constantes a serem também determinados em testes experimentais.

3.6 Implementação do algoritmo proposto

A fim de permitir que o algoritmo proposto seja implementado e validado, foi reproduzido um cenário em que um usuário realiza uma chamada VoIP para outro e, em ambas as extremidades, o referido aplicativo captura uma amostra do áudio recebido.

É importante notar que a amostra capturada tem apenas o sinal do áudio que está entrando, o qual sofreu degradação da rede IP e, assim, irá ser submetido ao algoritmo proposto. Dessa forma, o algoritmo será implementado em dois formatos: por meio de um aplicativo móvel, desenvolvido na plataforma *Android* e por um *Web Service*, desenvolvido na linguagem Java. Ambas as implementações serão abordadas a seguir.

3.6.1 Implementação do algoritmo em dispositivo móvel

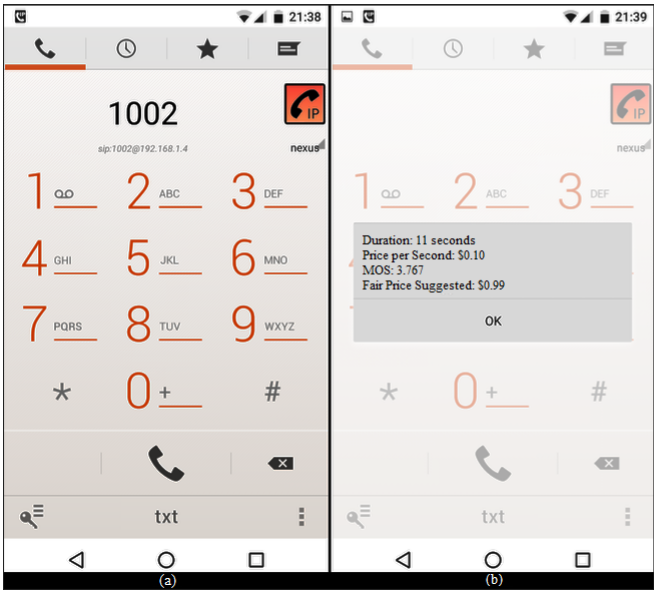
Com o objetivo de testar todo o modelo em uma aplicação VoIP funcional, a implementação do SQV_{mobile} com o algoritmo proposto foi feita no aplicativo móvel de código aberto *CSIPSimple*.

Na Fig. 3.4 observa-se a tela da aplicação móvel que se conecta a outro telefone em uma chamada VoIP e calcula o índice MOS com o algoritmo proposto ao final.

Para tornar possível a utilização do algoritmo da Rec. ITU-T P.563 em um dispositivo móvel, que é escrito na linguagem C, foram utilizados, em conjunto a Interface Nativa Java (*Java Native Interface*, JNI) e o *Kit* de Desenvolvimento Nativo (*Native Development Kit*, NDK). Para realizar a chamada VoIP, o *User Datagram Protocol* (UDP) foi utilizado como protocolo de transporte, a fim de garantir que as condições seriam semelhantes ao sistema real.

Ao final da ligação, o aplicativo apresenta o índice MOS calculado e também o preço justo sugerido da chamada, conforme os parâmetros indicados no trabalho de Rodriguez, Rosa

Figura 3.4 – Tela da aplicação móvel implementando o algoritmo proposto. (a) Tela inicial de discagem; (b) ao final da ligação; quando os dados são apresentados.



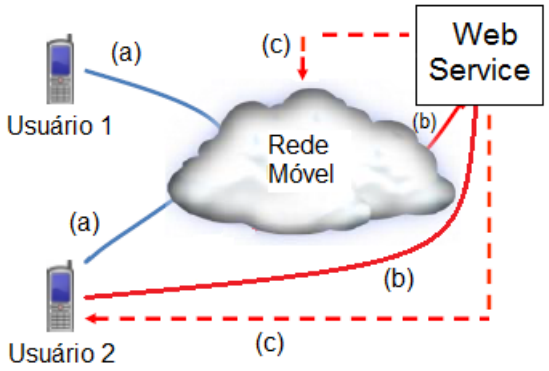
Fonte: Do autor (2017)

e Bressan (2013). O valor da tarifa a ser paga é hipotético, mas o aplicativo pode ser alterado, de forma que usuário configure esta informação, de acordo com seu plano.

3.6.2 Execução do algoritmo como *Web Service*

Para a segunda alternativa, na Fig. 3.5 está demonstrado o modo de operação do *Web Service* e como ele pode interagir com as aplicações que são suas potenciais clientes.

Figura 3.5 – Exemplo de uso da comunicação do *Web Service* com outros softwares em diferentes plataformas.



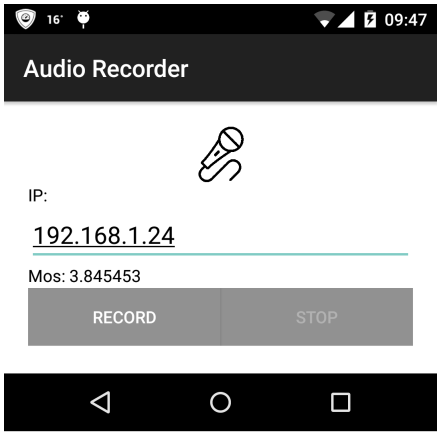
Fonte: Nunes et al. (2016)

Em (a), dois usuários estão em uma chamada VoIP, podendo esta ser de qualquer tipo de serviço (SMP, IMS, Skype, WhatsApp, etc.) e, assim, tanto o provedor do serviço quanto o dis-

positivo de um ou ambos os usuários (b) pode capturar uma amostra do sinal de voz degradado pela rede e enviá-lo ao *Web Service* que devolve (c) o índice MOS da referida amostra a quem o executou.

