UNIVERSIDADE ESTADUAL DO OESTE DO PARANÁ

CAMPUS DE FOZ DO IGUAÇU

PROGRAMA DE PÓS-GRADUAÇÃO EM
ENGENHARIA ELÉTRICA E COMPUTAÇÃO

DISSERTAÇÃO DE MESTRADO

DESENVOLVIMENTO DE UM PROTÓTIPO DE SISTEMA *WEB* PARA ELABORAÇÃO DE LAUDOS MÉDICOS UTILIZANDO SISTEMAS DE RECONHECIMENTO AUTOMÁTICO DE FALA

THIAGO FERREIRA DE TOLEDO

FOZ DO IGUAÇU
2017

Thiago Ferreira de Toledo

Desenvolvimento de um Protótipo de Sistema *Web* para Elaboração de Laudos Médicos Utilizando Sistemas de Reconhecimento Automático de Fala

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica e Computação como parte dos requisitos para obtenção do título de Mestre em Engenharia Elétrica e Computação. Área de concentração: Sistemas Dinâmicos e Energéticos.

Orientadora: Drª. Huei Diana Lee
Co-orientador: Dr. Wu Feng Chung
Co-orientador: Dr. Cláudio Saddy Rodrigues Coy

Foz do Iguaçu
2017

Catalogação na Publicação (CIP)
Sistema de Bibliotecas da UNIOESTE

T649 Toledo, Thiago Ferreira de
Desenvolvimento de um protótipo de sistema web para elaboração de laudos médicos utilizando sistemas de reconhecimento automático de fala / Thiago Ferreira de Toledo.-- Foz do Iguaçu, 2017.
137 p., il. : tabs. : gráfs.

Orientadora: Profª. Dr. Huei Diana Lee
Co-orientador: Prof. Dr. Wu Feng Chung
Co-orientador: Cláudio Saddy Rodrigues Coy
Dissertação (Mestrado) – Programa de Pós-Graduação em Engenharia Elétrica e Computação - Universidade Estadual do Oeste do Paraná.

1. Engenharia de software. 2. Informática na medicina. 3. Anamnese. 4. Word Wide Web (Sistema de recuperação da informação). 5. Software – Desenvolvimento. 6. UML (Computação). I. Título.

CDU 004.78:616-072
004.41

Miriam Fenner R. Lucas - CRB-9/268

Desenvolvimento de um Protótipo de Sistema *Web* para Elaboração de Laudos Médicos Utilizando Sistemas de Reconhecimento Automático de Fala

Thiago Ferreira de Toledo

Esta Dissertação de Mestrado foi apresentada ao Programa de Pós-Graduação em Engenharia Elétrica e Computação e aprovada pela Banca Examinadora:
Data da defesa pública: 18/08/2017.

Profª. Drª. **Huei Diana Lee** – (Orientadora)
Universidade Estadual do Oeste do Paraná – UNIOESTE

Prof. Dr. **Wu Feng Chung** – (Co-orientador)
Universidade Estadual do Oeste do Paraná – UNIOESTE

Prof. Dr. **Orlando Petrucci Junior**
Universidade Estadual de Campinas – UNICAMP

Prof. Dr. **Newton Spolaôr**
Universidade Estadual do Oeste do Paraná – UNIOESTE

Resumo

O propósito geral de sistemas de reconhecimento automático de fala é o de permitir a interação de seres humanos com dispositivos eletrônicos por meio da fala, por exemplo, a partir da fala do usuário, captada por um microfone, o seu conteúdo pode ser convertido em transcrição textual. Em geral, tais sistemas devem ser capazes de enfrentar adversidades como influência de ruídos, variabilidade do canal de comunicação, idade, sotaque e velocidade da fala, fala concorrente de outros oradores e fala espontânea. Mesmo diante desse cenário desafiador, este trabalho possui como propósito investigar essa tecnologia para ser empregada no âmbito médico. A investigação foi realizada por meio de uma revisão sistemática para o embasamento científico de conceitos e identificação de pesquisas relevantes. Para validar a viabilidade do uso desses sistemas na área médica, foi desenvolvido um Protótipo de Sistema *Web* com finalidade de gerar laudos médicos por meio do reconhecimento automático de fala para a Língua Portuguesa do Brasil. O protótipo foi desenvolvido com a técnica Entrega em Estágio, presente em Engenharia de *Software*. A seleção da tecnologia de reconhecimento automático de fala para ser integrada ao protótipo, foi baseada em uma avaliação da taxa de precisão de sete sistemas para a Língua Portuguesa do Brasil. O desenvolvimento desse trabalho permitiu concluir que esses sistemas de reconhecimento automático de fala podem ser empregados no ambiente médico, provendo suporte, não somente à confecção de laudos médicos, mas também atuando como um registro de monitoria durante um procedimento médico.

Palavras-chave: RAF, Laudo Médico, Transcrição Médica, Protótipo de Sistema *Web*, UML.

Abstract

The overall purpose of automatic speech recognition systems is to allow the interaction between humans and electronic devices through speech, for example, the content captured from user's speech using a microphone, can be converted into textual transcription. In general, such systems should be able to overcome adversities such as noise influence, communication channel variability, age, accent and speed of speech, concurrent speech from other speakers and spontaneous speech. Even with this challenging scenario, this work aims to investigate this technology to be employed in the medical field. The research was carried out through a systematic review for the scientific basis of concepts and identification of relevant research studies. To validate the feasibility of using these systems in the medical field, a Web System Prototype was developed to generate medical reports through automatic speech recognition for the Brazilian Portuguese Language. The prototype was developed with the application of Delivery in Stage technique, from Software Engineering. The selection of the automatic speech recognition technology to be integrated to the prototype was based on an evaluation of the accuracy rate of seven systems for the Brazilian Portuguese Language. The development of this work allowed to conclude that these systems of automatic speech recognition can be used in the medical environment, providing support not only for making medical reports, but also acting as a monitoring record during a medical procedure.

Keywords: ASR, Medical Report, Medical Transcription, Web System Prototype, UML.

Dedico este trabalho aos meus pais.

Agradecimentos

Agradeço especialmente aos meus pais, Carlos e Flávia, e aos meus irmãos, Vinícius e Fernanda, por todo o apoio, carinho, compreensão, paciência e amor. Sem vocês certamente o caminho que escolhi trilhar seria ainda mais difícil. Obrigado por tudo, por jamais medirem esforços em aconselhar-me, animar e encorajar a seguir em frente. A vocês, a minha eterna gratidão.

Aos meus amigos, Leonardo Pereira, Jonatan Liecheski, Pedro Coutinho e Rayan Chemin pela amizade sincera e sólida.

Aos condiscípulos da Fundação Logosófica de Foz do Iguaçu, Adriana, Audrey, Bruna, Elizabeth, Gustavo, João, Juliano, Jussara, Neri, Rafaela P., Rafaela Z., Simone, Tânia e Zélia, por me proporcionarem grandes momentos de evolução e valiosos ensinamento.

Aos meus dois grandes mestres e orientadores, professores Huei e Wu, pela paciência, no ensinamento do caminho da ciência e da docência.

Ao professor Newton por toda a sua dedicação, orientação e ajuda em diversos momentos. Aos grandes Labianos, Fonteque, Leandro, Paulo, Narco, Samia, Silvani e Weber.

Aos meus professores de mestrado, Carlos Farias, Carlos Rocha e Roberto Lotero, que me apoiaram e ensinaram ao longo dessa trajetória.

Também aos colegas de mestrado, Felipe Crestani e Itamar Nieradka, pelos momentos de estudo e de descontração.

À UNIOESTE e todos os seus funcionários. Em especial, agradeço à Fabiana Santos por todo o auxílio, profissionalismo e dedicação.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro.

Por fim, agradeço a Deus, a razão de tudo.

"Para triunfar é necessário vencer, para vencer é necessário lutar, para lutar é necessário estar preparado, para estar preparado é necessário prover-se de uma grande inteireza de ânimo e de uma paciência a toda a prova. Isto requer, por sua vez, levar constantemente ao íntimo da vida o incentivo da suprema esperança de alcançar aquilo que se anela como culminação feliz da existência."
Carlos B. G. Pecotche

Sumário

Lista de Figuras

Lista de Quadros

Lista de Siglas

ANN	*Artificial Neural Network*
API	*Application Programming Interface*
ARPA	*Advanced Research Projects Agency*
BD	Banco de Dados
CSS	*Cascading Style Sheets*
DAO	*Data Access Object*
dB	Decibel
DCT	*Discrete Cosine Transform*
DNN	*Deep Neural Network*
DP	Desvio Padrão
FFT	*Fast Fourier Transform*
GMM	*Gaussian Mixture Models*
HMM	*Hidden Markov Model*
HTK	*Hidden Markov Model Toolkit*
Hz	Hertz
JAR	Jar *ARchive*
JSF	*JavaServer Faces*
KS	Kolmogorov-Smirnov
LABI	Laboratório de Bioinformática
LSTM	*Long Short-Term Memory*
MA	Modelo Acústico
MFCC	*Mel-Frequency Cepstral Coefficients*
ML	Modelo de Linguagem
MLP	*Multi-layer Perceptron*
MVC	Modelo-Visão-Controlador – *Model View Control*

OOV	*Out-Of-Vocabulary*
PCM	*Pulse-code modulation*
PDF	*Portable Document Format*
PGEEC	Programa de Pós-graduação em Engenharia Elétrica e Computação
PSW	Protótipo de Sistema *Web*
RN	Regra de Negócio
RNN	*Recurrent Neural Network*
RS	Revisão Sistemática
SRAF	Sistema de Reconhecimento Automático de Fala
TIMIT	*Texas Instruments Massachusetts Institute of Technology*
UNICAMP	Universidade Estadual de Campinas
UNIOESTE	Universidade Estadual do Oeste do Paraná
WAR	*Web application ARchive*
WAV	*WAVEform audio format*
WER	*Word Error Rate*
WSJ	*Wall Street Journal*
xHTML	*eXtensible Hypertext Markup Language*
XML	*eXtensible Markup Language*

Lista de Tabelas

Capítulo 1

Introdução

Devido a importantes progressos ocorridos na área tecnológica, tornou-se possível a construção de sistemas cada vez mais complexos em virtude do grande aumento no poder de processamento e de armazenamento computacional. Uma das áreas a se beneficiar dessa evolução é a tecnologia de Sistema de Reconhecimento Automático de Fala (SRAF) (Yu & Deng, 2015). Um SRAF possui como finalidade o reconhecimento de palavras faladas e convertê-las no formato de texto (Donaj & Kačič, 2017).

Além dos SRAFs que permitem a interação de seres humanos com sistemas computacionais de maneira mais natural por meio da fala, também há os sistemas de diálogo e de síntese de fala. Os sistemas de diálogo são aplicações computacionais que conversam com um ser humano, enquanto os de síntese da fala permitem que dispositivos eletrônicos transformem um texto em fala (Donaj & Kačič, 2017).

O uso de SRAF se tornou um dos principais meios de interação de humanos com alguns equipamentos eletrônicos, como dispositivos móveis, dispositivos vestíveis – *wearables*[1] –, e dispositivos *infotainment*[2] (Yu & Deng, 2015).

A utilização da tecnologia de SRAF pode ser integrada a aplicações para controlar ambientes residenciais (Santos-Perez et al., 2013), reconhecer sentimentos da fala (Kaushik et al. 2013; Kaushik et al., 2017), minerar informações coletadas em conversas de uma central de atendimento (Clavel et al., 2013) e servir de guia de um museu ao responder as perguntas dos visitantes (Misu et al., 2012).

Esses sistemas também são utilizados para permitir a produção de documentos digitais a partir da fala (Revuelta-Martínez et al., 2012; Batista, 2013), realizar pesquisa de informações em arquivos multimídias (Law-To & Grefenstette, 2011; Varona et al., 2011; Adell et al., 2012) e para auxiliar no aprendizado, por exemplo, detectando erros de leitura de uma criança (Yilmaz et al. 2014) ou apoiar no ensino de uma segunda língua estrangeira (Ahn & Lee, 2016; Cavus & Ibrahim, 2016).

Outra utilidade desses sistemas é a de possibilitar a tradução entre diversas línguas por meio de dispositivos móveis, auxiliando turistas em países estrangeiros (Sakti et al., 2013; Ili, 2017; Matsuda et al., 2017; Waverly, 2017) e também a criar salas virtuais para interação entre pessoas de diferentes nacionalidades (Gopi et al., 2015; Microsoft, 2017a).

No contexto médico, os SRAFs podem ser utilizados para apoiar crianças com síndrome de *Down* a melhorar suas habilidades de leitura (Felix et al., 2017), prever se uma pessoa possui doença de Parkinson (Vasquez-Correa et al., 2016), apoiar, por meio conversacional,

[1] Tecnologias para vestir, como relógios inteligentes.
[2] Sistemas utilizados em veículos para oferecer informação e entretenimento.

pacientes com transtorno de estresse pós-traumático (Papangelis et al., 2013) ou facilitar a comunicação de uma pessoa que sofre de disfunção da fala (Balaji & Sadashivappa, 2015), encaminhar um paciente a uma determinada especialidade médica de acordo com as preocupações relatadas por ele verbalmente (Leuski et al., 2014) e avaliar a inteligibilidade da fala para pacientes com doença bocal (Riemann et al., 2016).

Já em hospitais, os SRAFs são utilizados em consultórios médicos durante uma consulta para aperfeiçoar a coleta de dados de informações durante o atendimento (Gür, 2012) ou como um sistema de tradução para facilitar o atendimento a imigrantes (Soller et al., 2012) e também na confecção de relatórios médicos (Prevedello et al., 2014).

Atualmente, os SRAFs mais modernos são construídos baseados na tecnologia de Redes Neurais Artificiais Profundas, como os SRAFs da Google (Sak et al., 2014), da IBM (Ganapathy et al., 2015) e da Microsoft (Hakkani-Tür et al., 2016). Essa tecnologia é baseada em uma arquitetura de aprendizagem de máquina composta de múltiplas camadas de processamento de dados (Li et al., 2016).

Apesar dos avanços na precisão desses sistemas, ainda existem limitações, como o ruído, que corresponde a distúrbios indesejados sobrepostos ao sinal da fala pretendido. As palavras fora do vocabulário de reconhecimento são as palavras que não estão presentes no vocabulário de treinamento do SRAF. Além disso, esses sistemas devem ser robustos para possibilitar o seu bom funcionamento em condições variadas (Li et al., 2016; Donaj & Kačič, 2017) e ser capazes de tratar as variações da voz do falante, da pronúncia e do ambiente (Huang et al., 2001).

Dessa maneira, é importante que os SRAFs reconheçam com a maior precisão possível, todas as palavras pronunciadas. No entanto, essa tarefa não é facilmente alcançada devido a erros ocorridos durante esse processo, como palavras não reconhecidas, palavras não pronunciadas que são inseridas no texto ou palavras substituídas (Donaj & Kačič, 2017).

Baseado nesse cenário, o Laboratório de Bioinformática (LABI) da Universidade Estadual do Oeste do Paraná (UNIOESTE/Foz do Iguaçu) em parceria com o Departamento de Coloproctologia da Faculdade de Medicina da Universidade Estadual de Campinas (UNICAMP), tem investigado o panorama geral da tecnologia de reconhecimento automático de fala para utilizá-la em sistemas no âmbito médico. Para validar o uso dessa tecnologia, foi desenvolvido um Protótipo de Sistema *Web* para gerar laudos médicos por meio de SRAFs. Nesse contexto, um laudo médico é um parecer escrito preenchido por um perito, no qual constam os resultados de um exame pericial. A sua estrutura padrão é composta por preâmbulo, perguntas a serem respondidas, histórico de doenças, descrição, discussão, conclusão e respostas a perguntas. Também devem incluir uma descrição de todos os sinais e sintomas, resultados dos testes realizados, tratamento adotado, evolução apresentada e esperada para o paciente (CRM-PR, 2008).

1.1 Objetivos

Como objetivo geral, este trabalho tem o propósito de investigar o panorama geral da tecnologia de reconhecimento automático de fala e o seu emprego no âmbito médico. Para alcançar esse objetivo foram definidos os seguintes objetivos específicos:

- Realização de uma revisão sistemática para embasamento científico de conceitos e identificar pesquisas relevantes na área;
- Avaliação do desempenho de SRAFs para a Língua Portuguesa do Brasil;
- Construção de um Protótipo de Sistema *Web* para geração de laudos médicos por meio de sistemas computacionais de reconhecimento automático de fala.

1.2 Assertiva

A utilização da tecnologia de reconhecimento automático de fala possibilita uma interação de maneira mais natural com dispositivos eletrônicos por meio da fala.

1.3 Hipótese

A tecnologia computacional de reconhecimento automático de fala pode ser utilizada como ferramenta de apoio para a redução de tempo para a confecção de laudos ou como um registro de monitoramento de acompanhamento durante um procedimento médico.

1.4 Organização do Trabalho

Este trabalho está organizado da seguinte maneira:

- **Capítulo 2:** são fundamentadas as características dos sons, bem como os conceitos do processo da fala e da sua compreensão;
- **Capítulo 3:** são apresentadas limitações acerca do funcionamento de SRAFs, uma breve evolução histórica desses sistemas e alguns trabalhos relacionados, bem como: a sua arquitetura geral, algumas fontes que resultam na variação acústica e métricas de avaliação de precisão;
- **Capítulo 4:** as principais tecnologias utilizadas em SRAFs são apresentadas, como Modelo Oculto de Markov, Perceptron Multicamadas e Rede Neural Recorrente com células de Memória Longa de Curto Prazo;

- **Capítulo 5:** são apresentados os materiais e os métodos utilizados para a realização deste trabalho, onde são detalhados o protocolo da revisão sistemática, os SRAFs avaliados, as tecnologias, as ferramentas e os equipamentos utilizados para o desenvolvimento do Protótipo de Sistema *Web*;
- **Capítulo 6:** para a construção do Protótipo de Sistema *Web* foi utilizado a técnica Entrega em Estágio, presente em Engenharia de *Software*, cujo método é apresentado nesse capítulo;
- **Capítulo 7:** são relatados os resultados e sua discussão referentes à revisão sistemática, à avaliação dos SRAFs e à apresentação do Protótipo de Sistema *Web*;
- **Capítulo 8:** são apresentadas as conclusões obtidas neste trabalho, bem como: suas contribuições, limitações e sugestões de trabalhos futuros.

Capítulo 2

A Fala Humana e a sua Compreensão

Para que os seres humanos possam se organizar, transmitir conhecimento e cultura para as próximas gerações é fundamental que haja comunicação adequada. Sendo assim, a modalidade mais importante é a comunicação oral, que requer a presença de dois mecanismos essenciais caracterizados pela produção da fala e pelo processo de sua compreensão.

Dessa maneira, neste capítulo, são abordadas as características dos sons (Seção 2.1), os sistemas e os principais órgãos envolvidos no processo da fala humana (Seção 2.2). Na Seção 2.3, é apresentada a sua compreensão, incluindo o sistema auditivo e os principais órgãos para essa finalidade. Os órgãos do sistema auditivo estão dispostos ao longo das quatro partes da orelha, sendo divididas em: externa, média, interna e via neural.

2.1 Características dos Sons

O som é a propagação de uma frente de compressão mecânica ou onda longitudinal que se propaga, de modo cíclico, tridimensionalmente pelo espaço em meios materiais (Bauer et al., 2013). Dessa maneira, é possível a mensuração do tempo de um ciclo desse fenômeno denominado período que está diretamente relacionado com a frequência. Tanto o período quanto a frequência podem ser representados por unidades como segundos e ciclos por segundo (Hertz), respectivamente. A percepção da frequência por um ouvinte é subjetiva, em que uma alta frequência é interpretada como um som agudo, e uma baixa frequência é interpretada como um som grave (Cutnell & Johnson, 2016).

Além desses atributos físicos, existem outras medidas no fenômeno de produção de sons, como a amplitude, que é uma medida de distância entre o ponto mais alto do padrão da onda com a posição não perturbada (Cutnell & Johnson, 2016). Por exemplo, na Figura 2.1 é representado o sinal de uma onda com o seu comprimento e amplitude. O comprimento da onda é determinado pela distância de um ponto em um ciclo da onda até o ponto idêntico no próximo ciclo. O eixo X representa o deslocamento no meio de propagação e o eixo Y representa a posição ao longo da onda.