Como pode ser observado, a grande vantagem no uso do *Web Service* é que ele é multi-plataforma e independente de linguagem de programação.

Figura 3.6 – Aplicação *Android* utilizada para testar o *Web Service*.



Fonte: Nunes et al. (2016)

Para consumir e testar o *Web Service*, validando assim, a arquitetura de *software* proposta, foi utilizado o aplicativo **Android** da Fig. 3.6. Nesse aplicativo de teste é possível realizar uma gravação simples, enviá-la ao *Web Service* que, então, processa e devolve o índice MOS.

3.7 Materiais

Para a realização das atividades que viabilizaram o produto deste trabalho, foram utilizados diversos *softwares*, todos disponibilizados gratuitamente, conforme a Tabela 3.2.

Tabela 3.2 – *Softwares* necessários para a realização do projeto

Nome do *software*	**Versão**	**Função no projeto**
NetBeans	8.0.2	IDE para desenvolvimento da ferramenta de geração de amostras degradadas
Android Studio	1.5	IDE para desenvolvimento da aplicação *Android*
Eclipse IDE	Luna	IDE para desenvolvimento do serviço *web* usando a plataforma Java EE
Astah *Community*	6.8.0	Ferramenta de UML para a diagramação do *software*
ITU-P.563	10/2007	*Software* escrito na linguagem C e de código aberto
ITU-P.862	10/2007	*Software* escrito na linguagem C e de código aberto
Wav2rtp	0.9.1	Emulador de redes livre e de código aberto

Fonte: Do autor (2017)

Os equipamentos utilizados foram um *notebook*, mas este pode ser substituído por um computador *desktop* convencional, desde que seja capaz de executar os *softwares* da Tabela 3.2, e dois *smartphones* com o sistema operacional *Android*, para executar o aplicativo móvel com o SQV_{mobile}.

4 RESULTADOS E DISCUSSÃO

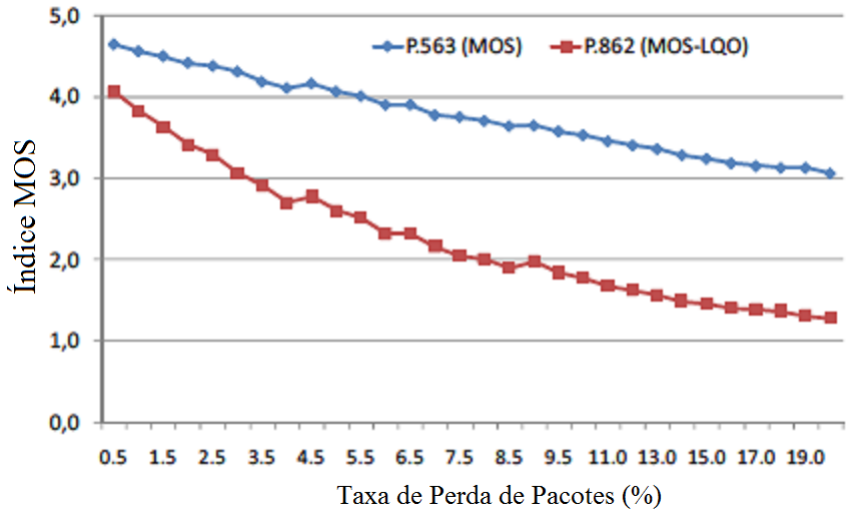
Os resultados alcançados são provenientes dos experimentos realizados, bem como do desenvolvimento do *Web Service* e do SQV_{mobile}, que deram suporte à realização deste trabalho. A seguir são apresentados os resultados dos experimentos obtidos com a função de ajuste, definida na seção 3.4 e, posteriormente, da remoção de silêncios naturais da fala humana, conforme seção 3.5.

4.1 Função de ajuste para perda de pacotes

Como citado anteriormente, a literatura destaca que a Rec. ITU-T P.862 é um dos métodos mais confiáveis para avaliação da qualidade de voz. Assim, testes preliminares foram realizados comparando-se os valores de MOS obtidos dos algoritmos das recomendações ITU-T P.563 e P.862, tanto para modelar quanto para validar a função.

Na Fig. 4.1 mostra-se um exemplo desta diferença, em que o MOS do arquivo m_27_en_c_se06 da BD ANITA foi gerado para alguns dos cenários de perda de pacotes. Outros arquivos da mesma base foram analizados e obtiveram resultados diferentes, mas com um padrão de comportamento semelhante.

Figura 4.1 – Valores de MOS obtidos por meio dos algoritmos das recomendações ITU-T P.563 e P.862 para 30 cenários de perda de pacotes.



Fonte: Pereira et al. (2015)

4.1.1 Modelagem da função de ajuste para perda de pacotes

Utilizou-se a Eq. 3.1, e os valores médios obtidos para cada cenário de perda de pacote foram $f'' = [1.02; 0.98; 0.94; 0.91; 0.88; 0.84; 0.81; 0.79; 0.76; 0.74; 0.73; 0.71; 0.70; 0.68; 0.67; 0.65; 0.64; 0.63; 0.62; 0.61; 0.60; 0.58; 0.57; 0.56; 0.55; 0.54; 0.53; 0.53; 0.52;0.52]$.

Com os valores de f'', a função $f(n)$ foi modelada utilizando-se a Eq. 3.2, em que $\alpha = -2x10^{-5}; \beta = 0.001; \gamma = -0.043; D = 1.059$.

Na Tabela 4.1 apresentam-se por meio do PCC e do Erro Absoluto Máximo (*Maximum Absolute Error*, MAE), os resultados da avaliação de desempenho do algoritmo da Rec. ITU-T P.563, com a função $f(n)$ e sem a mesma.

Tabela 4.1 – Avaliação de desempenho da função de ajuste proposta.

	PCC	MAE
P.563 Original vs P.862	0,9606	1,392
P.563 Ajustado com $f(n)$ vs P.862	0,9973	0,406

Fonte: Pereira et al. (2015)

É importante notar também que a função $f(n)$ pode aceitar diferentes valores de PLR, os quais são monitorados e extraídos de uma rede IP real. Além disso, a solução proposta tem baixa complexidade e pode ser usada para avaliar serviços de tempo real como VoIP. Apesar disso, consome baixos recursos de processamento, considerando-se as atuais características dos dispositivos eletrônicos.

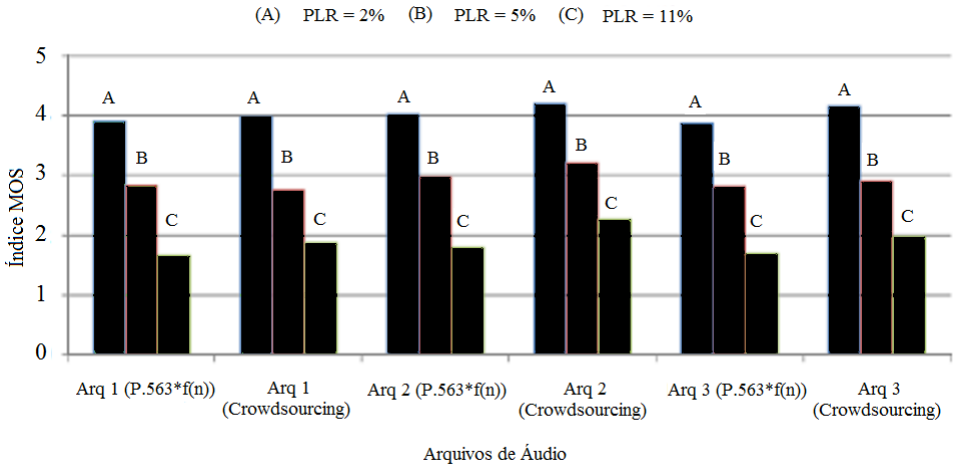
4.1.2 Validação da função de ajuste para perda de pacotes

A fim de validar o desempenho da solução proposta, três diferentes arquivos da BD ANITA foram avaliados. Cada um deles foi degradado com três diferentes cenários de taxas de perda de pacotes, que foram de 2%, 5% e 11%, obtendo-se nove diferentes arquivos de voz para serem testados. Assim, foram realizados testes subjetivos remotos utilizando-se o método de *crowdsourcing*, mais especificamente empregando-se uma plataforma comercial. Para esse teste, uma interface *web* foi construída, em que as instruções dos testes foram dadas e os arquivos para serem testados disponibilizados.

Cada arquivo de voz foi avaliado por 60 usuários remotos e, como pode ser observado na Fig. 4.2, os valores de índices MOS dados pelos usuários remotos são similares aos valores estimados pela solução proposta neste trabalho. Nesse caso, o MAE foi de 0,47, considerando

os mesmos cinco pontos da escala MOS. É importante observar que há poucas amostras para determinar um valor de PCC confiável.

Figura 4.2 – Avaliação de desempenho da solução proposta em relação aos testes subjetivos, com 9 cenários de perda de pacotes.



Fonte: Pereira et al. (2015)

Além disso, todo o conjunto da solução proposta foi validado com um número maior de amostras, como será detalhado nas seções a seguir.

4.2 Remoção de silêncios naturais da fala humana

As variáveis na função $SS_{PLR}(x)$, definida na Eq. 3.3, foram determinadas com base nos resultados dos testes experimentais realizados no processo de modelagem e confirmados em um segundo momento no processo de validação.

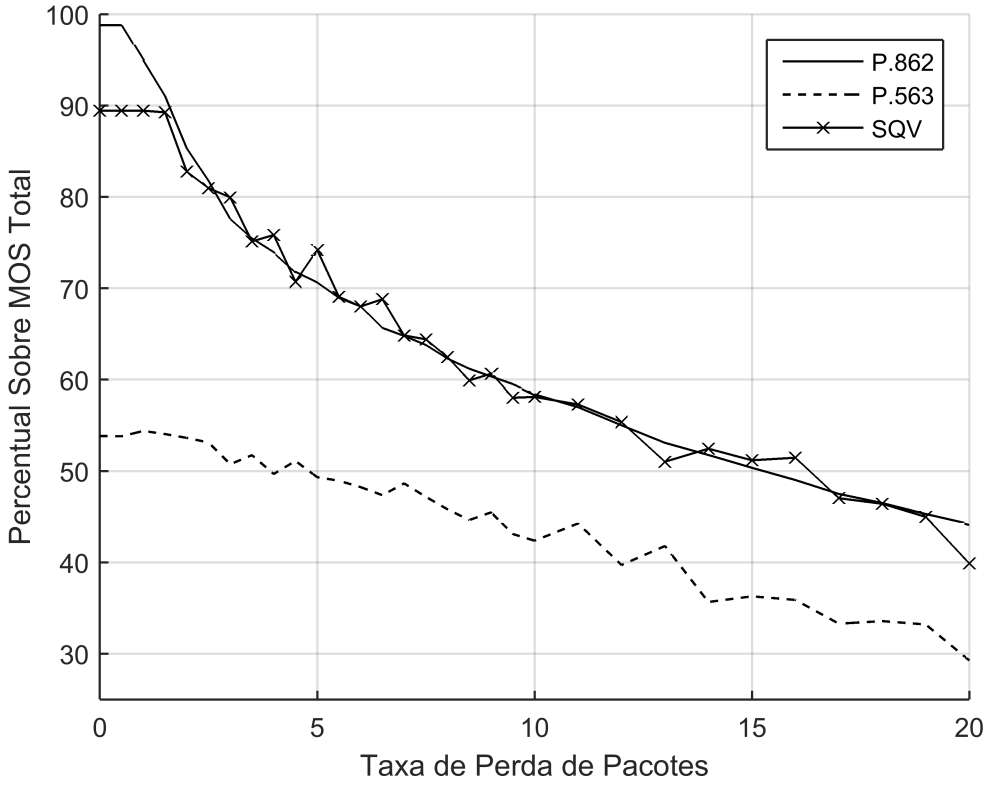
Como citado, essa função foi utilizada para estabelecer um limite para o comprimento dos segmentos de silêncios dentro do algoritmo de VAD externo, com base na PLR da rede. Tanto a modelagem quanto a validação desta função foram realizadas considerando o valor MOS de saída da solução proposta como um todo, incluindo a aplicação da Função de Ajuste, pois a amostra resultante do VAD externo, em que se aplica a referida função $SS_{PLR}(x)$, impacta em todas as demais etapas da solução. Desse modo, essas etapas de modelagem e validação já contemplam toda a solução proposta neste trabalho.