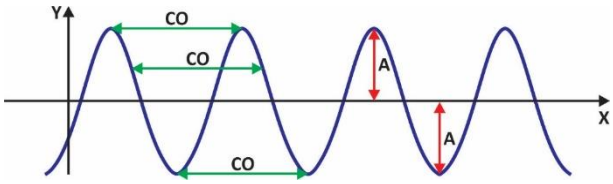

Figura 2.1: Representação da onda sonora (CO = comprimento da onda e A = amplitude).
Fonte: Adaptado de Kesten & Tauck (2015) e Cutnell & Johnson (2016).

Outras características essenciais no fenômeno físico do som é a intensidade, a potência e o ruído. A intensidade sonora (Equação 2.1) tem relação com a potência sonora *P* que atravessa de maneira perpendicular uma superfície dividida pela área *A* dessa superfície, resultando na intensidade *I* (Cutnell & Johnson, 2016).

$$I = \frac{P}{A}. \tag{2.1}$$

A intensidade é a qualidade para caracterizar se um som é forte ou fraco, sendo dependente da energia que a onda sonora transfere. Se a intensidade for expressa como o seu produto com unidade de área da superfície que recebe a energia vibratória, esta é definida como potência do som. Ou seja, potência é a quantidade de energia que uma onda sonora transfere por segundo (Bauer et al., 2013; Kesten & Tauck, 2015; Cutnell & Johnson, 2016).

Já o ruído é a ausência de periodicidade das ondas sonoras, isto é, ocorre uma mistura de diversos sons com diferentes frequências. O ruído também pode ser caracterizado como um som indesejado. Por exemplo, um som de baixa frequência produzido por um motor de avião a jato, sendo este som não desejado por um ouvinte (Bauer et al., 2013; Kesten & Tauck, 2015).

A unidade utilizada para o nível sonoro é o Bel (B), em que 1 B corresponde a 10 decibel (dB) (Bauer et al., 2013). A intensidade de pressão sonora que o ouvido humano pode suportar compreende a faixa entre 0 a 120 dB. Caso essa intensidade seja próxima ou superior a 120 dB, o nociceptor, que é um receptor sensorial, envia um sinal de dor em resposta a um estímulo que pode causar algum dano à audição. Para se ter uma noção dos diferentes níveis de dB, uma pessoa murmurando representa 30 dB, uma conversa habitual é de 60 dB, um trânsito pesado de rua equivale a 70 dB, um concerto de *rock* gera 120 dB e uma decolagem de um avião a jato equivale a 150 dB (Roberto-Douglas, 2009).

2.2 A Fala Humana

Para que seja possível a utilização de uma linguagem natural é fundamental possuir alguma língua natural específica, como a Língua Portuguesa ou a Língua Inglesa, que é a representação em palavras para haver uma comunicação por fala ou por escrita. Já a linguagem é a comunicação por algum meio. Por exemplo, a linguagem humana se

desenvolveu a partir de um sistema de comunicação gestual e, no atual estado de evolução, a linguagem predominante é por meio dos sons audíveis proveniente dos órgãos da fala (Lyons, 1987).

O processo da fala envolve, além do sistema respiratório, o envolvimento de centros específicos de controle da fala no córtex cerebral, de centros de controle respiratórios no cérebro e estrutura de articulações e ressonância da boca, bem como de cavidades nasais (Hall & Guyton, 2011). Na Tabela 2.1 são apresentados os sistemas corporais relacionados ao processo da fala.

Tabela 2.1: Sistemas corporais que possuem relação com o processo da fala humana. Fonte: (Fuller et al., 2014).

Sistema	Principais órgãos e tecidos	Função na fala
Circulatório	Coração, vasos sanguíneos e sangue	Irrigação sanguínea para o cérebro e outras partes do mecanismo de fala
Digestório	Lábios, dentes, língua, véu palatino, faringe, esôfago, estômago e intestinos	Mediação da articulação e da ressonância
Esquelético	Cartilagens, ossos, ligamentos, tendões e articulações	Estrutura para o mecanismo do processo da fala
Muscular	Músculos esqueléticos	Movimentação para realizar o processo da fala
Nervoso	Cérebro, medula espinal, nervos periféricos, gânglios e receptores sensoriais	Condução de estímulos nervosos associados ao processo da fala
Respiratório	Cavidades nasais, faringe, laringe, traqueia, brônquios e ramificações pulmonares	Fonte de energia para o processo da fala

De maneira geral, o processo da fala é composto por fonação, articulação e ressonância (Hall & Guyton, 2011):

- **Fonação:** a laringe é adaptada para funcionar como vibrador, tendo como elementos vibradores as cordas vocais. No momento da respiração as cordas vocais estão abertas para permitir a passagem de ar. Durante a fonação, as cordas vocais se movem juntas para permitir que a passagem de ar entre elas resulte em vibração. A ação dos músculos do interior das cordas vocais pode mudar o formato e a massa das bordas das cordas vocais para realizar a emissão de tons agudos ou graves.
- **Articulação e Ressonância:** os principais órgãos da articulação são lábios, língua e palato mole que se movimentam durante a fala e em outras vocalizações. Já os ressonadores são compostos pela boca, nariz, seios paranasais, faringe e cavidade torácica. A função dos ressonadores é a de permitir a mudança qualitativa da voz (timbre).

A fala humana pode ser caracterizada de acordo com algumas variantes para compor o aspecto estilístico da fala, por exemplo, de gênero, sendo masculino ou feminino; de faixa

etária e formal ou informal. Por exemplo, a fala formal é adequada para eventos formais, como em entrevista de emprego. Já a fala informal é utilizada em ocasiões onde existe o convívio íntimo entre pessoas. Há também as variantes geográficas, por exemplo, as variações entre a forma de falar em diferentes regiões do Brasil, por exemplo, o carioca, o gaúcho, o mineiro, o nordestino e o paulista, que possuem certas peculiaridades na fala. Outro fator que também influencia na variação da fala de um indivíduo está relacionado ao seu grau de instrução (Cristófaro-Silva, 2003).

2.3 Compreensão da Fala Humana

A anatomia do sistema auditivo humano utilizado para compreensão da fala pode ser dividida em quatro partes: orelha externa, orelha média, orelha interna e via neural (Fuller et al., 2014). Na Figura 2.2 é ilustrado as quatro partes da orelha.

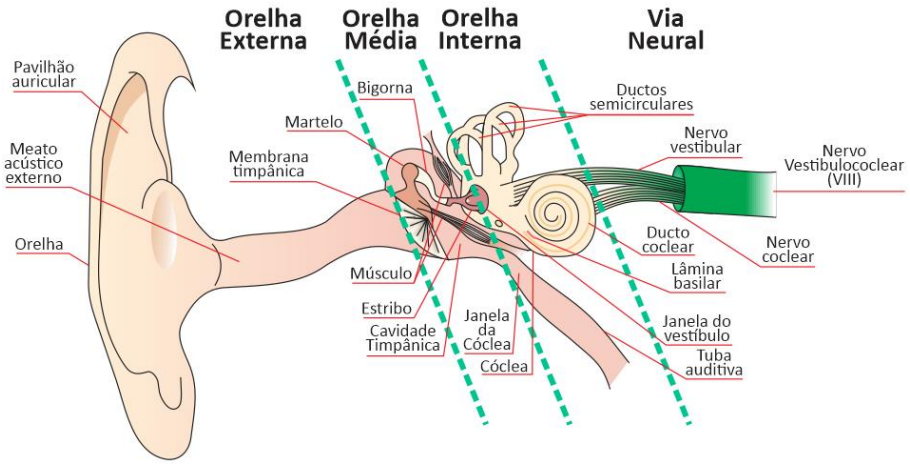

Figura 2.2: As quatro partes da orelha.
Fonte: Adaptado de Graaff, (2003); Fuller et al. (2014); Ward & Linden (2014).

Esquematicamente, a compreensão da fala é realizada da seguinte maneira (Fuller et al., 2014):

- **Orelha externa:** capta a onda sonora e a encaminha para o interior do meato acústico externo, que se estende até a membrana timpânica;
- **Orelha média:** a membrana timpânica separa a orelha média do meato acústico externo da orelha externa. Quando a membrana timpânica vibra, ocorre a movimentação dos ossículos da audição, composto pelo martelo, bigorna e estribo, que movimenta e transmite ondas sonoras por meio da cavidade timpânica em direção à janela do vestíbulo. Quando as ondas sonoras são transmitidas da membrana timpânica para a janela do vestíbulo ocorre a amplificação do som em aproximadamente 20 vezes (Graaff, 2003). A tuba auditiva equaliza a pressão de

ar no interior da orelha média e a pressão do ar atmosférico que preenche o canal auditivo;

- **Orelha interna:** a cóclea contém um órgão chamado ducto coclear e em seu interior há o órgão de Corti. Os receptores sonoros que transformam a vibração mecânica em impulsos nervosos estão localizados ao longo da lâmina basilar presente nesta estrutura. Com isso, é formada a unidade funcional da audição (Graaff, 2003);
- **Via neural:** o órgão de Corti que contém líquido, que de acordo com o balanço do estribo ocorre a vibração desse líquido criando impulsos nervosos sensoriais para que o nervo vestibulococlear – dividido nos nervos vestibular e coclear – possa conduzir a energia elétrica ao encéfalo. Com isso, esses impulsos são enviados para a via auditiva, composta pelo tronco encefálico e pelo córtex cerebral, no qual o som é percebido e interpretado.

Com base nisso, na Figura 2.3 é ilustrado o diagrama com os passos necessários para que possa haver a compreensão da fala humana.

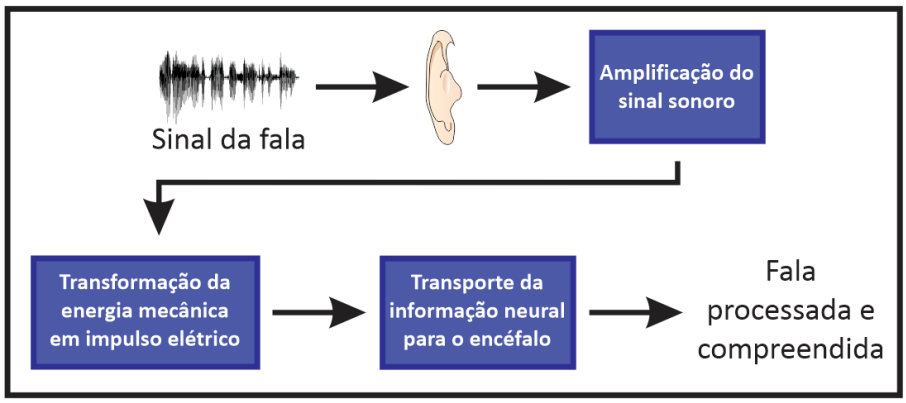

Figura 2.3: Diagrama com os passos para a realização da audição humana.

2.4 Considerações Finais

Neste capítulo foram abordadas as principais características dos sons, que incluem a frequência, a amplitude, a intensidade, a potência e o ruído. De acordo com a vibração das cordas vocais – fonação –, do movimento dos órgãos de articulação, como os lábios, a língua e o palato mole em conjunto com os órgãos da boca, nariz, seios paranasais, faringe e cavidade torácica – ressonadores –, é possível criar a fala ou outras vocalizações e mudar os aspectos qualitativos da voz.

Quando um ser humano emite uma fala, o aparelho auditivo capta esse som e o transforma em impulsos elétricos para que o encéfalo possa compreendê-lo. Dessa maneira, tendo em vista a maneira de como os seres humanos produzem e compreendem os sons, no

próximo capítulo será apresentado uma visão geral sobre sistemas de reconhecimento automático de fala e como essa tecnologia tenta simular o funcionamento da audição humana.

Capítulo 3

Sistemas de Reconhecimento Automático de Fala

Um Sistema de Reconhecimento Automático de Fala (SRAF) é uma tecnologia que permite aos seres humanos interagirem com dispositivos eletrônicos, por meio da sua fala, utilizando técnicas computacionais.

Dentre algumas das técnicas envolvidas nesse processo, tem-se a extração de características que capta a fala, por meio do microfone, e a converte em uma representação de dados que possa ser processada por computadores. O Modelo Acústico conta com áudios de diversas palavras pronunciadas por diferentes pessoas. O Decodificador compara o som da fala capturado pelo microfone com os áudios contidos no Modelo Acústico. O Modelo de Linguagem possui um conjunto de textos que será utilizado, também pelo Decodificador, para entregar as palavras pronunciadas pelo usuário.

Com isso, na Seção 3.1 são abordadas algumas limitações enfrentadas por esses sistemas. Uma evolução histórica desses sistemas (Seção 3.2) e na Seção 3.3 é relatado como tais sistemas estão sendo utilizados atualmente no contexto médico. Na sequência é apresentada a arquitetura geral e os principais componentes de um SRAF (Seção 3.4). Na Seção 3.5 são abordadas algumas fontes de variabilidade acústica que podem distorcer o sinal acústico. Por fim, algumas métricas para medir a precisão de um SRAF (Seção 3.6).

3.1 Limitações de Sistemas de Reconhecimento Automático de Fala

A utilização de tecnologias computacionais para o reconhecimento automático da fala, fez com que surgisse uma maneira mais natural de interação entre os seres humanos com os dispositivos eletrônicos, pois tornou-se possível com que as máquinas reconheçam o que está sendo falado (Donaj & Kačič, 2017).

Um SRAF tem como objetivo principal o de reconhecer corretamente as palavras faladas e devolvê-las em forma de texto. No entanto, essa tarefa não é facilmente alcançada, devido a erros que podem ocorrer durante esse processo, como palavras não reconhecidas, palavras não pronunciadas que podem ser inseridas no texto ou quando alguma palavra proferida é substituída por outra (Donaj & Kačič, 2017). Além disso, os SRAFs devem ser

capazes de lidar com as variações do falante, da pronúncia e das diferentes condições do ambiente (Huang et al., 2001).

Embora tenham ocorridos avanços na precisão dos SRAFs, há alguns problemas a serem superados no que se refere ao ruído, à robustez e às palavras fora do vocabulário – *Out-Of-Vocabulary* (OOV). O ruído é caracterizado por distúrbios indesejados sobrepostos ao sinal da fala pretendida. A robustez está relacionada com a capacidade de o sistema operar em condições variadas, incluindo questões imprevisíveis ou indisponíveis no momento do desenvolvimento do SRAF (Li et al., 2016). As palavras fora do vocabulário se referem às palavras não contempladas no treinamento de um SRAF. Assim, quanto maior for o vocabulário de treinamento, menor será a taxa de OOV (Donaj & Kačič, 2017).

3.2 Breve Evolução Histórica dos Sistemas de Reconhecimento Automático de Fala

Em 1961, pesquisadores do laboratório *Radio Research Lab* em Tóquio construíram um dispositivo de *hardware*, constituído por um circuito lógico, capaz de reconhecer vogais pronunciadas separadamente (Rabiner & Juang, 1993).

Após alguns anos de pesquisa na área de reconhecimento de fala, em 1975, James K. Baker da Universidade Carnegie-Mellon desenvolveu, de forma independente, o primeiro sistema de reconhecimento de fala contínua. Esse sistema, denominado DRAGON, utilizou um modelo probabilístico baseado em cadeias de Markov (Seção 4.2) para realizar o reconhecimento de fala (Baker, 1975). Uma cadeia de Markov é um processo de estado indeterminado, cujo estado atual dependerá apenas do estado anterior (Jurafsky & Martin, 2016).

Na década de 1970, o *Advanced Research Projects Agency* (ARPA) financiou um projeto para desenvolver sistemas de reconhecimento de fala capazes de reconhecer 1.000 palavras, obtendo taxas de erros inferiores a 10%. Dentre os sistemas financiados, apenas o Harpy, construído por Bruce Lowerre, durante o seu doutorado na Universidade Carnegie-Mellon (Lowerre, 1976), cumpriu o objetivo estipulado pelo projeto (Klatt, 1977).

Outros sistemas financiados pelo projeto ARPA foram o SDC (*System Development Corporation*[3]), o BBN HWIM (*Hear what I Mean*[4]) e o CMU Hearsay-2. O SDC gera uma transcrição fonética que inclui vários rótulos alternativos para cada segmento fonético (Klatt, 1977).

O HWIM é um sistema de compreensão de fala projetado para compreender expressões orais faladas de maneira natural para o domínio de tarefas relacionadas ao gerenciamento de orçamento de viagens (Wolf & Woods, 1977). Por fim, o CMU Hearsay-2 foi desenvolvido pela Universidade Carnegie-Mellon, tendo como objetivo compreender a fala contínua para

[3] Traduzindo para a Língua Portuguesa do Brasil: Corporação de Desenvolvimento de Sistemas.
[4] Traduzindo para a Língua Portuguesa do Brasil: Ouça o que eu quero dizer.

recuperar notícias diárias de um serviço de notícias mediante pedido por voz do usuário (Lesser et al., 1975).

Shozo Makino, Takeshi Kawabata e Ken'iti Kido desenvolveram o primeiro sistema de reconhecimento de fala utilizando Redes Neurais Artificiais (Seção 4.3), com o objetivo de reconhecer consoantes (Makino et al., 1983). As Redes Neurais Artificiais são técnicas computacionais baseadas em células nervosas de organismos inteligentes que possui capacidade de aprender para se obter algum resultado (Braga et al., 2000).

Notou-se que, para ser possível realizar a compreensão da fala contínua, seria necessária a construção de grandes bases de dados de áudio para maximizar a probabilidade de sistema conter melhores padrões. Então, em 1984, o ARPA financiou um segundo projeto, obtendo como resultado o *corpus*[5] TIMIT em 1986, que viria a ser o primeiro *corpus* padrão a ser amplamente utilizado para a Língua Inglesa (Gold, 2011). O TIMIT foi desenvolvido em um esforço conjunto entre *Massachusetts Institute of Technology* (MIT), SRI *International* e *Texas Instruments* (TI). A fala para a construção do *corpus* foi coletada no TI e as transcrições, ou seja, a conversão da fala do áudio para o formato de texto foi realizada no MIT, dando origem ao nome TIMIT (LDC, 2017a).

Em 1988, o *Defense Advanced Research Projects Agency*, em seu programa de reconhecimento de fala, desenvolveu outro *corpus*, chamado de *Resource Management*. Esse corpus foi construído com mais de 21.000 enunciados gravados a partir de 160 locutores com variedades de dialetos[6] da Língua Inglesa (Price et al., 1988). Posteriormente, o projeto ampliou o tamanho do *corpus* para realizar as avaliações do nível de precisão dos sistemas. O *corpus* teve como objetivo o reconhecimento de fala contínuo com um grande vocabulário para a Língua Inglesa (Pallett et al., 1994).

A Universidade Carnegie-Mellon desenvolveu, em 1990, o CMU *Sphinx* (Lee et al., 1990). Esse sistema é baseado em Modelo Oculto de Markov – *Hidden Markov Model* (HMM) – (Seção 4.2), que é uma função probabilística de uma cadeia de Markov para gerar uma sequência de saídas (Yu & Deng, 2015; Davis & Scharenborg, 2017). Atualmente, o CMU-*Sphinx* está em fase de desenvolvimento da sua quinta versão[7].

Em 1993, a Universidade de Cambridge desenvolveu o *Hidden Markov Model Toolkit* (HTK), que é um conjunto de ferramentas para a construção de sistemas utilizando modelos de Markov (HTK, 2016).

Pesquisas nessa área também foram realizadas pelo Laboratório de Processamento de Sinais da Universidade Federal do Pará, onde foi desenvolvido, em 2005, um SRAF para a Língua Portuguesa do Brasil. No entanto, os resultados não foram satisfatórios, alcançando Taxa de Erro de Palavra – *Word Error Rate* (WER) – (Seção 3.6), de 58,12%. Esse sistema foi construído utilizando Modelo Oculto de Markov como modelagem acústica e para o processo de extração de características foram utilizados os Coeficientes Cepstral de Frequência Mel – *Mel-Frequency Cepstral Coefficients* (MFCC) –, (Silva et al., 2005). Dando

[5] Conjunto de enunciados de uma determinada língua. O plural de *corpus* é *corpora*.
[6] Dialeto é a variação linguística para falantes de uma mesma língua, por exemplo, a Língua Portuguesa do Brasil possui o dialeto carioca, gaúcho, mineiro, nordestino e paulista.
[7] http://cmusphinx.sourceforge.net/

continuidade ao trabalho, em 2010, um novo sistema foi desenvolvido, denominado Coruja, alcançando WER de 32,87% (Silva, 2010).