4.2.1 Modelagem do VAD externo

Nesta etapa, os valores obtidos por meio dos testes realizados foram n=9, A_0=200, A_1=700, A_2=100, A_3=-100, A_4=-700, A_5=-110, A_6=-8, A_7=-2, A_8=-20, A_9=-56, a_0=0, a_1=2,9, a_2=3,4, a_3=3,9, a_4=4,1, a_5=6,4, a_6=9,4, a_7=16, a_8=15 and a_9=16.

Os resultados obtidos pelo SQV_{mobile}, que considera a função SS_{PLR} em conjunto com $f(n)$, são apresentados na Fig. 4.3, em que a comparação dos resultados é feita pelo percentual sobre a pontuação máxima da escala de qualidade usada em cada método de avaliação.

Figura 4.3 – Média do percentual sobre o valor total da escala de cada algoritmo obtido usando a BD da Rec. ITU-T P.862 por meio dos algoritmos das recomendações ITU-T P.862, P.563 e SQV, para as 620 amostras.



Fonte: Do autor (2017)

Os valores de PCC e RMSE obtidos pelo SQV, em comparação com o algoritmo da Rec. ITU-T P.862, são apresentados na Tabela 4.2.

4.2.2 Validação do VAD externo

Na sequência, a BD ITU-T *P.Supplement 23* foi utilizada para validar os valores modelados. Nesta etapa, 15 amostras foram degradadas em 31 diferentes PLRs de 0% a 20%, como

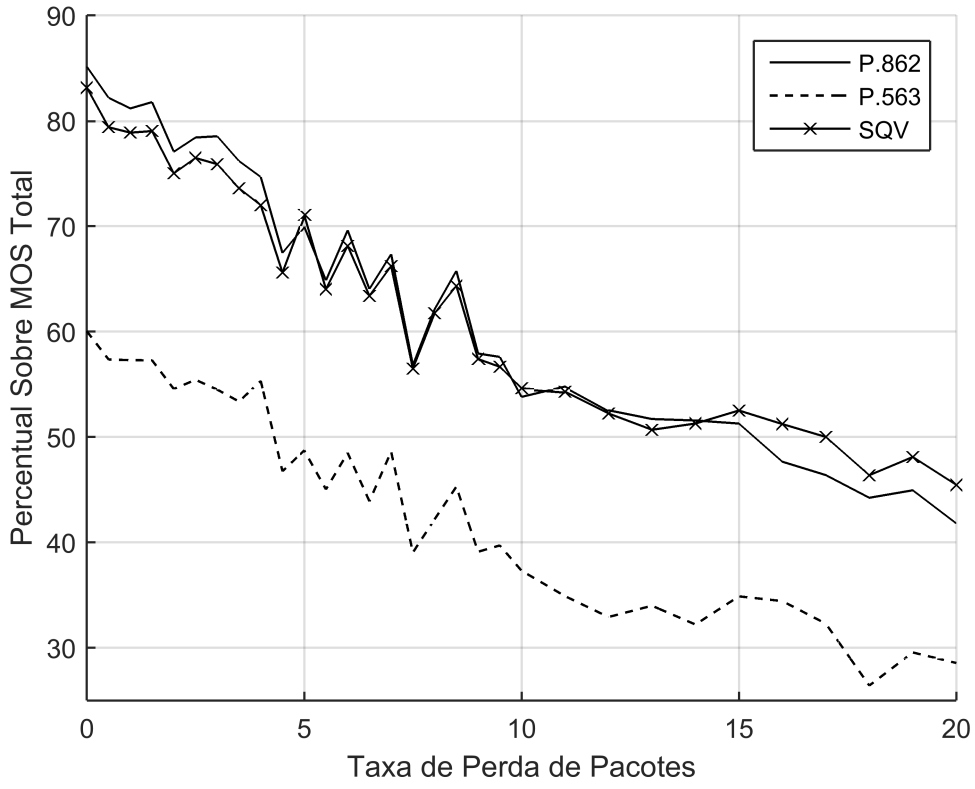
Tabela 4.2 – Avaliação de desempenho da solução proposta como um todo na etapa de modelagem.

	PCC	RMSE
P.563 vs P.862	0,9139	0,8238
SQV vs P.862	0,9846	0,3171

Fonte: Do Autor (2017)

na etapa anterior. Este processo resultou em um total de 465 amostras. Assim, na Fig. 4.4 é possível observar que os valores resultantes são mais próximos aos valores do algoritmo da Rec. ITU-T P.862, que os do algoritmo da Rec. ITU-T P.563.

Figura 4.4 – Média do percentual sobre o valor total da escala de cada algoritmo obtido usando a BD *P.Supplement 23*, por meio dos algoritmos das recomendações ITU-T P.862, P.563 e SQV, para as 465 amostras.



Fonte: Do autor (2017)

Na Tabela 4.3 é possível ver, por meio dos valores de PCC e RMSE, que o SQV atingiu melhor desempenho que o algoritmo da Rec. ITU-T P.563, confirmando, assim, a hipótese inicial deste trabalho, de que é possível melhorar o desempenho do referido algoritmo tratando os fatores de perda de pacote e silêncios naturais da fala.

Tabela 4.3 – Avaliação de desempenho da solução proposta como um todo na etapa de validação.

	PCC	RMSE
P.563 vs P.862	0,9913	0,6758
SQV vs P.862	0,9957	0,2983

Fonte: Do Autor (2017)

Além de avaliar a precisão da solução proposta, considerou-se importante avaliar também o impacto do algoritmo proposto ao ser implementado em um dispositivo móvel. Assim, alguns testes de desempenho foram realizados e são descritos na seção seguinte.

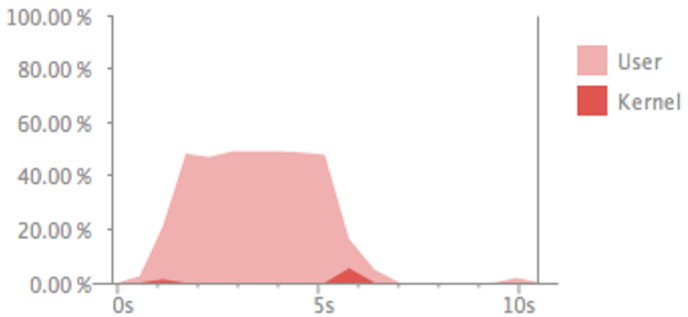
4.3 Testes de consumo de recursos em dispositivos móveis

Para avaliar o desempenho do algoritmo implementado, foi desenvolvida uma versão da aplicação que contava apenas com tal algoritmo, sem a solução completa que realiza as chamadas de VoIP.

Para monitorar o processador, a memória e o consumo de bateria durante a execução do SQV_{mobile}, uma ferramenta de análise de perfil nativa da *IDE Android Studio* foi utilizada. Os testes foram realizados em um dispositivo móvel comum com um processador *Quad-Core* de 2,3 GHz, 2 GB de memória e bateria de 2300 mAh de capacidade com voltagem nominal de 3,8V.

Na Fig. 4.5 observa-se o consumo de processamento ao analisar um segmento de voz com comprimento de 8 segundos.

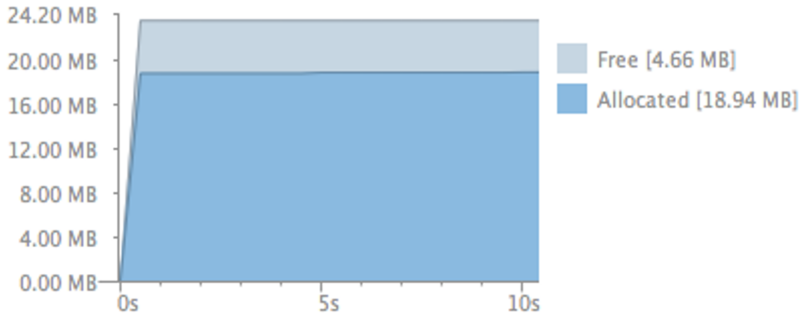
Figura 4.5 – Consumo de processamento ao analisar amostra com o SQV.



Fonte: Do autor (2017)

Na Fig. 4.6 estão demonstrados o consumo de memória para abrir a aplicação e o respectivo uso do SQV_{mobile} durante a sua execução. É importante observar que, uma vez a aplicação aberta, a execução do algoritmo não consome memória adicional.

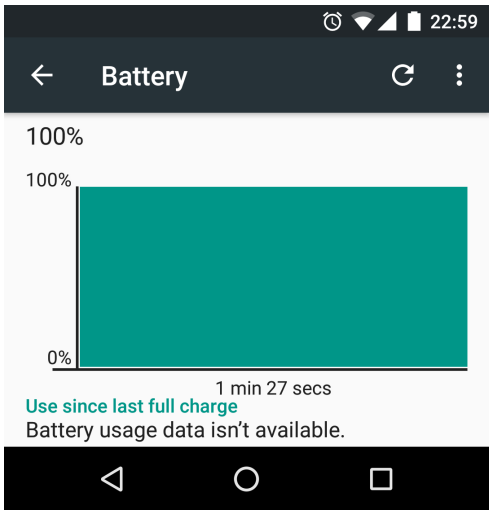
Figura 4.6 – Consumo de memória ao analisar amostra com o SQV.



Fonte: Do autor (2017)

Na sequência, foi empregada uma ferramenta para coletar informações sobre o consumo de bateria, que é nativa do sistema operacional *Android*. Esta ferramenta não pode medir quanta energia a aplicação usa para executar, mas pode determinar a energia total consumida pelo dispositivo durante um intervalo de tempo. Assim, com 100% de carga da bateria, o aplicativo é aberto e usado por 87 segundos, executando o algoritmo contra as amostras em intervalos de 5 segundos. Finalmente, a bateria estava até então carregada a 100%, como mostrado na Fig. 4.7.

Figura 4.7 – Condição da bateria após execução da aplicação móvel com o SQV.