Em 2010, os pesquisadores da Microsoft iniciaram pesquisas, em parceria com a Universidade de Toronto no Canadá, para investigar o uso da tecnologia híbrida, baseada em uma Rede Neural Profunda – *Deep Neural Network* (DNN) – (Seção 4.1) e em HMM para serem utilizadas na tarefa de transcrição da fala para texto em SRAFs de grande vocabulário (Yu et al., 2010; Dahl et al., 2012).

Os pesquisadores da Google, em parceria com a Universidade de Toronto, também avaliaram em 2012 a utilização da tecnologia híbrida DNN-HMM para o reconhecimento de fala contínua em dispositivo móvel baseado em nuvem (Jaitly et al., 2012). Nessa abordagem, o sinal da fala é modelado com HMM e as probabilidades de observações são estimadas por meio de DNNs (Yu & Deng, 2015).

Pesquisas com DNN também foram investigadas pela IBM, onde foi conduzido um experimento baseado em DNN utilizando vetores de identidades de falantes (*i*-vetores). O *i*-vetores possui um vetor dimensional fixo, utilizado para resumir as estatísticas obtidas de uma gravação. Nessa investigação, a união de DNN com *i*-vetores resultou em melhorias na precisão de reconhecimento. Essa abordagem tenta aprender os pesos do DNN de maneira a reduzir a variabilidade do falante na classificação de fonemas explorando as características do falante que melhor se encaixam no *i*-vetores (Ganapathy et al., 2015).

Ainda no contexto de Rede Neural Profunda, em 2010 foi projetada uma nova estrutura de reconhecimento de fala contínua com a tecnologia Memória Longa de Curto Prazo – *Long Short-Term Memory* (LSTM) – que possibilitou que uma Rede Neural Recorrente – *Recurrent Neural Network* (RNN) – (Seção 4.5) possa armazenar e recuperar informações durante longos períodos de tempo. Uma RNN possui estados internos que retêm um histórico de entradas anteriores, podendo processar sinais da fala que se desenvolvem ao longo do tempo (Wöllmer et al., 2010). O método LSTM foi proposto inicialmente em 1997 por Hochreiter & Schmidhuber (1997). Em 2014, os pesquisadores da Google utilizaram RNN em conjunto com LSTM em seu SRAF (Sak et al., 2014).

Em 2016, a Microsoft também utilizou a tecnologia RNN-LSTM em seu sistema de conversação para compreensão da fala. Esta abordagem foi testada no Cortana, um assistente virtual para a interação do usuário com o sistema operacional (Hakkani-Tür et al., 2016).

Na dissertação de Quintanilha (2017), foi proposto um SRAF para a Língua Portuguesa do Brasil, baseado em caracteres, utilizando a arquitetura RNN-LSTM. A extração de característica foi obtida por meio do MFCC. Nesse trabalho, também foi construído um *corpus* que consiste em aproximadamente 13 horas de dados de treinamento.

3.3 Trabalhos Relacionados

No contexto médico, os SRAFs podem ser utilizados no auxílio a pacientes com algum tipo de doença, como na interação de pacientes com transtorno de estresse pós-traumático, isto é, experimentar algum tipo de transtorno de ansiedade devido a algum evento traumático

(Papangelis et al., 2013). Para essa tarefa, Papangelis et al. (2013) desenvolveu um sistema capaz de interagir através de diálogo natural, em Língua Inglesa, com esses pacientes. O sistema é capaz de se adaptar a cada paciente individualmente, podendo operar em dois modos: (1) modo que armazena informações sobre sessões anteriores para proporcionar um sentimento de confiança e relacionamento com o paciente, e (2) modo que não armazena informações, a fim de preservar o anonimato. Segundo os autores, a novidade da abordagem é que o sistema acompanha o estado emocional do paciente ao armazenar informações de sessões anteriores. Outro objetivo do sistema é o de obter informações suficientes para fazer uma avaliação da condição do paciente de forma similar a um teste de autoavaliação.

No caso de pacientes que sofrem com a doença de Parkinson são desenvolvidas várias deficiências relacionadas ao processo de produção de fala, como a redução na capacidade de fonação, articulação prosódica e inteligibilidade (Vasquez-Correa et al., 2016). Diante desse cenário, foram propostas duas abordagens para analisar os déficits de inteligibilidade dos pacientes com doença de Parkinson. A primeira abordagem consiste em utilizar um SRAF para avaliar a quantidade de palavras pronunciadas corretamente. E a segunda abordagem corresponde à atribuição de uma pontuação de similaridade calculada entre as frases lidas pelos pacientes e a sequência de palavras reconhecidas pelo SRAF. Isso é realizado a partir da classificação automática da leitura das frases pelos pacientes versus a leitura de pessoas saudáveis que não possuem a doença de Parkinson. O sistema também realiza a predição automática do estado neurológico do paciente, que é realizada de acordo com uma escala de classificação de doença de Parkinson, atribuída por especialistas de neurologia. Dessa maneira, de acordo com os resultados, é possível diferenciar os pacientes com doença de Parkinson de pessoas saudáveis (Vasquez-Correa et al., 2016).

Para pacientes com desvio fonético, ou seja, distúrbio da fala que inclui problemas causados por distúrbios articulatórios e derivados das características acústicas (Chalegre-Paula & Neto, 2016), foi proposto um protótipo para auxiliar no tratamento de crianças. O tratamento consiste no treinamento da fala apoiado por um SRAF, em que o paciente pronúncia frases contendo palavras com fonemas previamente definidos. O módulo do SRAF do protótipo foi construído com base em falantes da Língua Inglesa. A validação das funcionalidades do protótipo foi realizada por três profissionais de fonoaudiologia da Universidade Federal de Pernambuco (Chalegre-Paula & Neto, 2016).

Com relação à utilização de SRAFs em ambiente hospitalar, no trabalho de Liu et al. (2011) foram avaliados dois SRAFs, o Nuance Dragon (nas versões Gen e Med) e o SRI Decipher. Segundo os autores, para a avaliação do Nuance, houve o treinamento do sistema com as respectivas vozes de cada participante. Com relação ao SRI foi realizada a adaptação do Modelo de Linguagem (Subseção 3.4.2) com mais de 4.000 questões clínicas.

Segundo os autores, o experimento dos SRAFs consistiu na gravação de leitura de 20 perguntas, em inglês, por nove estudantes do segundo ano de medicina da Universidade de Wisconsin-Milwaukee. Os participantes gravaram cada questão duas vezes, sendo a primeira gravação com a leitura textual e a segunda gravação com a fala espontânea. O SRI apresentou WER de 41,50% e após a adaptação do ML, o WER foi de 26,70% e para o Nuance Gen e Med, os WERs foram de 68,10% e 67,40%, respectivamente. Como resultado, os autores

relataram que os sistemas avaliados não funcionaram bem em questões clínicas quando não houve a adaptação dos SRAFs (Liu et al., 2011).

Outro SRAF utilizado no contexto hospitalar é o S-MINDS, desenvolvido pela empresa Fluential em colaboração com a Universidade da Califórnia, em San Francisco. A proposta do sistema é a de realizar a tradução de fala entre as Línguas Inglesa e Espanhola para pessoas com pouca proficiência na Língua Inglesa, e com isso, auxiliar os profissionais da saúde a ter uma melhor comunicação com os moradores latinos dos Estados Unidos da América. Uma das limitações do sistema é que o seu uso é restrito para atendimentos a pacientes com diabetes. O sistema opera em um *smartphone* e possui o seguinte fluxo de operação: (1) início da fala, em inglês, do médico para recomendação de um medicamento; (2) seleção em inglês de uma das três opções de medicamentos sugeridas; (3) de acordo com a escolha, o S-MINDS traduz a opção para o espanhol; (4) correção verbal da frase pelo paciente latino; (5) escolha realizada pelo paciente dentre as três opções em Língua Espanhola; e (6) de acordo com a escolha do paciente, o sistema informa a opção ao médico em Língua Inglesa (Soller et al., 2012).

Para aperfeiçoar a coleta de dados em atendimentos clínicos com o paciente, Gür (2012) propôs um sistema denominado Fairwitness para gerar transcrição por meio de SRAF, para a Língua Inglesa, durante o atendimento clínico a um paciente. O sistema tem como objetivo melhorar a qualidade das informações obtidas durante uma consulta. O seu funcionamento é realizado da seguinte maneira: (a) escuta uma conversa entre um paciente e um médico; (b) utiliza um SRAF para gerar a transcrição; (c) aplica métodos de processamento de linguagem natural para extrair os fatos clínicos importantes da conversa; (d) apresenta essas informações em tempo real ao médico, permitindo correção de erros de compreensão; e (e) organiza esses fatos em uma nota do atendimento que pode servir como um primeiro rascunho de anotações produzidas pelo clínico (Gür, 2012).

No trabalho de Prevedello et al. (2014) foi proposta uma implementação de um SRAF para a tarefa de confecção de relatórios de radiologia em Língua Inglesa, considerando o efeito no tempo de confecção desses documentos. O sistema foi implementado em um hospital comunitário com 150 leitos, entre maio de 2011 e julho de 2011. Segundo os autores, durante esse período, a implementação do SRAF resultou em uma redução de tempo de 24 horas para cerca de uma hora. Porém, não foi avaliado o conteúdo do relatório entre os períodos de pré-intervenção e pós-intervenção para avaliar possíveis alterações qualitativas em sua produção.

O SRAF da Nuance, para a Língua Alemã, foi utilizado no trabalho de Vogel et al. (2015), para descrever os efeitos do seu uso para documentação clínica referente a velocidade de documentação, tamanho do documento e satisfação do médico. A avaliação foi realizada no Hospital Universitário de Düsseldorf com 28 médicos. Após o término do documento, o médico realizava as correções necessárias, por exemplo, erros ortográficos e de reconhecimento, e finalizava informando o seu humor em uma escala de 1 a 3. O tempo foi calculado a partir do número de caracteres por minuto, o tempo de transferir o texto gerado pelo SRAF até o sistema de documentação do hospital e do tempo de correções.

Segundo os autores, a documentação médica com a assistência do SRAF baseado na *Web* leva a um aumento na velocidade, na quantidade da documentação e na melhora do humor do médico em comparação com a digitação. Além da economia de tempo, também houve aumento na quantidade de documentação produzida (Vogel et al., 2015).

No trabalho de Ahlgrim et al., (2016), foi proposto um estudo de caso para a implementação de um SRAF em um departamento ambulatorial. A abordagem consiste em criar um dicionário específico para o contexto de medicina esportiva e uma função de aprendizagem de vocabulário compartilhado. A inclusão das palavras no dicionário foi baseada a partir da análise de documentos clínicos e a função de aprendizagem de vocabulário consiste em extrair vocabulário de um usuário para permitir atualizações regulares da lista geral de palavras. A avaliação da abordagem foi realizada considerando a satisfação do usuário, por meio da aplicação de um questionário antes e dez semanas após a implementação do SRAF, e o seu impacto no tempo até o documento médico final ser salvo no sistema. Como resultado, os autores relataram redução para a confecção do documento médico de oito dias para quatro dias. Também foi relatado que a utilização de SRAF pode ser empregada nos departamentos ambulatoriais quando os usuários são colocados em foco.

Por fim, por se tratar de um SRAF de código aberto – *open source* – e desenvolvido especificamente para a Língua Portuguesa do Brasil, o Coruja também foi incluído em trabalhos relacionados. Segundo o autor, a taxa WER reportada nesse trabalho foi de 39,57% para o modelo independente de locutor e de 22,29% para o modelo dependente de locutor, isto é, o modelo do sistema foi adaptado para reconhecer a fala de um indivíduo específico (Batista, 2013).

De acordo com Batista (2013), o Coruja pode ser utilizado como um sistema de ditado e também como uma aplicação para o atendimento automático em uma central de atendimento telefônico. O aplicativo denominado de SpeechOO permite o ditado, edição e formatação de textos por meio de comandos de fala. Os comandos de fala para edição incluem as correções de pontuação, navegação, fonte, cor, formatação e comandos alternativos como salvar e corrigir. O comando corrigir permite a correção de alguma palavra reconhecida erroneamente. O SpeechOO é utilizado no processador de texto *Writer* do pacote LibreOffice.

Segundo o autor, o Coruja também foi integrado ao *software* Asterix para atuar como uma unidade de resposta audível da central de atendimento telefônico do Disque Denúncia Nacional, que é um serviço da Secretária de Direitos Humanos da Presidência da República. A sua utilização consiste em ouvir o nome da cidade pronunciada pelo usuário, e após ter o seu nome reconhecido, é retornado o número de telefone do conselho tutelar da cidade solicitada (Batista, 2013).

3.4 Arquitetura Geral de Sistemas de Reconhecimento Automático de Fala

O primeiro componente a ser considerado para realizar o reconhecimento de fala é o microfone responsável por transformar a onda de pressão sonora para um sinal elétrico. O microfone também serve como pré-amplificador sonoro para que o sinal possa ser transmitido ou digitalizado de maneira satisfatória. Quando a pressão de onda sonora atinge a membrana do microfone ocorre a sua vibração. Como resultado, a onda sonora é convertida em um sinal elétrico analógico. Esse sinal é transportado ao longo de um canal de digitalização que o entrega como um sinal digital ao SRAF (Virtanen et al., 2012).

De maneira geral, o funcionamento de um SRAF é realizado do seguinte modo: o sinal digital da fala de entrada passa pela Extração de Características, que pré-processa o sinal para gerar características espectrais que correspondem à informação do comportamento do sinal de onda. Isto é, o sinal é representado por números para alimentar um sistema digital. As características espectrais são passadas para um estimador de probabilidade de fonemas que estima a melhor unidade de som que corresponda a um fonema localizado no Modelo Acústico. Em seguida, as palavras formadas a partir dos fonemas que estão contidos no Léxico, são comparadas com as palavras contidas no Modelo de Linguagem. Por fim, o Decodificador converte as palavras para o formato de texto (Yu & Deng, 2015).

Na Figura 3.1, é ilustrado o diagrama da arquitetura geral de um SRAF, no qual é representada a sequência de processos necessários, a partir da entrada do sinal de fala, para sua decodificação e transcrição.

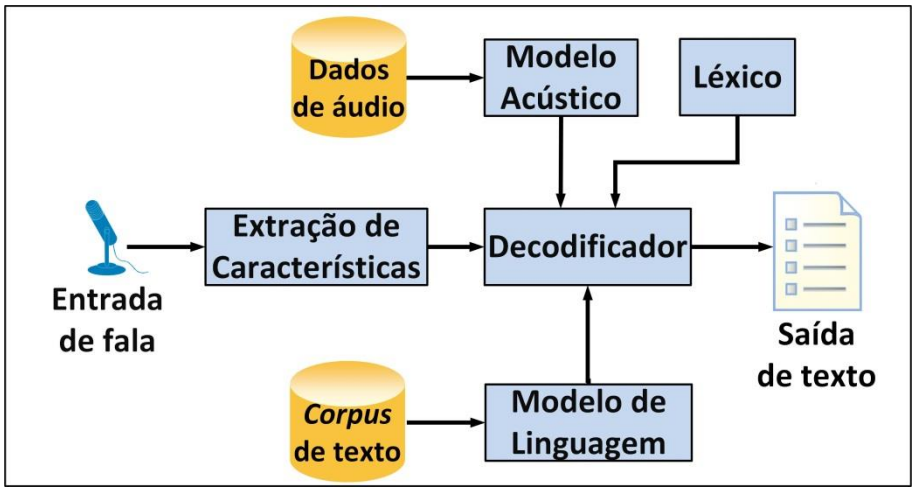

Figura 3.1: Arquitetura Típica de um Sistema de Reconhecimento Automático de Fala.
Fonte: Adaptado de Schalkwyk et al. (2010); Gold (2011); Yu & Deng (2015).

A seguir, serão descritos em detalhes os quatro componentes gerais de um SRAF. Na Subseção 3.4.1, a Extração de Características e o método Coeficientes Cepstral de Frequência

Mel são abordados. Na Subseção 3.4.2 é descrito o Modelo de Linguagem. Na Subseção 3.4.3 é apresentado o Modelo Acústico e na sequência é descrito o Decodificador (Subseção 3.4.4).

3.4.1 Extração de Características

O reconhecimento de fala não é realizado diretamente com o sinal da fala. Inicialmente, a informação contida no sinal da fala possui conteúdo espectral com modulação – processo de variação da amplitude, intensidade, frequência e comprimento da onda –, que varia ao longo do tempo. Dessa maneira, o componente Extração de Características do SRAF calcula uma sequência de vetores de características para capturar as características espectrais mais relevantes do sinal da fala (Virtanen et al., 2012).

Para a análise do sinal espectral é necessário realizar a extração de características para reduzir a variabilidade do sinal. Em particular, o cálculo do espectro de curta duração reduz a variabilidade ao suavizar o espectro detalhado, e assim, são eliminadas várias fontes de informações, ou seja, ocorre a remoção das variações aleatórias dos dados do espectro (Anusuya & Katti, 2011).

Portanto, a extração de características tem como objetivo representar o sinal da voz (Figura 3.2A) através da captura de informações relevantes do seu espectro (Figura 3.2B). De maneira geral, as técnicas de extração de características são classificadas de modo a realizar a análise temporal do formato da onda sonora.

Na Figura 3.2 é ilustrada a frase “gerar laudo por reconhecimento de fala” pronunciada por um falante. As linhas contínuas na vertical representam a divisão entre a pronúncia de cada palavra e as linhas tracejadas na vertical representam, de maneira aproximada, a fronteira onde há ocorrência de pausa durante a pronúncia.

Na fala contínua, duas palavras podem ser pronunciadas sem pausa entre uma palavra e outra, conforme ilustrado na Figura 3.2 em que a pronúncia das palavras “gerar” e “laudo” são pronunciadas sem pausa entre elas.

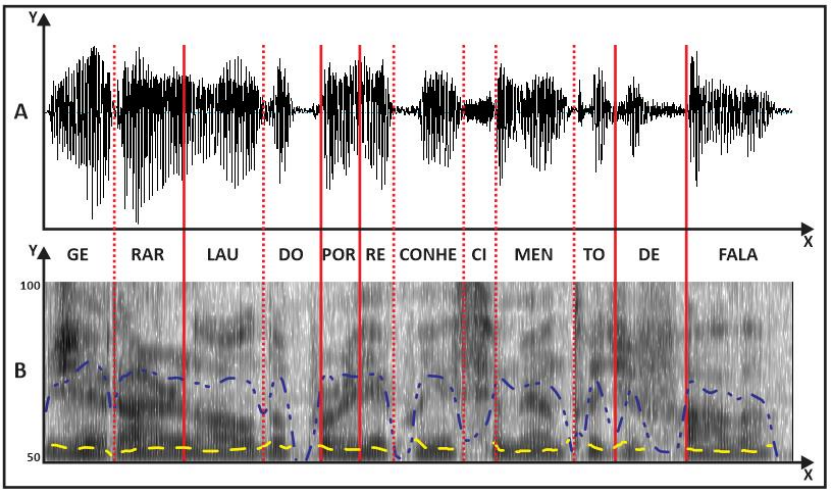

Figura 3.2: (A) Sinal acústico de uma onda sonora; e (B) Espectro de uma onda sonora. Ambos correspondentes à pronúncia da frase “gerar laudo por reconhecimento de fala”.

Na Figura 3.2A e 3.2B, o eixo Y representa a amplitude, ou seja, a oscilação da onda. A Figura 3.2B representa o espectro do sinal acústico da fala. O eixo Y possui variação de 50 a 100 dB, cuja representação simboliza a quantidade de energia carregada ao longo da onda sonora. As partes mais escuras indicam as palavras pronunciadas com maior energia, isto é, maior quantidade de dB. Já a as partes mais claras representam as pronúncias com menor energia. A linha tracejada, com traços de mesmo tamanho, representa a intensidade da pronúncia. A linha tracejada seguida por dois traços menores representa a frequência da pronúncia.