Fonte: Do autor (2017)

4.4 Discussão

A efetividade da solução proposta neste trabalho é confirmada comparando-se seus resultados ao de outros trabalhos publicados, apresentados a seguir.

4.4.1 Comparação com outros trabalhos

Mossavat, Petkov e Klejin (2011) trabalharam o problema identificado por eles sobre a heterogeneidade dos dados. Como solução, adotam um *Bayesian Framework* (BF) e seus resultados mostram um índice médio de PCC de 0,89 e um RMSE de 0,3225. No entanto, a métrica utilizada como referência por eles não foi especificada.

Os estudos relacionados a seguir trabalharam na solução de problemas relacionados apenas à perda de pacotes usando diferentes métodos.

Dubey e Kumar (2015) e Falk e Chan (2006) utilizaram GMM e obtiveram um PCC de 0,95 e 0,82, e RMSE de 0,14 e 0,30, respectivamente. Ambos os trabalhos geraram PCC, comparando suas soluções com MOS de métricas objetivas.

Cherif et al. (2012) utilizaram uma RNA e obtiveram um PCC de 0,98, comparando sua solução a valores dados pelo algoritmo da Rec. ITU-T P.862.

Raja et al. (2007) empregaram programação genética em seu trabalho e obtiveram índices de PCC e RMSE de 0,96 e 0,038, respectivamente.

Os resultados agrupados dos trabalhos descritos são mostrados na Tabela 4.4.

Tabela 4.4 – Comparação entre outros trabalhos e seus resultados.

Autor(es)	Método	PCC	RMSE
Mossavat, Petkov e Klejin (2011)	BF	0,89	0,3225
Dubey e Kumar (2015)	GMM	0,95	0,14
Falk e Chan (2006)	GMM	0,82	0,3
Cherif et al. (2012)	RNA	0,98	-
Raja et al. (2007)	PG	0,96	0,038
Modelo Proposto - SQV	Função de ajuste e remoção de silêncios	0,99	0,29

Fonte: Do Autor (2017)

Nenhum dos trabalhos citados tratou o fator dos silêncios e, por isso, é difícil fazer uma comparação direta, principalmente porque os valores obtidos podem variar em função da BD e da metodologia utilizada nos processos de modelagem e validação. Assim, uma comparação de desempenho somente baseada nos valores de PCC e RMSE pode não ser apropriada.

Apesar disso, observando-se os resultados de outros autores e os comparando-os com os deste trabalho, observa-se que os valores obtidos pela solução proposta são melhores em alguns casos e muito próximo de outros.

Deve ser destacado que o modelo proposto é limitado a cenários de perda de pacotes na rede IP da ordem de, no máximo, 20%, não tendo limite inferior.

5 CONCLUSÃO

Conclui-se, por meio deste trabalho, que os silêncios naturais da fala humana devem ser considerados em algoritmos objetivos e não intrusivos, especificamente em serviços de comunicação. Este fator impacta de forma significativa no desempenho da avaliação da qualidade de voz porque, ao considerá-lo, os resultados são mais confiáveis e próximos aos das métricas intrusivas, consideradas como referência. Assim, o monitoramento por parte das operadoras de telefonia poderá ser mais eficiente na garantia da QoE dos usuários.

O desempenho do algoritmo da Rec. ITU-T P.563 foi avaliado e, para melhorar seu desempenho, duas funções foram determinadas. A primeira foi nomeada $f(n)$ e é uma função de ajuste, cujo objetivo é aproximar os resultados da solução proposta aos do índice MOS dado pelo algoritmo da Rec. ITU-T P.862 em cenários com perda de pacotes. A segunda função foi nomeada $SS_{PLR}(X)$, a qual, por meio do PLR, determina o comprimento máximo do segmento de silêncio, em milésimos de segundo, permitido na amostra de voz sendo analisada.

A avaliação de desempenho de toda a solução, composta pelas duas funções, atingiu resultados seguros, como mostrado na Fig. 4.4, com um índice de PCC de 0,9957 e RMSE de 0,2983. Deve-se destacar também que a base de dados de áudio utilizada nos testes de desempenho é publicamente acessível e dada pela ITU-T como uma referência para a criação de métricas de avaliação de voz.

Finalmente, o SQV$_{mobile}$ é um meio de baixo custo de prover monitoramento em tempo real que pode melhorar indiretamente a QoE dos usuários. Além disso, aplicações móveis são muito comuns hoje em dia e a adoção deste tipo de solução é extremamente viável. O SQV$_{mobile}$ tem uma interface amigável e apresenta baixo nível de complexidade. Portanto, a solução pode ser implementada em dispositivos móveis atuais.

Como sequência deste trabalho sugere-se a continuidade de experimentos visando encontrar novas deficiências no algoritmo da Rec. ITU-T P.563, uma vez que a literatura é rica em publicações apontando diversas limitações em cenários de VoIP.

REFERÊNCIAS

ABAREGHI, M. et al. Improved ITU-T P.563 non-intrusive speech quality assessment method for covering voip conditions. In: INTERNATIONAL CONFERENCE ON ADVANCED COMMUNICATION TECHNOLOGY, 10., 2008, Gangwon-Do. **Proceedings...** Gangwon-Do: IEEE, 2008. p. 354–357.

ABDULLAH-AL-MAMUN, K.; SARKER, F.; MUHAMMAD, G. A high resolution pitch detection algorithm based on AMDF and ACF. **Journal of Scientific Research**, Bangladesh, v. 1, n. 3, p. 508–515, Aug. 2009.

ANTONS, J. N. et al. Impact of perceived quality and other influencing factors on emotional video experience. In: IEEE INTERNATIONAL WORKSHOP ON QUALITY OF MULTIMEDIA EXPERIENCE, 70., 2015, Pylos-Nestor. **Proceedings...** Pylos-Nestor: IEEE, 2015. p. 1–6.