Para capturar a dinâmica dos movimentos do trato vocal, o espectro de curta duração é tipicamente calculado a cada período de 10 a 20 milissegundos (ms) utilizando uma janela entre 20 e 30 ms. O espectro de curta duração é calculado para simular o funcionamento da audição humana, similar à cóclea do ouvido, que ocorre em uma escala de frequência não linear, conhecida como escala Bark ou escala Mel. Essa escala é aproximadamente linear até cerca de 1.000 Hertz (Hz) e aproximadamente logarítmica posteriormente (Anusuya & Katti, 2011).

O espectro pode ser representado diretamente em termos dos coeficientes de Fourier, isto é, o espectro do sinal representado no domínio de frequência é transformado para uma função de tempo para que possa ser possível decompor uma função de soma de um número potencialmente infinito de componente de frequência para uma onda. Outra forma de representar o espectro é por meio de um conjunto de valores de energia nas saídas de um Banco de Filtros, que é um filtro de resposta a um impulso com duração finita dispostos linearmente ao longo da escala Mel (Anusuya & Katti, 2011).

Uma maneira de representação acústica em SRAFs é o fonema utilizado como passo intermediário entre sinais acústicos e palavras específicas. Os modelos acústicos representam fonemas ou trifonemas, permitindo que palavras sejam formadas a partir de sequências dessas unidades (Virtanen et al., 2012; Davis & Scharenborg, 2017).

Dentre algumas das técnicas de extração de características propostas na literatura, pode-se citar o método Percepção Linear Preditiva – *Perceptual Linear Predictive* (PLP) – e o Coeficientes Cepstral de Frequência Mel – *Mel-Frequency Cepstral Coefficients* (MFCC) (Anusuya & Katti, 2011).

A técnica de extração de características PLP estima o espectro sonoro para SRAFs independentes de falante. Essa técnica surgiu 10 anos após o MFCC (Hermansky, 1990), no entanto, a grande maioria dos SRAFs da atualidade utilizam o método de extração de características baseados em MFCC. Sendo assim, na seção seguinte é apresentada a técnica MFCC (Clark et al., 2010; Virtanen et al., 2012).

Coeficientes Cepstral de Frequência Mel

O ouvido humano tem a capacidade de resolver frequências não lineares por meio do espectro da fala. Similarmente a essa característica, os SRAFs que operam com essa mesma capacidade possuem desempenho de reconhecimento superior em relação a outras técnicas,

dessa maneira, o MFCC é uma escolha adequada para tarefas de reconhecimento baseada em recursos de áudio (Smaragdis et al., 2009), o qual foi proposto pela primeira vez por Davis & Mermelstein (1980).

A utilização do MFCC para a extração de características é realizada a partir da entrada de um sinal de fala. A etapa de segmentação (quadros) divide o sinal em quadros com duração de 30 ms com sobreposição entre os quadros de 10 ms. Cada quadro é multiplicado pela função de Hamming para manter a continuidade do sinal da fala (Smaragdis et al., 2009). A aplicação de uma janela de Hamming auxilia a produzir uma estimativa suavizada da energia em regiões onde a energia muda rapidamente. Na sequência, a Transformada Rápida de Fourier – *Fast Fourier Transform* (FFT) – determina as frequências predominantes em um dado segmento, convertendo cada segmento do domínio de tempo para o domínio de frequência para gerar o espectro da fala. Sendo assim, a FFT irá permitir o desmembramento dos estímulos selecionando o som com determinadas frequências. Em seguida, aplica-se o Banco de Filtro correspondente à escala Mel, que separa o sinal de entrada em vários componentes, ao espectro obtido no FFT para transformar a escala de frequência em uma escala linear. A Transformada Discreta de Cosseno – *Discrete Cosine Transform* (DCT) – reduz a dimensionalidade do espectro mantendo apenas as frequências de maior energia. Nessa fase final, o espectro é convertido novamente para o domínio de tempo para proporcionar uma boa representação das propriedades espectrais locais do sinal para a análise de quadros. Com isso, tem-se um vetor de características de MFCC, composto por unidades de fonemas de fala (Smaragdis et al., 2009; Anusuya & Katti, 2011; Lopes-Filho, 2013; Li et al., 2016). Na Figura 3.3 é ilustrado os processos necessários para a realização do cálculo do MFCC.

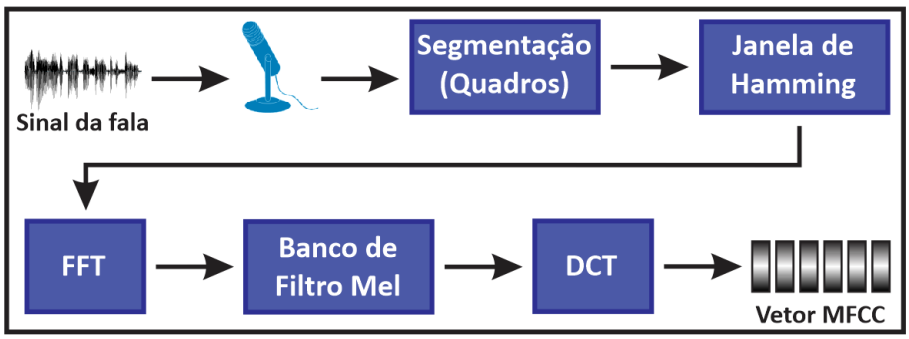

Figura 3.3: Diagrama do processo Coeficientes Cepstral de Frequência Mel – *Mel-Frequency Cepstral Coefficients* (MFCC).

Na Figura 3.4 é ilustrado a comparação entre o funcionamento do processo de audição humana com a técnica de extração de característica baseada em MFCC. Como resultado, são obtidas as unidades fonêmicas.

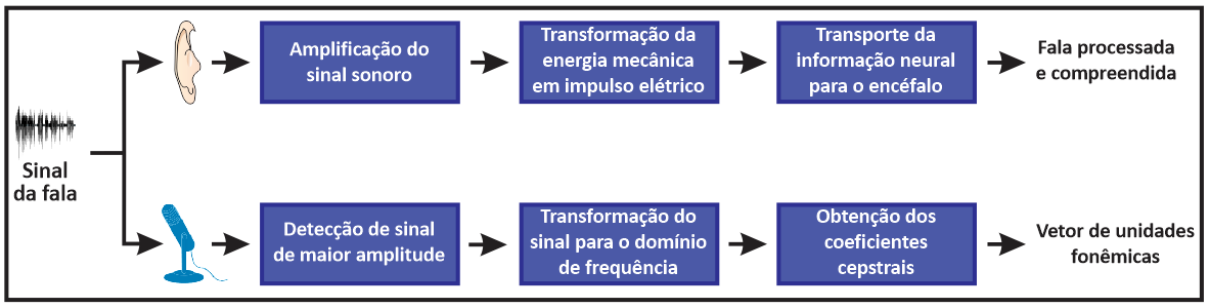

Figura 3.4: Diagrama de comparação entre a audição humana com a técnica Coeficientes Cepstral de Frequência Mel – *Mel-Frequency Cepstral Coefficients* (MFCC).

3.4.2 Modelo de Linguagem

O conhecimento da linguagem é importante para realizar o reconhecimento e o entendimento da fala natural. Para isso, é necessário o conhecimento lexical, que é a definição do vocabulário e a sua respectiva pronúncia para cada palavra de um idioma. Além do léxico, também é necessário o conhecimento da sintaxe e da semântica da linguagem, ou seja, das regras que determinam quais sequências de palavras são gramaticalmente adequadas em termos de formulação e significância. Outro aspecto importante a ser considerado é o conhecimento da pragmática da linguagem que é o que as pessoas podem falar em contextos específicos (Huang et al., 2001).

De maneira geral, existem dois tipos de Modelo de Linguagem (ML). Um ML baseado em regras fixas definidas manualmente, no qual é descrita uma sequência de palavras como certas ou erradas. Uma desvantagem é que essa abordagem pode ser usada apenas em sistemas com um pequeno vocabulário com um número limitado de diferentes sequências de palavras possíveis. No caso de sistemas com grandes vocabulários, onde não é possível implementar todas as regras contidas em uma gramática, é utilizado ML denominado de modelo estocástico, baseado em estatística ou probabilidade. Tal abordagem é utilizada para estimar a probabilidade de que a sequência de palavras dada esteja presente no ML (Donaj & Kačič, 2017).

Os MLs baseados em modelo probabilístico utilizam a técnica de aproximação *n*-grama, que é uma subsequência de n elementos para formar uma sequência para calcular a probabilidade de sequência de palavras, cuja palavra atual depende somente de $n-1$ palavras anteriores. Basicamente, o modelo *n*-grama é uma cadeia de Markov de ordem $n-1$. No contexto de ML, a ordem do modelo é definida pelo número de palavras do *n*-grama, podendo ser unigrama (ordem 1), bigrama (ordem 2) e trigrama (ordem 3). Para ordens superiores, utiliza-se o dígito, por exemplo, 4-grama, 5-grama, e assim sucessivamente (Arısoy et al., 2008; Donaj & Kačič, 2017).

Um ML probabilístico estima valores de probabilidade prévia, $P(W)$, para sequências de palavras W em um vocabulário (Equação 3.1). No entanto, é computacionalmente custoso estimar todas as possíveis sequências de palavras caso seja necessário procurar em um modelo

probabilístico, normalizado de maneira adequada, dada uma sequência de palavras com comprimento finito. Uma maneira de garantir a normalização adequada do modelo é a de decompor a probabilidade de sentenças e certificar-se de que o símbolo de fim de frase, </s>, seja previsto com probabilidade não nula e em qualquer contexto. Com $W = w_1, w_2, \dots, w_n$, em que cada $w_i, i = 1, \dots, n$, representa uma palavra. O conjunto, $P(w_i|w_1, w_2, \dots, w_{i-1})$, representa a probabilidade de $w_i$ser a sequência de palavra correspondente. Dessa maneira, tem-se (Clark et al., 2010):

$$P(W) = \prod_{i=1}^{n} P(w_i|w_1, w_2, \dots, w_{i-1}). \tag{3.1}$$

A sequência de palavras encontradas em aplicações práticas possui comprimento finito. A probabilidade de distribuição, $P(W)$, deve atribuir probabilidade 0 a sequências de palavras de comprimento infinito e 1 para o conjunto da sequência de palavras de comprimento finito. Do ponto de vista da modelagem, em uma situação prática, o texto é dividido em frases, e o ML precisa prever o símbolo distinto de fim de frase para garantir que o modelo atribua probabilidade 1. Dessa maneira é garantido que a sequência de palavras não possua comprimento infinito (Clark et al., 2010).

Para exemplificar, a Google desenvolveu um ML para realizar pesquisa por fala treinado com mais de 230 bilhões de palavras usando um vocabulário de 1 milhão de palavras. O tamanho desse conjunto de dados apresenta desafios únicos. Em contrapartida, oferece oportunidades para melhorar a modelagem de linguagem (Schalkwyk et al., 2010). Essa grande quantidade de palavras em um ML resulta em um modelo capaz de fornecer maiores probabilidades de sequências de palavras corretas.

No entanto, por maior que seja a quantidade de palavras utilizadas para realizar o treinamento de um ML, nunca haverá dados suficientes para representar todas as relações estatísticas entre as palavras formuladas durante uma comunicação por fala (Gold, 2011), devido ao fato de que durante a fala existem infinitos padrões de palavras que podem ser formuladas pelo orador (Huang et al., 2001).

Outro agravante para a construção de ML se refere a vocabulários em línguas aglutinantes[8], por exemplo, Finlandês, Estoniano, Húngaro e Turco, nas quais o desempenho de reconhecimento não se assemelha ao desempenho alcançado por sistemas como para a Língua Inglesa. A principal razão dessa degradação do desempenho é devido à rica estrutura morfológica. Em línguas aglutinantes, as palavras são formadas principalmente por concatenação de vários sufixos com as raízes e, juntamente com composição e inflexões, resultando em milhões de formas de palavras diferentes. Portanto, é praticamente impossível construir um ML que cubra todas as palavras relevantes dos idiomas aglutinantes (Arısoy et al., 2008; Clark et al., 2010).

[8] Línguas em que ocorre a união de duas ou mais palavras distintas em uma só palavra para formar uma nova.

Devido a essas dificuldades para a construção de MLs, surgiu a necessidade de minimizar as taxas de erros. Isso é realizado atribuindo menores pontuações a frases improváveis. Com isso, os MLs probabilísticos estimam, por meios estatísticos, uma determinada sequência de palavras, com base em quantas vezes essa sequência, ou parte dela, foi encontrada no *corpus* de treinamento (Donaj & Kačič, 2017).

3.4.3 Modelo Acústico

Um Modelo Acústico (MA) é um modelo matemático utilizado para estimar a probabilidade do MA conter um determinado fonema ou uma palavra inteira correspondente a um dado sinal de áudio de entrada. A probabilidade de ocorrência de um sinal de áudio é realizada pela maximização da sequência de palavras W, representada por $max\ W$ (Equação 3.2), dado um valor máximo estimado para a sequência de palavras com a observação de um vetor de características O (Huang et al., 2001; Donaj & Kačič, 2017):

$$\tilde{W} = max\ W\ P(W)P(O|W), \tag{3.2}$$

onde, $P(W)$, é a probabilidade de sequência de palavras do ML e, $P(O|W)$, é a probabilidade de sequência de palavras do MA. Quando a equação alcançar a melhor probabilidade de sequência de palavras, tem-se o resultado, representado por $\tilde{W}$.

Para a construção de um MA preciso, é necessário criar um modelo eficiente para realizar o reconhecimento de fala em sistemas com um grande vocabulário. Uma maneira é decompor a representação de uma palavra em uma sequência de subpalavras. Assim, $P(O|W)$ está intimamente relacionada à modelagem fonética, a qual deve levar em consideração as variações do falante, da pronúncia, do ambiente e também as variações de coarticulação fonética dependente de contextos. O processo de decodificação para encontrar a melhor sequência de palavras W, correspondente ao sinal de entrada de fala O, é mais complexo do que um problema simples de reconhecimento de padrão. Isso ocorre, pois no reconhecimento de fala contínua existem infinitos padrões de palavras para buscar (Huang et al., 2001).

3.4.4 Decodificador

A tarefa para se reconhecer a fala contínua é um problema de reconhecimento e de busca de padrões. A ideia é decodificar um sinal previamente codificado pelo processo de origem e transmitido através do canal de comunicação. Dessa maneira, esse processo objetiva encontrar uma sequência de palavras encontradas nos MAs e MLs que melhor correspondam ao sinal de entrada (Huang et al., 2001).

O Decodificador é o componente principal de um SRAF. A sua função consiste em determinar a melhor sequência de palavras $\tilde{W} = w_1, w_2, \ldots, w_n$, cujo comprimento é desconhecido a partir de todas as sequências de palavras possíveis, W, que maximiza a probabilidade do MA e do ML. Isso é realizado de maneira ponderada, podendo atribuir peso

fixado em 1 para o MA, dessa maneira, é necessário a atribuição de apenas um peso. O peso para o ML é representado por α. Devido à preferência do algoritmo de busca do Decodificador em selecionar palavras curtas ao invés de palavras longas, é adicionado um fator heurístico chamado de penalização de inserção de palavra, β, dado o número de palavras N. Com isso, o algoritmo de busca do Decodificador pode ser resolvido de acordo com a Equação 3.3 (Donaj & Kačič, 2017):

$$\tilde{W} = max\, W\{\alpha\, log\, P(W) + \beta N + log\, P(O|W)\}, \tag{3.3}$$

onde $P(W)$ representa a probabilidade do Modelo de Linguagem e $P(O|W)$ a probabilidade do Modelo Acústico.

Para as várias sentenças possíveis são atribuídas pontuações. A sentença com maior pontuação é a probabilidade máxima. Todas as várias sentenças são chamadas de hipóteses. Durante o reconhecimento, o algoritmo de busca acompanha um número limitado de hipóteses parciais. A hipótese com melhor pontuação será o resultado final do algoritmo de busca. Em outros casos, o resultado pode ser uma lista, uma rede de palavras ou alguma outra representação com várias hipóteses (Donaj & Kačič, 2017).

3.5 Fontes de Variabilidade Acústica

Os sinais da fala são frequentemente influenciados por fatores externos que não possuem relação com o que foi pronunciado pelo falante. Essas influências podem resultar em incompatibilidade entre as distribuições dos dados a serem classificados e, mesmo que a classificação seja realizada com as distribuições adequadas, isso pode resultar em aumento de erro durante o reconhecimento (Virtanen et al., 2012).

A variação do sinal acústico pode ser agrupada em quatro áreas principais (Clark et al., 2010):

- **Domínio da Tarefa:** o domínio da tarefa inclui o idioma e o tamanho do vocabulário, afetando o processo de transcrição para o reconhecimento da fala. Diferentes idiomas apresentam desafios distintos para um SRAF, como língua aglutinante e não aglutinante (Arısoy et al., 2008);
- **Características do Falante:** diferença na produção da fala devido à anatomia e à fisiologia individual de cada pessoa, como ritmos, uso de linguagem com graus de variabilidade intrínseca e variações sistemáticas, como o sotaque[9] e a idade do falante;
- **Estilo da Fala:** os SRAFs contínuos devem ser capazes de identificar o limite entre cada palavra, ou seja, deve haver a segmentação e rotulação das palavras;

[9] Variação na pronúncia entre indivíduos.

- **Ambiente:** o ambiente acústico no qual a fala é gravada, juntamente com qualquer canal de transmissão, pode impactar de maneira significativa o nível de precisão de um SRAF. Por exemplo, na parte externa dos escritórios e laboratórios silenciosos, geralmente há múltiplas fontes acústicas, incluindo ruído ambiente, outros oradores e dispositivos elétricos ou mecânicos. Em muitos casos, é um problema significativo separar os diferentes sinais acústicos encontrados em um ambiente. Além disso, há também as variações nos canais de transmissão, que ocorrem devido à movimentação de cabeça do locutor durante a gravação do seu discurso. Provavelmente, a maior disparidade entre a precisão de um SRAF em comparação com o reconhecimento de fala por uma pessoa ocorre em situações onde existem ruídos, como múltiplas fontes acústicas ou ambientes com reverberações.

A reverberação é um fator agravante na captura do sinal da fala, pois ocorre a persistência do som logo após ter sido extinto. Geralmente a reverberação ocorre em ambientes fechados ou parcialmente fechados (Virtanen et al., 2012).

Devido à necessidade de sistemas centrados no usuário, ou seja, sistemas que se adequam às características específicas para cada usuário, tem sido exigido que os SRAFs se tornem cada vez mais robustos no contexto relacionado ao ruído do ambiente real e outras condições de distorção do canal acústico, por exemplo (Li et al., 2016):

- Reduzir o nível de ruído explorando o *hardware* utilizando informações espaciais ou direcionadas do microfone e dos princípios dos transdutores para a conversão da onda sonora em sinal elétrico, como a utilização de microfones com cancelamento de ruídos;
- Utilizar processamento algorítmico, por meio de *software*, aproveitando a separação espectral e temporal entre os sinais da fala e a interferência.

Também existe o ruído aditivo em que o sinal gravado é somado a diferentes fontes sonoras. Por exemplo, quando ocorre a interposição com outros sinais indesejados, como sons de falantes concorrentes, rádio, ar-condicionado ou fonte de sons difusos como o de um automóvel. Outros ruídos são introduzidos pelo próprio equipamento de gravação, mesmo quando não há fonte de ruído externa (Virtanen et al., 2012).

3.6 Métricas de Avaliação de Desempenho para Sistemas de Reconhecimento Automático de Fala

Em SRAFs é fundamental avaliar o desempenho do reconhecimento. Uma métrica que pode ser utilizada é a taxa de palavras reconhecidas – *Word Recognition Rate* (WRR) –, que consiste em contar o número de palavras geradas corretamente pelo SRAF (Gavat et al., 2008). Outra medida amplamente utilizada é a taxa de erro de palavra – *Word Error Rate*

(WER). Nessa avaliação, normalmente, existem três tipos de erros (Huang et al., 2001; Clark et al., 2010; Virtanen et al., 2012):

- **Substituição:** uma palavra correta foi substituída por outra;
- **Deleção:** uma palavra correta foi omitida na frase reconhecida;
- **Inserção:** uma palavra não pronunciada pelo orador foi adicionada no texto gerado pelo SRAF.

A WER é proveniente da distância de Levenshtein, ou distância de edição entre duas sequências de palavras, em que o número mínimo, ou soma ponderada, de inserções, deleções e substituições necessárias para transformar a sequência de palavras em outra, é verificada (Levenshtein, 1966).

Para determinar a taxa WER (Equação 3.4), não se pode apenas comparar duas sequências de palavras uma a uma. Ou seja, é preciso alinhar uma sequência de palavra reconhecida com a sequência de palavra correta, somando o número de substituições (S), deleções (D) e inserções (I) dividido por N_r, que é o número total de palavras de referência na frase (Huang et al., 2001):

$$WER = 100\frac{S+D+I}{N_r}. \tag{3.4}$$

Por exemplo, na frase “novo laudo médico gerado” caso seja reconhecida como “laudo médico não gerado”, se houver uma comparação de palavra por palavra, a taxa de erro é de 75%, pois se compara “novo” versus (vs.) “laudo”, “laudo” vs. “médico”, “médico” vs. “não” e “gerado” vs. “gerado”, com apenas uma palavra correta “gerado”. Ao utilizar a WER, a taxa de erro é de 50% com uma deleção da palavra “novo” e uma inserção da palavra “não”, conforme é apresentado na Equação 3.5:

$$WER = 100\frac{0+1+1}{4} = 50. \tag{3.5}$$

3.7 Considerações Finais

Neste capítulo foram apresentadas algumas limitações presentes na tecnologia de SRAFs, bem como, uma evolução histórica e como esses sistemas estão sendo utilizados em trabalhos recentes.

A evolução da tecnologia para se realizar o processo de reconhecimento de fala em sistemas computacionais viabilizou a sua utilização em diversas áreas da sociedade. Desde os primeiros ensaios, com uma tecnologia capaz de simular a maneira como o ser humano compreende os sons acústicos até os SRAFs mais modernos, houve uma série de melhorias nas técnicas computacionais utilizadas.

Algumas dessas melhorias incluem a extração de características, que embora haja algumas limitações, já é capaz de suprimir alguns tipos de ruídos e de variabilidade acústica. Com relação ao treinamento para o Modelo de Linguagem, utiliza-se um conjunto cada vez maior de textos e de diferentes áreas de conhecimento. O mesmo ocorre com o Modelo Acústico, sendo treinado com cada vez mais dados de áudio, devido ao crescente número de áudios e vídeos disponibilizados na *Internet*. Por fim, o uso de técnicas avançadas nos algoritmos de busca dos Decodificadores permite que esses sistemas alcancem níveis de precisão cada vez maiores.

Além disso, foram descritas as fontes de variabilidade acústica que degradam o sinal, prejudicando o reconhecimento da fala. Por fim, foram apresentadas as principais métricas para avaliar a precisão de SRAFs.

Com base na arquitetura de um SRAF e dos componentes que o compõem, no próximo capítulo serão descritos os principais algoritmos de aprendizado de máquina, utilizados para esses sistemas.

Capítulo 4

Aprendizado de Máquina para Sistemas de Reconhecimento Automático de Fala

O aprendizado de máquina busca fazer com que as tecnologias computacionais aprendam uma determinada função que, geralmente, é realizada por pessoas, para que se possa agilizar e automatizar alguns processos.

Mais especificamente, em reconhecimento automático de fala, vários algoritmos foram propostos para viabilizar o seu uso em condições reais. A busca por melhores soluções levou ao desenvolvimento de tecnologias cada vez mais eficientes.

Um resumo de algumas das principais tecnologias para SRAF são apresentadas na Seção 4.1. Nas seções seguintes são descritas em mais detalhes, as seguintes tecnologias: Modelo Oculto de Markov (Seção 4.2) e Rede Neural Artificial (Seção 4.3), com as suas variações: Perceptron Multicamadas (Seção 4.3.1) e Rede Neural Recorrente (Seção 4.3.2).

4.1 Principais Algoritmos Utilizados em Sistemas de Reconhecimento Automático de Fala

No decorrer dos anos surgiram técnicas para fazer com que os computadores aprendessem a partir de um conjunto de experiências. Tom Mitchel definiu aprendizado de máquina como sendo um programa de computador que aprende dada uma experiência *E* com relação a alguma classe de tarefas *T* e uma medida de desempenho *D*, cujo desempenho *D*, em uma tarefa *T*, melhora o desempenho *D*, a partir de uma experiência *E* (Mitchell, 1997).

Uma das técnicas de aprendizado de máquina comumente utilizadas em SRAFs é baseada em cadeia de Markov, dando origem ao Modelo Oculto de Markov – *Hidden Markov Model* (HMM) – (Jurafsky & Martin, 2016; Davis & Scharenborg, 2017).

Outra técnica de aprendizado de máquina empregada nesse contexto é o Modelo de Mistura Gaussiana – *Gaussian Mixture Models* (GMM), modelo utilizado para representar a distribuição dos dados para a modelagem acústica em sistemas convencionais de reconhecimento automático de fala. Quando as formas de ondas da fala são processadas pela transformada de Fourier, a distribuição da mistura Gaussiana é utilizada para ajustar tais características da fala quando a informação sobre a ordem temporal é descartada. Isto é, pode-se utilizar a distribuição de mistura Gaussiana como um modelo para representar características da fala baseadas em quadros – *frames* (Yu & Deng, 2015).

O GMM também pode ser utilizado em conjunto com o HMM, originando o método híbrido GMM-HMM, que é um modelo estatístico utilizado para descrever dois processos aleatórios dependentes, sendo um processo observável e um processo oculto de Markov. A sequência observável é gerada por cada estado oculto de acordo com a distribuição da mistura Gaussiana (Yu & Deng, 2015).

Durante a década de 1980, surgiram os primeiros esforços de pesquisa para explorar o uso de redes neurais artificiais para a tarefa de reconhecimento de fala. Tais redes são sistemas de multicamadas de processamento, que calculam a soma ponderada de suas entradas e são passadas para camadas subsequentes de processamento (Davis & Scharenborg, 2017).

Uma Rede Neural Artificial – *Artificial Neural Network* (ANN) possuem estrutura de multicamadas de Perceptron Multicamadas – *Multi-layer Perceptron* (MLP) – (Seção 4.3.1), isto é, com várias camadas ocultas, e como entrada da rede tem-se um vetor de características acústicas formado por um conjunto de quadros sucessivos. A saída da ANN é um vetor de probabilidades posteriores, contendo um elemento para cada fonema (Li et al., 2016).

Com relação a arquitetura de aprendizagem profunda, Rede Neural Profunda – *Deep Neural Network* (DNN) –, possui um conjunto de algoritmo de aprendizagem de máquina para modelar abstrações de alto nível em dados, utilizando para isso, uma arquitetura composta de múltiplas camadas. Cada camada de processamento utiliza a saída da camada anterior como processamento para a camada seguinte. Uma DNN combina extração de características acústicas com a classificação fonética da fala em uma única estrutura. Dessa maneira, as DNNs garantem que tanto a extração de características como a classificação sejam otimizadas (Li et al., 2016).

A Rede Neural Recorrente (Seção 4.3.2) se mostrou um método apropriado para codificar a estrutura temporal para representar a entrada sequencial da fala (Davis & Scharenborg, 2017), pois essa arquitetura possibilita o aprendizado sequencial das dependências estendidas ao longo de um período de tempo (Li et al., 2016).

Com relação ao método híbrido baseado em ANN, tem-se a ANN-HMM, a ANN substituí o modelo acústico GMM para a tarefa de avaliação da pontuação de verossimilhança. Já o modelo híbrido baseado em DNN-HMM aplica a capacidade de aprendizado de uma DNN com a modelagem sequência de um HMM, superando o modelo GMM-HMM em tarefas de reconhecimento de fala contínuo de grande vocabulário. No modelo DNN-HMM, a dinâmica do sinal da fala é modelada com HMM e as probabilidades de observações são estimadas através de DNNs. Cada neurônio de saída é treinado para estimar a probabilidade posterior da densidade contínua do estado HMM, dadas as observações acústicas (Yu & Deng, 2015).

4.2 Modelo Oculto de Markov

As cadeias de Markov constituem extensões de autômatos finitos. Um autômato finito ponderado é definido por um conjunto de estados e um conjunto de transições entre estados,

no qual cada arco é associado a um peso. Em outras palavras, a cadeia de Markov é um caso especial de um autômato ponderado, cujos pesos são probabilidades em que a soma dos arcos que deixam um nó deve ser igual a 1. O autômato não pode representar problemas de natureza ambígua, uma vez que a cadeia de Markov é útil apenas para atribuir probabilidade a sequências inconfundíveis (Jurafsky & Martin, 2016).

Para ilustrar um exemplo relacionado a transição de estados, a Figura 4.1 aborda três estados possíveis: S_1, S_2 e S_3. O a_{ij}, com i = 1, 2, 3 e j = 1, 2, 3, representa as possíveis transições para cada um dos três estados. Cada uma das transições possui um peso associado.

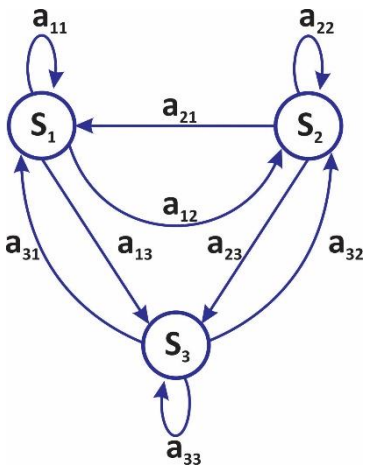

Figura 4.1: Sequência de observação de estados.

O HMM é um modelo probabilístico cujo sistema é modelado por uma cadeia de Markov. O ponto central de um HMM é o conceito de estado, que é em si uma variável aleatória que normalmente possui valores discretos. A extensão de uma cadeia de Markov para um HMM envolve a adição de incertezas ou uma distribuição estatística em cada um dos estados da cadeia de Markov (Yu & Deng, 2015). Um HMM contém como saída uma sequência de símbolos. Cada estado da cadeia de Markov possui uma distribuição de probabilidade associada com os símbolos de saída (Davis & Scharenborg, 2017).

No caso do processamento de língua natural, as probabilidades de sequências de palavras são modeladas com *n*-grama (Subseção 3.4.2) e as probabilidades dos sinais acústicos são modelados com o HMM. Para a modelagem HMM, usualmente são utilizados fonemas, pois o número de fonemas é muito inferior em relação à quantidade de palavras em um idioma. Por isso, geralmente existem dados de treinamento em quantidade suficiente para treinar os HMMs para todos os fonemas (Virtanen et al., 2012).

No reconhecimento de fala, o HMM utiliza modelos que possuem distribuição linear de estados e probabilidades com valores diferentes de 0 e apenas para a transição de um estado para si próprio ou para algum estado mais próximo. Os símbolos de saída são vetores de características e os estados ocultos da cadeia de Markov podem representar palavras ou frases. Para o caso de SRAFs com grande vocabulário, são utilizados vários estados em sequências para representar trifones, ou seja, fonemas classificados com base nos fonemas anterior e do

próximo, com o objetivo de encontrar a sequência de estados ocultos que possa ter produzido a melhor sequência de observação de vetores de características (Davis & Scharenborg, 2017).

O HMM possui duas hipóteses simplificadoras. A primeira hipótese é a de que a probabilidade de um estado particular, da cadeia de Markov, depende apenas do estado anterior, conhecido como pressuposto de Markov. A segunda hipótese consiste no fato de que a probabilidade de uma observação de saída depende somente do estado que produziu a observação e não de qualquer outro estado ou observação (Jurafsky & Martin, 2016).

Os três problemas fundamentais de um HMM são (Yu & Deng, 2015; Jurafsky & Martin, 2016):

- **Problema 1 (probabilidade):** é uma tarefa básica necessária para aplicações de processamento de fala envolvendo HMM que utiliza sequências ocultas de Markov para aproximar características dos vetores da fala, ou seja, a probabilidade é definida a partir de uma sequência de observações;
- **Problema 2 (decodificação):** é para encontrar, de maneira eficiente, a melhor sequência dos estados do HMM dado uma sequência arbitrária de observações. Ou seja, calcular a probabilidade da sequência de observação dada a sequência de estados ocultos;
- **Problema 3 (aprendizagem):** dada uma sequência de observações e de um conjunto de estados no HMM, aprender os parâmetros do HMM. Isso é realizado a partir de uma sequência de observações e de um vocabulário de potenciais estados ocultos.

O diagrama de um SRAF que utiliza o HMM é ilustrado na Figura 4.2, no qual as palavras sublinhadas representam os fonemas. A palavra e a sentença em itálico foram selecionadas como sendo a melhor hipótese encontrada no Léxico e Modelo de Linguagem, respectivamente.

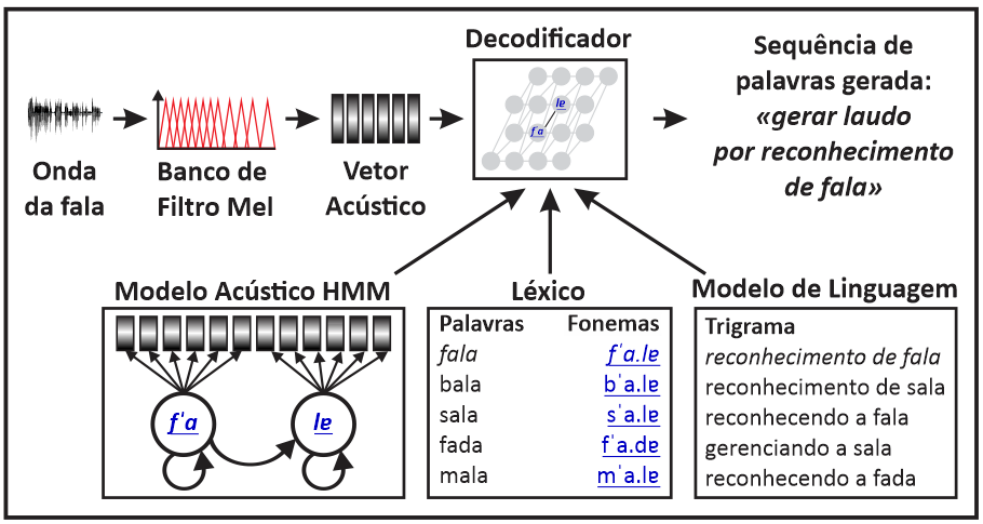

Figura 4.2: Diagrama da arquitetura de um Sistema de Reconhecimento Automático de Fala que utiliza Modelo Oculto de Markov – *Hidden Markov Model* (HMM).
Fonte: Adaptado de Davis & Scharenborg (2017).

A Figura 4.2 ilustra a arquitetura de um típico SRAF, utilizando HMM para identificar a melhor sequência de fonemas, comparando o sinal de onda da fala com os dados de treinamento do Modelo Acústico. Dois processos-chave podem ser considerados: pré-processamento acústico e algoritmo de busca (Decodificador). No pré-processamento acústico o sinal da onda da fala é passado por um Banco de Filtro Mel para gerar o vetor de característica acústica, o qual representa a forma de onda da fala. O Decodificador, utilizando pontuações, é responsável por combinar informações do Modelo Acústico do HMM, do Léxico e do Modelo de Linguagem. O Modelo de Linguagem atribuirá maiores pontuações a sequências de fonemas com maior probabilidade de ocorrência, dada a comparação de similaridade entre o vetor de característica acústica e os fonemas treinados no Modelo Acústico. O Léxico é um dicionário de pronúncia que contém a palavra e a sua respectiva transcrição fonética, em que cada palavra é convertida numa sequência de sons básicos, os fonemas. O Modelo de Linguagem irá estimar a probabilidade de sequência de palavras diferentes, geradas a partir da sequência observada de vetores acústicos. Com isso, a sequência de máxima probabilidade é a melhor hipótese avaliada pelo SRAF.

4.3 Rede Neural Artificial

A Rede Neural Artificial – *Artificial Neural Network* (ANN) – foi desenvolvida pelo neurofisiologista McCulloch e pelo matemático Walter Pitts, da Universidade de Illinois. O conceito de redes neurais consiste em uma analogia entre células nervosas vivas e um processo eletrônico binário (Braga et al., 2000; Oliveira-Junior et al., 2007; Graves, 2012).

As ANNs são utilizadas em soluções de problemas devido a sua capacidade de aprender. O treinamento é a capacidade da rede de ajustar os parâmetros para alcançar os resultados esperados, dado um conjunto de padrões específicos. Tais padrões de treinamento são constituídos de informações que se espera que a rede aprenda. Os ajustes dos parâmetros são realizados com a atribuição de pesos das conexões que interligam os neurônios (Braga et al., 2000; Oliveira-Junior et al., 2007).

O treinamento de uma rede pode ser supervisionado ou não supervisionado. No treinamento supervisionado, o ajuste dos parâmetros é realizado de maneira que, dado uma entrada padrão, o valor é calculado para gerar uma saída. As entradas e a saída desejadas são fornecidas por um supervisor. Isso é feito para ajustar os parâmetros da rede para encontrar uma ligação entre os pares de entrada e saída fornecidos. Ou seja, são fornecidos, previamente para a rede, os valores de entrada e saída. Já no treinamento não supervisionado, o conjunto padrão de treinamento possui apenas entradas, cuja saída padrão não é previamente conhecida (Braga et al., 2000; Oliveira-Junior et al., 2007).

Uma ANN contém uma estrutura com pequenas unidades de processamento (ou neurônios) que são unidos entre si por conexões ponderadas. Ainda seguindo a analogia do modelo biológico, essas unidades de processamento representam os neurônios e os pesos das conexões representam a força das sinapses entre os neurônios. A rede é ativada fornecendo

uma entrada para alguns ou todos os neurônios. Essa ativação se espalha por toda a rede ao longo das suas conexões ponderadas e a atividade elétrica dos neurônios biológicos atinge "picos". Portanto, para a rede neural artificial simular essa característica de atingir picos, utiliza-se da ativação de um neurônio, que é projetada para modelar a taxa média de disparo desses picos (Graves, 2012).

Em outras palavras, a ANN é uma combinação de neurônios artificiais, conexões e algoritmo de aprendizagem. A caracterização do agrupamento de neurônios é dada pela topologia da rede e deve considerar alguns aspectos, como o número de camadas da rede, o número de neurônios por camada, o tipo de conexão e o grau de conexidade entre os neurônios (Braga et al., 2000; Oliveira-Junior et al., 2007).

Com relação ao número de camadas da ANN, é possível uma rede conter uma camada entre as camadas de entrada e de saída. As redes de múltiplas camadas são compostas de camadas de entrada, duas ou mais camadas ocultas e uma camada de saída. Os neurônios podem ter conexões *Feedforward* ou *Feedback*. A conexão *Feedforward* não permite que a saída de um neurônio seja utilizada como entrada em um neurônio de uma camada anterior a sua, diferentemente da conexão *Feedback* que permite a entrada a um neurônio de camada anterior à sua. Com relação à conexidade entre os neurônios da rede, os neurônios podem ser parcialmente ou completamente conectados (Braga et al., 2000).

Uma ANN é constituída por duas fases de processamento: a fase de aprendizagem ou de treinamento e a fase de utilização. No processo de treinamento, são realizados os ajustes dos pesos das conexões em resposta ao estímulo apresentado à rede neural, tornando-a capaz de modificar-se de acordo com a necessidade de aprender uma determinada informação que lhe é destinada. Já o processo de utilização, após treinamento, é caracterizado pela resposta da rede neural baseada em um estímulo de entrada. Entre as camadas de entrada e saída, ocorre o processamento dos dados. Na Equação 4.1 são acumulados, de maneira ponderada, os pesos dos dados recebidos das entradas do neurônio em S_j (Oliveira-Junior et al., 2007).

$$S_j = \sum_{i=1}^{n} x_i w_{ji}, \tag{4.1}$$

onde x_i representa os dados de entrada e w_{ji} representa o peso para cada uma das entradas.