BENYASSINE, A. et al. ITU-T Recommendation G.729 Annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications. **IEEE Communications Magazine**, New York, v. 35, n. 9, p. 64–73, Aug. 1997.

BRASIL. Agência Nacional de Telecomunicações. **Relatório anual 2015**. Brasília, DF, 2016. 76 p.

CHERIF, W. et al. A_PSQA: PESQ-like non-intrusive tool for QoE prediction in VoIP services. In: IEEE INTERNATIONAL CONFERENCE ON COMMUNICATIONS, 2012, Ottawa. **Proceedings...** Ottawa: IEEE, 2012. p. 2124–2128.

DEITEL, P.; DEITEL, H. **Java**: how to program. 9^{th} ed. Boston: Prentice Hall, 2012. 1535 p.

DUBEY, R. K.; KUMAR, A. Non-intrusive speech quality assessment using multi-resolution auditory model features for degraded narrowband speech. **IET Signal Processing**, London, v. 9, n. 9, p. 638–646, Dec. 2015.

EADS TELECOM. **Audio enhancement in secured telecom applications reference database description**. Boston, 2003. 57 p.

FALK, T. H.; CHAN, W. Y. Enhanced non-intrusive speech quality measurement using degradation models. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, 2006, Toulouse. **Proceedings...** Toulouse: IEEE, 2006. p. 1.

FIEDLER, M. et al. Quality of experience. **Annals of Telecommunications**, New York, v. 65, n. 1, p. 1–2, Feb. 2010.

FIELDING, R. T. **Architectural styles and the design of network-based software architectures**. 2000. 180 p. Thesis (Ph.D. in Information and Computer Science)-University of California, Irvine, 2000.

GOOGLE. **A História do Android**. Mountain View, 2015. Disponível em: <http://www.android.com/intl/ptBR_br/history/>. Acesso em: 22 ago. 2015.

GRANCHAROV, V. et al. Low-complexity, nonintrusive speech quality assessment. **IEEE Transactions on Audio, Speech, and Language Processing**, New York, v. 14, n. 6, p. 1948–1956, Nov. 2006.

HAN, Y.; MUNTEAN, G. M. Hybrid real-time quality assessment model for voice over IP. In: IEEE INTERNATIONAL SYMPOSIUM ON BROADBAND MULTIMEDIA SYSTEMS AND BROADCASTING, 10., 2015, Ghent. **Proceedings...** Ghent: IEEE, 2015. p. 1–6.

HARTPENCE, B. **Packet guide to voice over IP**: a system administrator's guide to VoIP technologies. Sebastopol: O'Reilly Media, 2013. 242 p.

HOBFELD, T.; HEEGAARD, P. E.; VARELA, M. QoE beyond the MOS: added value using quantiles and distributions. In: INTERNATIONAL WORKSHOP ON QUALITY OF MULTIMEDIA EXPERIENCE, 70., 2015, Pylos-Nestor. **Proceedings...** Pylos-Nestor: IEEE, 2015. p. 1–6.

IMANKULOV, R. **Wav2RTP**. Porto, 2008. Disponível em: <http://wav2rtp.sourceforge.net/index.html>. Acesso em: 11 fev. 2017.

INSTITUTE OF ELECTRICAL AND ELECTRONIC ENGINEERS. **IEEE Std 3333.1.1-2015**: IEEE standard for quality of experience (QoE) and visual-comfort assessments of three-dimensional (3D) contents based on psychophysical studies. New York, 2015. 46 p.

INTERNATIONAL TELECOMMUNICATION UNION - TELECOMMUNICATION STANDARDIZATION SECTOR. **Framework of a Service Level Agreement**. Genebra, 2002. Disponível em: <https://www.itu.int/rec/T-REC-E.860-200206-I/en>. Acesso em: 23 dez. 2016.

INTERNATIONAL TELECOMMUNICATION UNION - TELECOMMUNICATION STANDARDIZATION SECTOR. **Method for the subjective assessment of intermediate quality levels of coding systems**. Genebra, 2015. Disponível em: <http://www.itu.int/rec/R-REC-BS.1534>. Acesso em: 31 jan. 2017.

INTERNATIONAL TELECOMMUNICATION UNION - TELECOMMUNICATION STANDARDIZATION SECTOR. **Methods for subjective determination of transmission quality**. Genebra, 1996. Disponível em: <http://www.itu.int/rec/T-REC-P.800/en>. Acesso em: 23 dez. 2016.

INTERNATIONAL TELECOMMUNICATION UNION - TELECOMMUNICATION STANDARDIZATION SECTOR. **Perceptual evaluation of speech quality (PESQ)**: an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. Genebra, 2001. Disponível em: <http://www.itu.int/rec/T-REC-P.862/en>. Acesso em: 23 dez. 2016.

INTERNATIONAL TELECOMMUNICATION UNION - TELECOMMUNICATION STANDARDIZATION SECTOR. **Perceptual Objective Listening Quality Assessment (POLQA)**. Genebra, 2014. Disponível em: <http://www.itu.int/rec/T-REC-P.863/en>. Acesso em: 23 dez. 2016.

INTERNATIONAL TELECOMMUNICATION UNION - TELECOMMUNICATION STANDARDIZATION SECTOR. **Pulse code modulation of voice frequencies**. Genebra, 1993. Disponível em: <https://www.itu.int/rec/T-REC-G.711/en>. Acesso em: 31 jan. 2017.