A função de transferência do neurônio, ou seja, a relação entre a entrada e a saída de um sistema recebe o resultado da soma ponderada para ativar ou não o neurônio. O neurônio será ativado caso o somatório seja igual ou maior do que o valor de um determinado limiar. Caso o neurônio seja ativado, a sua saída terá valor 1, caso contrário, a saída do neurônio terá valor 0 (Braga et al., 2000; Oliveira-Junior et al., 2007).

A ANN *Feedforward* mais difundida é a Perceptron Multicamadas (Graves, 2012), conforme explicado na subseção a seguir.

4.3.1 Perceptron Multicamadas

Uma das principais características da arquitetura MLP se refere a função de ativação, que é responsável por ativar ou não a saída de um neurônio, de acordo com o valor da soma ponderada das suas entradas (Braga et al., 2000). A função de ativação mais utilizada é a sigmoidal logística (Braga et al., 2000; Yu & Deng, 2015).

A quantidade de níveis e de neurônios que uma rede possui representa a sua dimensão, enquanto a quantidade de neurônios da camada de entrada e de saída da rede, normalmente, é conhecida pelo usuário (Oliveira-Junior et al., 2007).

Na Figura 4.3, é ilustrada uma MLP, em que as múltiplas camadas estão delimitadas entre as linhas tracejadas. Nessas camadas ocultas é realizado o processamento dos dados de entrada X_m, gerando os dados de saída Y_n, onde m = 1, 2, ..., m e n = 1, 2, ..., n. Cada círculo representa um neurônio da rede.

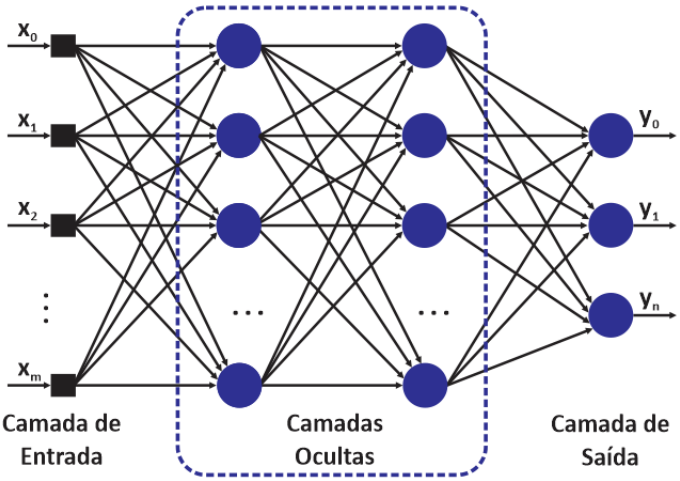

Figura 4.3: Rede Neural com múltiplas camadas.
Fonte: Adaptado de Oliveira-Junior et al. (2007).

Em uma MLP, o processamento realizado por cada neurônio sofre influência do processamento dos neurônios anteriores conectados a ele. Devido a isso, conforme ocorre o processamento dos neurônios de cada camada em direção à saída, as funções se tornam mais complexas. Sendo assim, pode-se dizer que as camadas ocultas desempenham a função de detectores de característica, pois geram uma codificação interna dos padrões de entrada e que servirão para gerar a saída da rede. Com isso, a partir de um número suficientemente grande de camadas ocultas é possível representar qualquer conjunto de padrões de entrada. No entanto, é necessário ter controle da quantidade de camadas ocultas, pois quanto maior o número de camadas, maior a chance de propagação de erros durante o treinamento da rede para as camadas posteriores, resultando em redução na precisão da rede (Braga et al., 2000).

Uma MLP com duas camadas ocultas com quantidade de neurônios suficientes pode aprender qualquer mapeamento arbitrário a partir de um conjunto de dados de entrada para gerar uma saída (Gold, 2011).

O processo de estimação de parâmetros ou o treinamento de uma MLP pode ser especificado por um critério de treinamento e um algoritmo de aprendizagem (Yu & Deng, 2015). Os algoritmos de treinamento utilizados podem ser estáticos ou dinâmicos. O algoritmo estático não altera a estrutura da rede, o que variam são os valores dos pesos. Já o algoritmo dinâmico tem a capacidade de reduzir ou de aumentar o tamanho da rede, podendo alterar o número de camadas, o número de neurônios nas camadas ocultas e o número de conexões (Braga et al., 2000).

O algoritmo de aprendizado para o treinamento da MLP amplamente conhecido é o *back-propagation* que consiste em um algoritmo supervisionado para ajustar os pesos da rede (Haykin, 2008).

O treinamento da MLP ocorre em duas fases. A fase *forward* percorre a rede da camada de entrada até a camada de saída e o sentido *backward* realiza o caminho inverso. A fase *forward* é composta pelos passos: (1) o valor de entrada é apresentado à primeira camada oculta; (2) para cada camada subsequente, os neurônios processados resultam em um valor de saída que é utilizado como entrada para o próximo neurônio da camada seguinte; e (3) a saída produzida pelos neurônios da última camada é comparada às saídas desejadas. A fase *backward* compreende as etapas: (1) é realizado o ajuste dos pesos, iniciando na última camada até chegar à camada de entrada, para reduzir seus erros; e (2) o erro de um neurônio das camadas ocultas é calculado a partir dos erros dos neurônios das próximas camadas conectados a ele. Esse cálculo é ponderado de acordo com os pesos das conexões entre eles (Braga et al., 2000).

Outra arquitetura de aprendizado profunda eficaz para o reconhecimento de fala é baseada em Rede Neural Recorrente – *Recurrent Neural Network* (RNN) – podendo conter uma estrutura celular especial, Memória Longa de Curto Prazo – *Long Short-Term Memory* (LSTM) –, para armazenar informações por um maior tempo. A arquitetura RNN-LSTM também é eficaz em reconhecimento de fala em ambientes com reverberação (Li et al., 2016).

4.3.2 Rede Neural Recorrente

O ponto central de uma RNN é que as conexões recorrentes permitem que uma "memória" de entradas anteriores persista no estado interno da rede exercendo influência na saída da rede (Graves, 2012).

As RNNs possuem conexões *Feedback*, que as tornam adequadas para o uso em modelagem sequencial, sendo utilizadas com sucesso em tarefas de rotulagem e predições de sequências, como em reconhecimento de manuscrito, modelagem de linguagem e rotulagem fonética de quadros acústicos (Sak et al., 2014). No entanto, em uma RNN, não é possível interpretar o significado de atividade em um único neurônio isoladamente. Em vez disso, o significado da atividade em qualquer neurônio depende das atividades de outros neurônios (Yu & Deng, 2015).

A representação interna de características da fala é formada discriminativamente alimentando as características acústicas entre a camada oculta juntamente com as

características ocultas recorrentes de histórias passadas. A RNN possui conexões entre suas unidades, formando um ciclo direcionado. Esse ciclo é associado com um atraso de tempo que origina a estrutura de memória, expressada como estado interno. Ou seja, o ciclo recorrente associado com um atraso de tempo permite criar a estrutura para a memória da rede. Uma RNN cria novas camadas na rede de acordo com o comprimento da fala de entrada (Yu & Deng, 2015).

Uma RNN é obtida a partir de uma ANN *Feedforward*, em que a saída de um neurônio é conectada a suas entradas (Siddique & Adeli, 2013), conforme ilustrado na Figura 4.4. Essa rede possui uma camada oculta e profundidade, p_n, em que dada as entradas, X_m, o valor é conectado a todos os demais neurônios, N_n, com diferentes pesos, $W_{n,m}$, onde m = 1, 2, ... m e n = 1, 2, ..., n. Antes de gerar a saída Y_n a rede é realimentada, representada pelo símbolo, Δ, que simboliza o atraso de tempo da rede. O atraso de tempo possui significado simbólico, referindo-se a uma analogia ao período refratário de um modelo de neurônio biológico elementar (Siddique & Adeli, 2013), no qual uma célula não responde a algum estímulo ou responderá apenas se o estímulo for consideravelmente forte (Fuller et al., 2014).

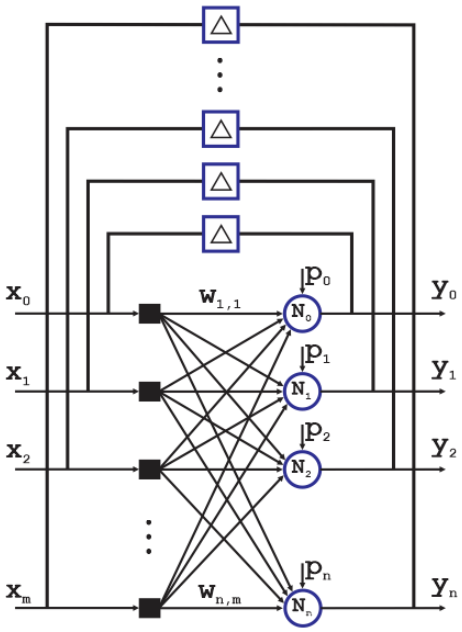

Figura 4.4: Rede Neural Recorrente com uma camada.
Fonte: Adaptado de Siddique & Adeli (2013).

Na Figura 4.5 é ilustrado o diagrama de um SRAF que utiliza RNN para encontrar a melhor sequência de fonemas. Na imagem, as palavras sublinhadas representam os fonemas. A palavra e a sentença em itálico foram selecionadas pelo sistema como sendo a melhor hipótese encontrada no Léxico e no ML, respectivamente.

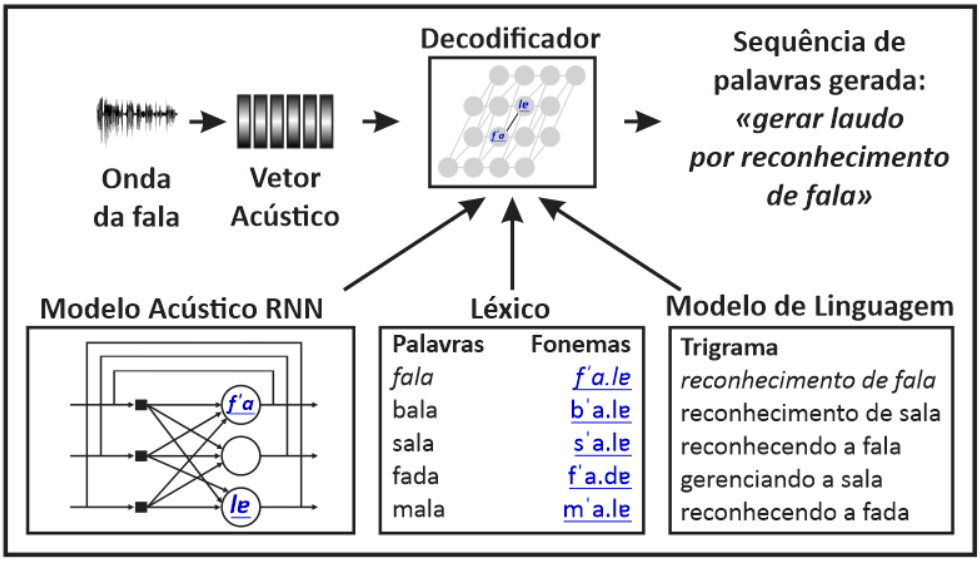

Figura 4.5: Diagrama da arquitetura de um Sistema de Reconhecimento Automático de Fala que utiliza Rede Neural Recorrente – *Recurrent Neural Network* (RNN).
Fonte: Adaptado de Siddique & Adeli (2013) e Davis & Scharenborg (2017).

Conforme ilustrado na Figura 4.5, a partir da entrada da fala, é realizado o seu tratamento para gerar o vetor acústico composto por fonemas (Subseção 3.4.1), que são então utilizados para alimentar a RNN que escolherá os neurônios cujos fonemas sejam compatíveis com os fonemas de entrada. A saída da RNN será composta pela sequência de fonemas que será enviada ao Decodificador. O Decodificador irá pesquisar as sequências de fonemas de máxima probabilidade utilizando as informações recebidas da RNN, do Léxico e do Modelo de Linguagem para gerar a melhor sequência de palavras.

Em uma RNN é criada uma nova camada oculta para cada entrada (Graves, 2012; Yu & Deng, 2015). No entanto, uma arquitetura padrão de RNN tem uma capacidade limitada de acesso às primeiras camadas de informações da rede. Essa dificuldade de acesso ocorre devido à fuga de gradiente (Graves, 2012), conforme exemplificado na Figura 4.6.

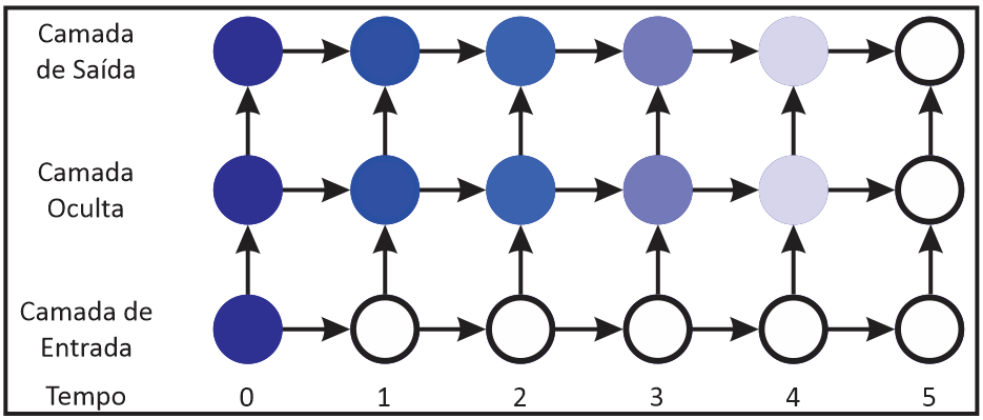

Figura 4.6: Fuga de gradiente de uma Rede Neural Recorrente.
Fonte: Adaptado de Graves (2012).

Na Figura 4.6, as variações entre os círculos escuros para os círculos claros indicam a sensibilidade variando no tempo. A sensibilidade do neurônio representa a sua capacidade de armazenar informações na rede. Com isso, é possível perceber que, conforme o tempo avança, a sensibilidade diminui, sendo representada pelos círculos claros. Isso ocorre à medida que novas entradas substituem as ativações da camada oculta anterior, e assim, as primeiras entradas da rede são "esquecidas".

Para atenuar o problema de fuga de gradiente, surgiu um novo tipo de rede, baseada em blocos de Memória Longa de Curto Prazo – *Long Short-Term Memory* (LSTM). Os blocos LSTM são responsáveis por armazenar e acessar informações durante longos períodos de tempo. Uma rede LSTM é semelhante a uma RNN padrão, exceto que as unidades de soma na camada oculta são substituídas por blocos de memória (Graves, 2012).

Rede Neural Recorrente Baseada em Memória Longa de Curto Prazo

Uma RNN-LSTM é uma arquitetura RNN que possui blocos LSTM. Um bloco LSTM pode ser considerado como uma unidade de rede complexa e inteligente, pois cada um deles é capaz de lembrar informações referentes a um longo período de tempo. A capacidade que um bloco de memória possui para lembrar informações é realizada pela estrutura associada que determina quando a entrada é significativa o suficiente para lembrar, quando deveria continuar a lembrar ou esquecer as informações, e quando deve produzir a saída de informação (Yu & Deng, 2015).

Cada bloco de memória LSTM contém um ou mais blocos de memória autoligadas, e três unidades multiplicativas, representada por: portão de entrada, portão de esquecimento e portão de saída. Analogamente, essas unidades podem representar respectivamente as operações de gravação (entrada), reinicialização (esquecimento) e leitura (saída). As unidades multiplicativas permitem que os blocos de memória LSTM armazenem e acessem informações durante longos períodos de tempo. Por exemplo, enquanto o portão de entrada permanece fechado, a ativação do bloco de memória não será substituída pelas novas entradas que cheguem à rede. Com isso, a informação contida nesse bloco pode ser disponibilizada à rede posteriormente. Essa informação é disponibilizada à rede ao abrir o portão de saída (Graves, 2012).

As unidades multiplicativas de um bloco LSTM permitem o armazenamento e o acesso às informações durante um longo período de tempo, reduzindo assim, o problema de fuga de gradiente. Por exemplo, enquanto o portão de entrada permanece fechado, possuindo ativação perto de 0, a informação contida no bloco de memória não será substituída por novas entradas, permanecendo disponível para utilização posterior pela rede, e em seguida, o portão de saída é aberto. Essa preservação das informações de dados, ao longo do tempo em uma rede LSTM, é ilustrada na Figura 4.7:

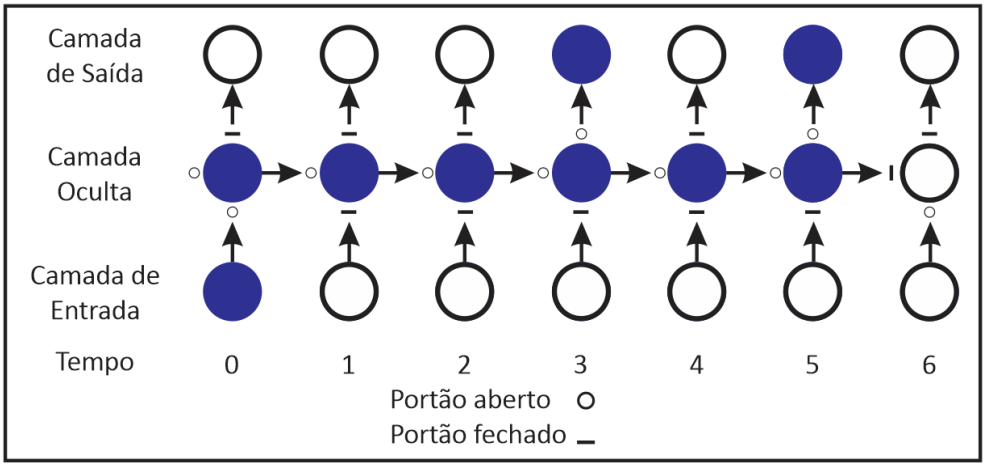

Figura 4.7: Preservação das informações do gradiente por Memória Longa de Curto Prazo.
Fonte: Traduzido de Graves (2012).

Na Figura 4.7 é ilustrada a preservação de informações de dados em uma rede LSTM. Os neurônios de cores escuras indicam máxima sensibilidade, enquanto os neurônios claros indicam ausência de sensibilidade. Os círculos escuros são neurônios que foram ativados, pois o seu portão estava aberto. Os círculos claros são referentes às células inativas. A sensibilidade da camada de saída pode ser ligada e desligada pelo portão de saída sem afetar o neurônio.

4.4 Considerações Finais

Conforme apresentado, o HMM é uma técnica amplamente utilizada em SRAFs. No entanto, as Redes Neurais Artificiais, com sua variação RNN, são utilizadas em SRAFs modernos pelas principais empresas de tecnologia da atualidade.

O HMM identifica melhores sequências de fonemas, comparando o sinal acústico de entrada com os dados de áudio do Modelo Acústico. Isso é realizado ao assumir valores discretos a variáveis aleatórias, tornando o HMM uma função probabilística de uma cadeia de Markov. Já as ANNs são alimentadas por fonemas que são encaminhados para os neurônios da rede que tenham sidos treinados com fonemas correspondentes aos de entrada.

Para que as Redes Neurais Artificiais possam operar maiores quantidades de dados de entrada com maior precisão, foram desenvolvidas células de memórias LSTM que as permitem armazenar informações ao longo das camadas mais profundas da rede.

Por fim, as principais tecnologias e algoritmos utilizados em SRAFs foram descritas nos Capítulos 3 e 4. No capítulo seguinte são descritos os materiais e os métodos utilizados para o desenvolvimento deste trabalho.

Capítulo 5

Materiais e Métodos

Com o intuito de orientar a condução deste trabalho seguindo um conjunto de métodos e técnicas, foram considerados alguns mecanismos como a definição de um protocolo para a realização da revisão sistemática, além de métodos para a coleta e tratamento dos arquivos, de áudios para a avaliação dos SRAFs, as tecnologias e os métodos empregados no desenvolvimento do Protótipo de Sistema *Web*.

Nesse sentido, o protocolo da revisão sistemática permitiu a seleção de trabalhos por meio de critérios bem definidos. Após a seleção dos trabalhos, foi proposta uma maneira para coletar e tratar os arquivos de áudio para serem utilizados na avaliação dos SRAFs. Para o desenvolvimento do Protótipo de Sistema *Web* foi empregado o método de Engenharia de *Software* Entrega em Estágio juntamente com a técnica Modelo-Visão-Controlador (MVC) para modelar a arquitetura do sistema.

Neste capítulo, o protocolo da revisão sistemática é detalhado na Seção 5.1. Na Seção 5.2 é apresentada a maneira como foi realizada a coleta e o tratamento dos arquivos de áudio. A avaliação dos SRAFs é descrita na Seção 5.3. As tecnologias, as ferramentas, os métodos e as etapas necessárias para o desenvolvimento do Protótipo de Sistema *Web* são abordados na Seção 5.4.

5.1 Protocolo da Revisão Sistemática

A Revisão Sistemática (RS) consiste em um método com a finalidade de identificar, avaliar e interpretar pesquisas relevantes disponíveis para uma determinada questão de pesquisa. Contudo, antes de realizar uma RS é necessário garantir a sua real necessidade, identificar e analisar quaisquer revisões sistemáticas existentes para o fenômeno de interesse (Kitchenham & Charters, 2007).

O estudo de viabilidade pode ser conduzido conforme ilustrado no fluxograma da Figura 5.1. Após constatada a necessidade de se realizar a RS, é possível utilizar um protocolo aplicado conforme o fluxograma da Figura 5.2 com o intuito de sintetizar as evidências relevantes no contexto de SRAFs e encontrar as ferramentas de estado-da-arte a serem acopladas ao Protótipo de Sistema *Web* a ser desenvolvido nesse trabalho.

No protocolo adotado neste trabalho foram definidas três etapas para a condução da RS: planejamento, execução e descrição dos resultados (Kitchenham & Charters, 2007).

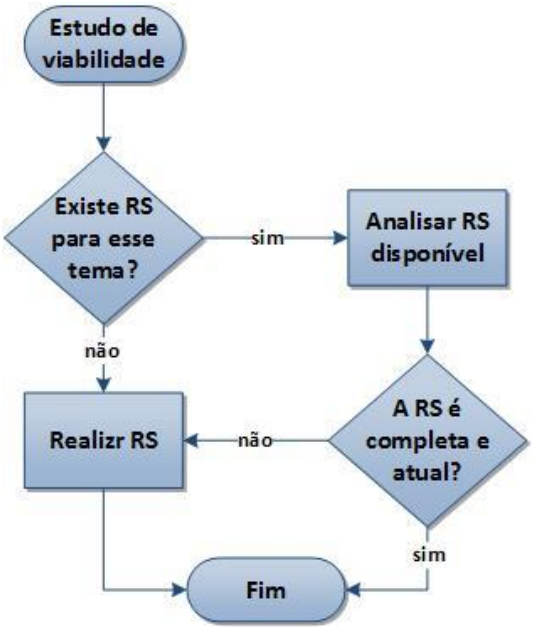

Figura 5.1: Fluxograma do estudo de viabilidade para execução da Revisão Sistemática.
Fonte: Baseado em Kitchenham & Charters (2007).

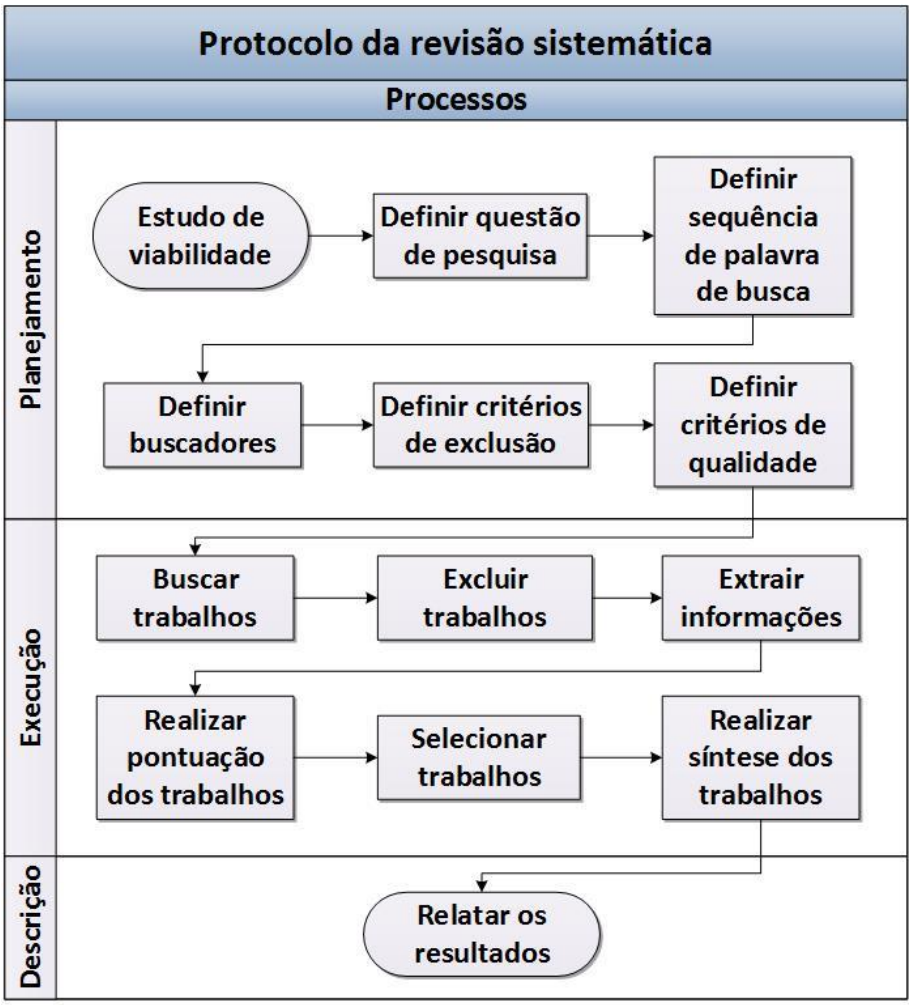

Figura 5.2: Fluxograma da Revisão Sistemática.
Fonte: Baseado em Kitchenham & Charters (2007).

A etapa de planejamento tem como objetivo organizar e direcionar a condução da RS. Nessa fase são definidas as estratégias de busca e os locais de busca são utilizados para a seleção dos trabalhos, bem como, os critérios de exclusão e de qualidade que são aplicados aos trabalhos.

Na fase de execução, são descartados os trabalhos não relevantes baseados nos critérios de exclusão. Os trabalhos selecionados são avaliados pelos critérios de qualidade, e na sequência, sintetizados. Na fase de descrição, os resultados são relatados. As três etapas da RS adotadas neste trabalho são descritas a seguir.

1. **Planejamento:**
   - **Definir questão de pesquisa:** formular e responder à questão de pesquisa que motivou a realização da RS;
   - **Definir estratégia de pesquisa:** criar a sequência de palavra de busca para ser reaplicada em todos os buscadores e modo de sua aplicação;
   - **Definir buscadores:** selecionar repositórios de publicações para realizar a busca dos trabalhos utilizando a sequência de palavra de busca;
   - **Definir critérios de exclusão:** criar critérios para excluir os trabalhos considerados não relevantes;
   - **Definir critérios de qualidade:** criar critérios para avaliar a qualidade dos trabalhos.
2. **Execução:**
   - **Buscar trabalhos:** realizar a busca dos trabalhos nos repositórios de artigos utilizando a sequência de palavra de busca;
   - **Excluir trabalhos:** realizar a exclusão de trabalhos que atenderam algum critério de exclusão;
   - **Extrair informações:** realizar a extração de informações relevantes dos trabalhos que não foram excluídos;
   - **Realizar pontuação dos trabalhos:** atribuir pontuação para cada critério de qualidade satisfeito;
   - **Selecionar trabalhos:** selecionar os trabalhos mais relevantes para realizar a síntese;
   - **Realizar síntese dos trabalhos:** resumir a ideia central dos trabalhos selecionados.
3. **Descrição:** analisar os resultados e discuti-los.

Neste trabalho a realização da RS teve como objetivo responder à seguinte questão de pesquisa: “quais sistemas computacionais ou *software* de reconhecimento automático de fala estão sendo utilizados em trabalhos rescentes ?”

Com relação à estratégia de pesquisa, foi definida a seguinte sequência de palavra de busca: ((“*automatic speech recognition*”) *AND* (“*open source*” *OR application OR system OR software OR program OR tool*)). A escolha de “*automatic speech recognition*” é baseada no termo utilizado para identificar trabalhos relacionados a SRAFs (Yu & Deng, 2015). Como o intuito foi a busca por sistemas computacionais, além de “*automatic speech recognition*”, o trabalho também deve conter pelo menos uma das seguintes variações: “*open source*” ou “*application*” ou “*system*” ou “*software*” ou “*program*” ou “*tool*”.

Essa sequência de palavra foi utilizada para realizar pesquisa nos seguintes repositórios: ACM *Digital Library*, CAPES, CiteSeerX, IEEE Xplorer, PubMed, Scopus, *Web of Science* e Wiley, entre os anos de 2011 até 2017. A primeira pesquisa foi realizada em setembro de 2015 compreendendo todas as bases de busca. A segunda pesquisa foi realizada em junho de 2017 compreendendo as bases de busca da ACM *Digital Library*, IEEE Xplorer e PubMed até o ano de 2016. A terceira pesquisa foi realizada em agosto de 2017, compreendendo os trabalhos complementares de 2015 e de 2016 para as demais bases de busca, com exceção do CiteSeerX (ver Seção 7.1) e a quarta pesquisa foi realizada em setembro de 2017 compreendendo os trabalhos até essa data para todas bases de busca, também com exceção do CiteSeerX.

A RS realizada neste trabalho, baseada no protocolo da Figura 5.2, foi dividida em quatro fases, conforme pode ser verificado na Figura 5.3. As quatro fases são utilizadas para selecionar os trabalhos mais relevantes.

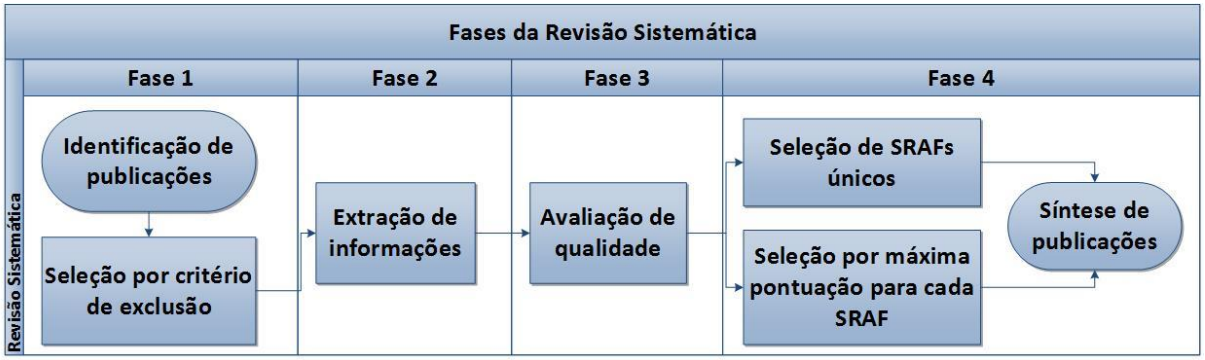

Figura 5.3: Fluxograma com as quatro fases da Revisão Sistemática.

Na primeira fase da RS, os trabalhos são identificados, excluindo os que atendam a pelo menos um dos seguintes critérios:

- Título duplicado, isto é, duas ou mais publicações com o mesmo título;
- Trabalho que trata exclusivamente de Modelo de Linguagem, Modelo Acústico, Léxico, *corpus* ou extração de características;
- Trabalho relacionado ao reconhecimento de voz, *speaker* (*identification* ou *dependent* ou *recognition* ou *verification* ou *adaptative* ou *diarization*), *text-to-speech*;
- Trabalhos que tratam de *audio-visual automatic speech recognition*, *visual speech recognition*, leitura labial e leitura gestual;
- Patentes, incluindo patentes de domínio público;
- O nome de *Workshops* ou de Conferências;
- Trabalho que propõe métodos ou técnicas para o melhoramento de *hardware* para o reconhecimento de fala, por exemplo, para melhorar o desempenho de processadores, unidade de processamento gráfico, entre outros;
- Trabalhos que criam modelos ou propõe melhoria para SRAF, porém, não contribuem com a construção de um sistema;

- Trabalho que não atende a questão de pesquisa, por exemplo, trabalhos que tratam de implantes cocleares, que são dispositivos eletrônicos que simulam sensações auditivas, semelhantes ao sistema fisiológico humano;
- Trabalho cujo acesso institucional não foi liberado ou publicações que não são atribuídas a um texto completo acessível a partir do acesso institucional;
- Trabalho escrito em línguas diferentes do Português, do Espanhol ou do Inglês;
- Publicação que utiliza uma língua de reconhecimento para o SRAF diferente da Língua Itálica (Espanhol, Francês, Italiano, Português e Romeno) e Língua Germânica (Alemão, Dinamarquês, Escocês, Frísio, Holandês, Inglês, Islandês e Sueco) (Beekes, 2011).

Para os trabalhos selecionados, ou seja, os que não são excluídos na fase anterior, na segunda fase são extraídos as seguintes informações:

- País em que a pesquisa foi desenvolvida e o idioma de treinamento do SRAF;
- Tecnologia do sistema, como nome do SRAF ou decodificador, modelagem acústica, extração de características e *corpus*;
- Medida de avaliação de desempenho, por exemplo, WER;
- Reconhecimento contínuo, ou seja, o SRAF reconhece fala contínua e não apenas palavras-chave, palavras isoladas ou dígitos. Caso o reconhecimento seja contínuo, o campo é marcado com "não".

Na terceira fase é avaliado o nível de qualidade dos trabalhos de acordo com os seguintes critérios:

- Idioma de reconhecimento do SRAF;
- Apresenta discussão de resultados?;
- Apresenta taxa de avaliação de precisão do reconhecimento?;
- Realiza o reconhecimento de fala contínua?;
- Permite ser treinado com outros *corpora*?;
- Propõe técnica para melhorar a tecnologia do SRAF?;
- A pontuação de cada publicação foi aplicada da seguinte maneira:
    - Três pontos: publicação que utilizaram um SRAF para a Língua Portuguesa;
    - Dois pontos: publicação que contém um SRAF para a Língua Espanhola ou para a Língua Inglesa;
    - Um ponto para cada um dos seguintes requisitos satisfeitos: (a) publicação que utiliza outro idioma de reconhecimento do SRAF, (b) artigo que apresenta discussão sobre os resultados, (c) apresenta medida de avaliação de precisão, (d) realiza o reconhecimento de fala contínuo, (e) permite ser treinado com outros *corpora*, e (f) propõe técnica para melhorar a tecnologia do SRAF.

Ao final, na quarta fase são selecionados os trabalhos com base em SRAF únicos, ou seja, sistemas que apareceram em uma única publicação e em critérios de qualidade selecionando todos os trabalhos para cada SRAF com máxima pontuação.

O critério de SRAF único foi utilizado para que não houvesse a possibilidade de ignorar um sistema, mesmo que não tenha atingido pontuação elevada. Com relação aos sistemas que são utilizados em duas ou mais publicações, são selecionadas as publicações que obtém as pontuações mais altas.

5.2 Coleta e Tratamento dos Arquivos de Áudio

Para a avaliação dos SRAFs neste trabalho foram utilizados arquivos de áudio coletados especificamente para este fim, com cada arquivo contendo a gravação da leitura de um texto de referência de 617 palavras (Varella, 2011) (Apêndice A), cujo conteúdo é relacionado com a área médica.

A coleta desses áudios foi realizada por 30 voluntários que colaboraram com a leitura do texto de referência em velocidade normal, ou seja, com a sua leitura habitual. As gravações foram realizadas individualmente em sete salas: sala 1 (15 voluntários), sala 2 (três voluntários), sala 3 (seis voluntários), sala 4 (três voluntários), sala 5 (um voluntário), sala 6 (um voluntário) e sala 7 (um voluntário). No momento da leitura do texto, a gravação do áudio foi coletada com o *software Voice Recorder* 10.1611.3051.0 presente no sistema operacional Windows 10.

Esse grupo de 30 voluntários foi composto por 15 homens e 15 mulheres, com idades entre 19 e 50 anos. A média de idade foi de 26,7 anos e desvio padrão de 7,60, sendo todos moradores da cidade de Foz do Iguaçu, Paraná.

As características dos voluntários são apresentadas na Tabela 5.1, na qual são descritas uma identificação genérica, a idade, a sala e a formação de cada um dos voluntários. As linhas em itálico compreendem os voluntários que participaram do experimento preliminar (Seção 7.2).

Tabela 5.1: Características dos voluntários e sala de realização das gravações (Masc. = Masculino e Fem. = Feminino).

Voluntários	**Idade**	**Sala**	**Formação**
Masc. 1	*23*	*3*	*Engenheiro Eletricista*
Masc. 2	*23*	*6*	*Engenheiro Mecânico*
Masc. 3	*21*	*1*	*Estudante de Ciência da Computação*
Masc. 4	*24*	*7*	*Tecnólogo em Análise e Desenvolvimento de Sistemas*
Masc. 5	*39*	*2*	*Cientista da Computação*
Fem. 1	*20*	*2*	*Estudante de Engenharia Elétrica*
Fem. 2	*25*	*3*	*Matemática*
Fem. 3	*22*	*1*	*Estudante de Engenharia Elétrica*
Fem. 4	*22*	*3*	*Estudante de Matemática*
Fem. 5	*43*	*2*	*Cientista da Computação*

(Continuação)

Voluntários	Idade	Sala	Formação
Masc. 6	22	1	Cientista da Computação
Masc. 7	21	1	Engenheiro Eletricista
Masc. 8	33	3	Engenheiro Eletricista
Masc. 9	20	3	Estudante de Engenharia Elétrica
Masc. 10	23	5	Estudante de Engenharia Química
Masc. 11	25	1	Estudante de Engenharia Elétrica
Masc. 12	22	1	Engenheiro Mecânico
Masc. 13	25	1	Estudante de Ciência da Computação
Masc. 14	30	1	Cientista da Computação
Masc. 15	20	3	Estudante de Ciência da Computação
Fem. 6	29	1	Bacharela em Direito
Fem. 7	50	4	Dentista
Fem. 8	33	1	Turismóloga
Fem. 9	22	1	Estudante de Ciência da Computação
Fem. 10	28	4	Enfermeira
Fem. 11	24	1	Engenheira Eletricista
Fem. 12	23	1	Engenheira Eletricista
Fem. 13	19	1	Estudante de Enfermagem
Fem. 14	31	1	Estudante de Enfermagem
Fem. 15	39	4	Turismóloga

O tratamento dos arquivos de áudio consistiu em converter o formato e as características de digitalização sonora, de acordo com a exigência de cada SRAF avaliado. A digitalização sonora tem como finalidade representar o sinal analógico de maneira digital, podendo ser utilizado a Modulação de código de pulso – *Pulse-code modulation* (PCM). Esse processo envolve a taxa de amostragem e a profundidade de bit. A taxa de amostragem se refere à quantidade de amostras coletadas de um sinal analógico para converter o arquivo de áudio em formato digital, ou seja, a quantidade de vezes em que é medida a amplitude da onda. Já a profundidade se refere ao número de bits contido em cada amostra.

Para a conversão entre os diferentes tipos de extensão dos arquivos de áudio, foi utilizado o *software* FormatFactory[10] 3.6.0. Os formatos de áudio utilizados neste trabalho foram: (a) formato de áudio de forma de onda – *WAVEform audio format* (WAV) – e (b) Codec livre de áudio sem perdas – *Free Lossless Audio Codec* (FLAC). O WAV é um formato-padrão para armazenar arquivos de áudio da Microsoft e da IBM que utiliza PCM para a digitalização sonora (IBM-Microsoft, 1991). O formato FLAC também utiliza o PCM (FLAC, 2014).