INTERNATIONAL TELECOMMUNICATION UNION - TELECOMMUNICATION STANDARDIZATION SECTOR. **Single-ended method for objective speech quality asssessment in narrow-band telephone applications**. Genebra, 2004. Disponível em: <http://www.itu.int/rec/T-REC-P.563/en>. Acesso em: 23 dez. 2016.

INTERNET ENGINEERING TASK FORCE. **Request for comments 7159**: Javascript object notation (JSON) data interchange format. Fremont, 2014. Disponível em: <https://tools.ietf.org/html/rfc7159>. Acesso em: 20 jan. 2016.

KUIPERS, F. et al. Techniques for measuring quality of experience. In: INTERNATIONAL CONFERENCE ON WIRED/WIRELESS INTERNET COMMUNICATIONS, 8., 2010, Lulea. **Proceedings...** Lulea: Springer Berlin Heidelberg, 2010. p. 216–227.

LAGHARI, K. U. R.; CONNELLY, K. Toward total quality of experience: a QoE model in a communication ecosystem. **IEEE Communications Magazine**, New York, v. 50, n. 4, p. 58–65, Apr. 2012.

LI, Q. et al. Non-intrusive quality assessment for enhanced speech signals based on spectro-temporal features. In: IEEE INTERNATIONAL CONFERENCE ON MULTIMEDIA AND EXPO WORKSHOPS, 5., 2014, Chengdu. **Proceedings...** Chengdu: IEEE, 2014. p. 1–6.

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**, New York, v. 14, n. 6, p. 1924–1934, Oct. 2006.

MARZINZIK, M.; KOLLMEIER, B. Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. **IEEE Transactions on Speech and Audio Processing**, New York, v. 10, n. 2, p. 109–118, Aug. 2002.

MEIRELLES, F. S. **Pesquisa Anual do Uso de Tecnologia da Informação**. 27. ed. São Paulo: Fundação Getúlio Vargas, 2016. 24 p.

MELLOUK, A.; HOCEINI, S.; TRAN, H. A. Quality of experience vs. quality of service: Application for a CDN architecture. In: INTERNATIONAL CONFERENCE ON SOFTWARE, TELECOMMUNICATIONS AND COMPUTER NETWORKS, 21., 2013, Primosten. **Proceedings...** Primosten: IEEE, 2013. p. 1–8.

MONTEIRO, J. B. **Google Android**: crie aplicações para celulares e tablets. São Paulo: Casa do Código, 2012. 327 p.

MOSSAVAT, I.; PETKOV, P. N.; KLEJIN, W. B. A hierarchical bayesian approach to modeling heterogeneity in speech quality assessment. **IEEE Transactions on Audio, Speech, and Language Processing**, New York, v. 20, n. 1, p. 136–146, Jan. 2011.

NEMER, E.; GOUBRAN, R.; MAHMOUD, S. Robust voice activity detection using higher-order statistics in the LPC residual domain. **IEEE Transactions on Speech and Audio Processing**, New York, v. 9, n. 3, p. 217–231, Aug. 2001.

NUNES, R. D. et al. Real-time evaluation of speech quality in mobile communication services. In: IEEE INTERNATIONAL CONFERENCE ON CONSUMER ELECTRONICS, 2016, Las Vegas. **Proceedings...** Las Vegas: IEEE, 2016. p. 425–426.

PEREIRA, C. H. et al. Improving the performance of a non-intrusive metric of voice quality assessment considering IP network parameters. In: BRAZILIAN SYMPOSIUM OF TELECOMMUNICATION, 33., 2015, Juiz de Fora. **Anais...** Juiz de Fora: SBRT, 2015. p. 513–517.

POIKSELKA, M.; MAYER, G. **The IMS**: ip multimedia concepts and services. 3^{rd} ed. Chichester: J. Wiley, 2013. 533 p.

POLACKY, J.; POCTA, P. An analysis of the impact of packet loss, codecs and type of voice on internal parameters of P.563 model. In: INTERNATIONAL CONFERENCE ON DIGITAL TECHNOLOGIES, 10., 2014, Zilina. **Proceedings...** Zilina: IEEE, 2014. p. 281–284.

RAJA, A. et al. Real-time, non-intrusive evaluation of VoIP. In: EUROPEAN CONFERENCE, 10., 2007, Valencia. **Proceedings...** Berlin: Springer Berlin Heidelberg, 2007. p. 217–228.

RAMIREZ, J.; GÓRRIZ, J. M.; SEGURA, J. C. **Voice activity detection**: fundamentals and speech recognition system robustness. Vienna: INTECH Open Access, 2007. 460 p.

RODRIGUEZ, D. Z.; ROSA, R. L.; BRESSAN, G. A billing system model for voice call service in cellular networks based on voice quality. In: IEEE INTERNATIONAL SYMPOSIUM ON CONSUMER ELECTRONICS, 2013, Hsinchu. **Proceedings...** Hsinchu: IEEE, 2013. p. 89–90.

SADJADI, S. O.; HANSEN, J. H. L. Unsupervised speech activity detection using voicing measures and perceptual spectral flux. **IEEE Signal Processing Letters**, New York, v. 20, n. 3, p. 197–200, Jan. 2013.

TANENBAUM, A. S.; WETHERALL, D. J. **Computer Networks**. 5^{th} ed. Boston: Prentice Hall, 2010. 960 p.

TSIARAS, C. et al. Towards evaluating type of service related quality-of-experience on mobile networks. In: WIRELESS AND MOBILE NETWORKING CONFERENCE, 7., 2014, Vilamoura. **Proceedings...** Vilamoura: IEEE, 2014. p. 1–8.

WORLD WIDE WEB CONSORTIUM. **Web Services Activity**. Cambridge, 2015. Disponível em: <https://www.w3.org/2002/ws/Activity>. Acesso em: 24 dez. 2016.