Após a conversão de formato, o *software* Audacity[11] 2.1.3 foi utilizado para dividir os arquivos de áudio em partes menores, devido à limitação de alguns SRAFs de não permitir a transcrição da fala utilizando arquivos de áudio com maior tempo de duração. Além disso, esse *software* permite a gravação de áudio ao vivo por meio de um microfone, edição de áudio

[10] http://www.pcfreetime.com/formatfactory/index.php?language=pt
[11] http://www.audacityteam.org/

como copiar, colar, recortar, deletar e adicionar efeitos, como tons, silêncio, ruído, instrumentos ou faixas de ritmo, entre outras funcionalidades.

5.3 Avaliação dos Sistemas de Reconhecimento Automático de Fala

A avaliação dos SRAFs é realizada em uma avaliação preliminar e uma avaliação final. A avaliação preliminar contém todos os SRAFs selecionados na RS com foco na Língua Portuguesa do Brasil e a avaliação final compreendem os SRAFs selecionados a partir da avaliação preliminar.

Para avaliar o desempenho dos SRAFs foi utilizada a métrica de taxa de erro de palavra (Seção 3.6). Essa métrica mensura a taxa de erro em que quanto menor for essa taxa, expressa em porcentagem, mais preciso é o sistema.

O arquivo de áudio em avaliação é então submetido a cada um dos SRAFs. Os textos gerados pelos sistemas são denominados de hipótese, sendo comparados com o texto de referência durante a avaliação.

Os arquivos de hipóteses, gerados pelos SRAFs, a partir das gravações de áudio, passaram por uma padronização das seguintes palavras:

- *1* ou *um* precedido por *cm* ou *centímetro* para *1 cm*;
- *à* ou *há* para *a*;
- *1os.* ou *1º* para *primeiro*;
- *pré-maligna* para *pré maligna*;
- *d n a* para *dna*;
- *trinta* para *30*;
- *mil* ou *1.000* para *1000*;
- *1* para *um*;
- *2* para *dois*;
- *cinco* para *5*;
- *dez* para *10*;
- *cinquenta* para *50*;
- *cinquenta e cinco* para *55*;
- *cinquenta e seis vírgula sete* para *56,7*;
- *dois mil quatrocentos e trinta e seis* ou *duas mil quatrocentos e trinta e seis* para *2436*;
- *dezesseis* para *16*;
- *por cento* para *%*;
- *centímetro* para *cm*;
- *um vírgula cinco* para *1,5*.

Com relação aos números com dois ou mais dígitos gerados por extenso também foram padronizados para o seu equivalente numeral, por exemplo, "*cinquenta e seis* para *56*". Por questão de padronizar são retiradas as pontuações e deixada apenas a primeira letra em maiúscula para todos os arquivos de hipóteses.

Os SRAFs para a Língua Portuguesa do Brasil considerados para os experimentos preliminares são: *Web Speech Application Programming Interface* (API)[12]; Bing *Speech* API[13]; IBM *Speech to Text*[14]; Audimus[15]; duas versões do Voxsigma *Speech to Text*[16], sendo uma versão estável e uma versão beta; e Coruja versão 0.2[17].

O SRAF *Web Speech* API foi desenvolvido pela Google em linguagem de programação JavaScript e sua especificação foi definida pelo grupo *Speech* API *Communit Group*[18]. Esse grupo comunitário tem por objetivo desenvolver possíveis padronizações futuras para a *Internet*, pertencentes ao *World Wide Web Consortium*[19] (W3C), que é uma organização com o objetivo de definir padrões para o desenvolvimento e a interpretação de conteúdos utilizados na *Internet* (Shires & Wennborg, 2014).

O Bing *Speech* API foi desenvolvido pela Microsoft, o qual permite a transcrição de áudios longos de até dez minutos de duração com os seguintes modos de reconhecimento: interativo, ditado e conversacional. O modo interativo permite a transcrição de áudio de até 15 segundos e oferece transcrições parciais de reconhecimento. O modo ditado permite a transcrição de áudio longo, porém, até o momento da escrita desta dissertação, esse modo não retorna resultados parciais de reconhecimento, retornando apenas o resultado final após o término da fala. E no modo conversacional, a API se adequa para reconhecer a fala informal (Microsoft, 2017b).

O IBM *Speech to Text* conta com retornos de resultado alternativos e intermediários de reconhecimento. O primeiro fornece diferentes alternativas de hipóteses para o reconhecimento, e o resultado intermediário representa hipóteses intermediárias conforme o envio de áudio. Em ambos os casos, é retornado o resultado final representado pela transcrição de maior confiança (IBM, 2017).

O Audimus é um motor de reconhecimento de fala desenvolvido pela Voiceinteraction que permite desenvolver aplicações para os sistemas operacionais Windows. Também é possível realizar transcrição de fala de maneira *offline* (Voiceinteraction, 2017).

O Voxsigma *Speech to Text* é desenvolvido pela Vocapia que oferece um conjunto de *softwares* de transcrição de fala para texto, para sistemas operacionais Linux e também como um serviço *Web*. No momento do experimento preliminar, em maio de 2016, o Voxsigma contava com uma versão estável e uma versão beta. Uma das características do Voxsigma é a sua capacidade adaptativa que permite a transcrição da fala ruidosa, como a fala com música de fundo (Vocapia, 2017).

[12] https://dvcs.w3.org/hg/speech-api/raw-file/tip/webspeechapi.html
[13] https://docs.microsoft.com/pt-br/azure/cognitive-services/speech/home
[14] https://www.ibm.com/watson/developercloud/speech-to-text.html
[15] http://www.voiceinteraction.com.br/?page_id=423
[16] http://www.vocapia.com/voxsigma-speech-to-text.html
[17] http://www.laps.ufpa.br/falabrasil/downloads.php
[18] https://www.w3.org/community/speech-api
[19] http://www.w3c.br/Home/WebHome

Por fim, o Coruja é um *software* de código aberto de reconhecimento de fala específico para a Língua Portuguesa do Brasil desenvolvido pela Universidade Federal do Pará. O Coruja é compatível para ser utilizado no desenvolvimento de aplicações destinadas aos sistemas operacionais Windows e Linux e também para desenvolver aplicações em linguagem de programação Java (Silva, 2010).

Para os testes dos SRAFs foi necessário ajustar os arquivos de áudio de acordo com as especificidades de cada sistema. Os ajustes compreendem a conversão do formato, taxa de amostragem, amplitude da onda e tempo de duração de cada arquivo. As características dos arquivos de áudio esperados por cada SRAF podem ser verificadas na Tabela 5.2.

Tabela 5.2: Características dos arquivos de áudio para os Sistemas de Reconhecimento Automático de Fala.

SRAF	**Formato do áudio**	**Digitalização sonora**	**Taxa de amos-tragem (Hz)**	**Amplitude (bits)**	**Canal**
Web Speech API	FLAC	PCM	44.100	16	Mono
Bing *Speech* API	WAV	PCM	44.100	16	Mono
IBM *Speech to Text*	WAV	PCM	44.100	16	Mono
Audimus	WAV	PCM	16.000	16	Mono
Voxsigma *Speech to Text*	WAV	PCM	16.000	32	Mono
Voxsigma *Speech to Text* beta	WAV	PCM	16.000	32	Mono
Coruja	WAV	PCM	22.050	16	Mono

As avaliações dos SRAFs realizados neste trabalho compreendem os experimentos preliminar e final. Para o experimento preliminar são selecionados dez voluntários para avaliar os sete SRAFs, cujas transcrições foram obtidas em maio de 2016.

Para o experimento preliminar, os testes com os SRAFs da Audimus e as duas versões do Voxsigma foram realizadas pelas respectivas empresas. Para isso, foram enviados os arquivos de áudio para as respectivas empresas e na sequência, tanto a Audimus quanto a Voxsigma retornaram os resultados das transcrições. Os demais SRAFs foram avaliados com arquivos de áudio de acordo com a capacidade de transcrição máxima de cada sistema (Tabela 5.3). Isso foi realizado levando em consideração um tamanho que não resultasse em travamentos durante a transcrição dos arquivos de áudio.

Tabela 5.3: Tamanho dos arquivos de áudio dos Sistemas de Reconhecimento Automático de Fala avaliados no experimento preliminar (s. = segundos).

***Web Speech* API**	**Bing *Speech* API**	**IBM *Speech to Text***	**Coruja**
Até 12 s.	Aproximadamente 30 s.	Arquivo completo	Aproximadamente 30 s.

Como o *Web Speech* API permite a transcrição de áudio de no máximo 12 segundos, no experimento final (Seção 7.3), todos os arquivos de áudio do Bing *Speech* API foram divididos também em no máximo 12 segundos. Dessa maneira, as transcrições para o experimento final foram realizadas em abril de 2017 para todos os arquivos de áudio do Bing *Speech* API e para os demais 20 arquivos de áudio do *Web Speech* API.

Para os testes do Microsoft Bing *Speech* API, todas as partes dos arquivos de áudio foram importadas para o projeto de teste disponibilizado pela empresa[20]. Essa aplicação é baseada em C#, que é uma linguagem de programação orientada a objetos para criar aplicativos executados em .NET *Framework*[21]. O .NET é um modelo de programação para desenvolver aplicativos no sistema operacional Windows para celulares, *Web* e *desktops* (Microsoft, 2017c).

O projeto de teste foi executado no Visual Studio[22] *Community* 2017 (Figura 5.4), que é um ambiente de desenvolvimento utilizado para a criação de aplicativos nos sistemas operacionais Android, iOS e Windows e aplicações para a *Web* e serviços em nuvem (Microsoft, 2017d).

A melhor hipótese de reconhecimento é retornada quando não há mais hipóteses prévias (Figura 5.4), sendo formuladas em tempo real durante o envio de áudio. As hipóteses prévias se adequam com base na pronúncia de novas palavras. Essa adequação é baseada em *n*-grama (Subseção 3.4.2).

Figura 5.4: Resultado da transcrição do arquivo de áudio do Microsoft Bing *Speech* API.

O SRAF IBM *Speech to Text* é testado submetendo o arquivo de áudio em um *site* disponibilizado pela empresa[23], conforme é ilustrado na Figura 5.5.

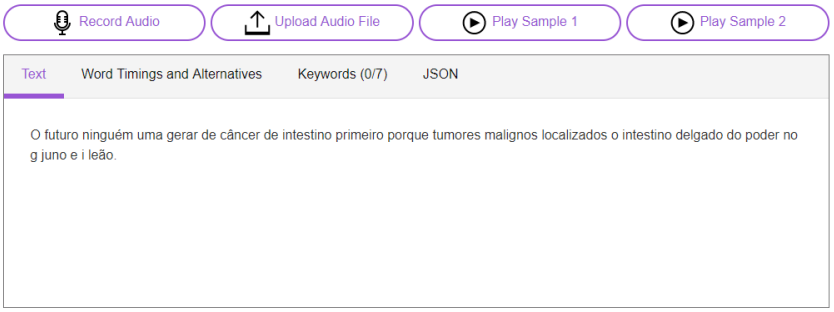

Figura 5.5: Resultado da transcrição do arquivo de áudio do IBM *Speech to Text*.

[20] https://github.com/Azure-Samples/Cognitive-Speech-STT-Windows
[21] https://www.microsoft.com/net/download/framework
[22] https://www.visualstudio.com/pt-br/vs/
[23] https://speech-to-text-demo.mybluemix.net/

O Coruja é testado com a API disponibilizada para o sistema operacional Linux[24] (Figura 5.6), instalada no Ubuntu 11.04 de 32 bits, seguindo o tutorial de Domingues (2013) para a sua configuração.

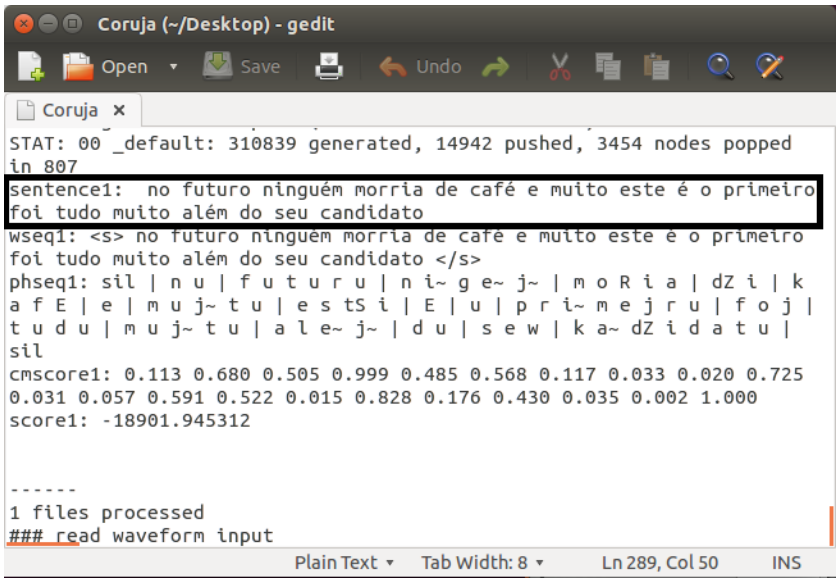

Figura 5.6: Resultado da transcrição do arquivo de áudio do Coruja.

Para o teste do Google *Web Speech* API, as partes dos arquivos de áudio foram importadas para um projeto do Visual Studio em C# (código-fonte no Apêndice B). Após a execução da aplicação, foi retornado o resultado da transcrição no console do Visual Studio, conforme ilustrado na Figura 5.7:

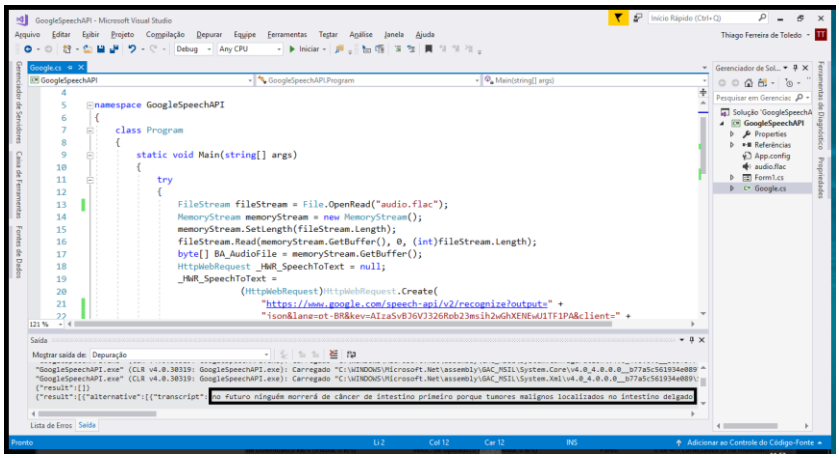

Figura 5.7: Resultado da transcrição do arquivo de áudio do Google *Web Speech* API.

Como mencionado, a taxa de precisão dos SRAFs foi medida utilizando a taxa de erro de palavra – *Word Error Rate* (WER). O cálculo foi realizado utilizando o *software* Sclite, um módulo do NIST *Scoring Toolkit* versão 0.1[25]. O Sclite realiza o cálculo da WER a partir do

[24] http://www.laps.ufpa.br/falabrasil/files/Coruja_Linux.rar
[25] https://www.nist.gov/itl/iad/mig/tools

alinhamento entre o texto de hipótese com o texto de referência. O primeiro passo para o alinhamento é realizar a segmentação de listas de palavras de referências e de hipóteses ao localizar áreas comuns, ou seja, o arquivo de hipótese é cortado em regiões correspondentes aos segmentos do arquivo de referência. Na sequência, o alinhamento é realizado utilizando a métrica de distância de palavras de Levenshtein (1966) para contabilizar, de maneira ponderada, o custo de palavras corretas, inseridas, deletadas e substituídas (ICSI, 2015).

Após a obtenção de todas as taxas WERs, os resultados foram avaliados utilizando métodos estatísticos para a análise dos dados. O *software* estatístico utilizado para avaliar os resultados foi o GraphPad InStat[26] versão 3.10.

Os testes estatísticos são realizados considerando o nível de significância de 0,05. Para verificar se os dados apresentam uma distribuição normal é aplicado o teste de normalidade de Kolmogorov-Smirnov (KS).

5.4 Desenvolvimento do Protótipo de Sistema *Web*

Nesta seção são apresentadas as tecnologias (Subseção 5.4.1), as ferramentas (Subseção 5.4.2), bem como: os equipamentos (Subseção 5.4.3) e na Subseção 5.4.4, os métodos utilizados para o desenvolvimento do Protótipo de Sistema *Web* (PSW).

5.4.1 Tecnologias

As tecnologias utilizadas para o desenvolvimento do PSW foram:

- **CSS[27] (*Cascading Style Sheets*[28]):** linguagem de folhas de estilo para adicionar estilos (fontes, espaçamento, entre outros) em documentos *Web* escritos em linguagem de marcação;
- **Hibernate[29]:** *framework* de código aberto – *open source* – para o mapeamento de objeto-relacional para a linguagem de programação Java, ou seja, transformação de classes em Java para tabelas do Banco de Dados (BD) (Alves, 2015);
- **Java[30]:** linguagem de programação interpretada e orientada a objetos (Winder & Graham, 2009);
- **JavaScript[31]:** linguagem de programação orientada a objetos baseada em *scripts* para serem integradas nas páginas *Web*;

[26] https://www.graphpad.com/scientific-software/instat/
[27] http://www.w3.org/Style/CSS/
[28] Traduzindo para a Língua Portuguesa do Brasil: folhas de estilo em cascata.
[29] http://hibernate.org/orm
[30] http://docs.oracle.com/javase/7/docs/technotes/guides/language/
[31] http://www.ecma-international.org/ecma-262/5.1/

- **JavaServer Faces (JSF) 2[32]:** especificação Java para construção de componentes para simplificar a criação de interfaces de usuários para o desenvolvimento *Web* em Java;
- **PrimeFaces[33]:** *framework* de código aberto de componentes para construção de interfaces de usuários para JSF;
- **Spring *Security*[34]:** *framework* de segurança para garantir a autenticação e a autorização dos usuários, permitindo acesso ao sistema apenas para usuários cadastrados;
- **XHTML [35] (*eXtensible Hypertext Markup Language* [36]):** reformulação da linguagem de marcação HTML (*HyperText Markup Language*), baseada em XML (*eXtensible Markup Language*[37]), que combina as *tags* de marcação HTML com as regras do XML.

5.4.2 Ferramentas

O PSW foi desenvolvido utilizando as seguintes ferramentas:

- **Apache Tomcat 8[38]:** servidor *Web* para hospedar aplicações desenvolvidas em linguagem de programação Java;
- **Eclipse Mars.2 (4.5.2)[39]:** ambiente de desenvolvimento para a linguagem de programação Java;
- **Maven[40]:** *software* de automação e gerenciamento para auxiliar a organização dos arquivos-fontes, gerenciando a dependência de bibliotecas e Jar ARchives (JARs) do projeto;
- **MySQL *Communuty Server* 5.7[41]:** BD para armazenar os registros do PSW.

5.4.3 Equipamentos

Os seguintes equipamentos foram utilizados para o desenvolvimento do PSW, execução dos experimentos e hospedagem do PSW:

- **Avaliação do experimento preliminar dos SRAFs e início do desenvolvimento do PSW:** *notebook* Lenovo G40-80 com Windows 10 *Home* de 64 bits, processador Intel Core i5-5200U de 2.2 Gigahertz (GHz), placa gráfica AMD

[32] http://www.oracle.com/technetwork/java/javaee/javaserverfaces-139869.html
[33] http://www.primefaces.org/
[34] https://projects.spring.io/spring-security/
[35] https://www.w3.org/TR/xhtml2/
[36] Traduzindo para a Língua Portuguesa do Brasil: linguagem de marcação de hipertexto extensível.
[37] Traduzindo para a Língua Portuguesa do Brasil: linguagem de marcação extensível.
[38] http://tomcat.apache.org/
[39] https://eclipse.org/mars/
[40] https://maven.apache.org/
[41] http://dev.mysql.com/downloads/mysql/

Radeon R5 M230 de 2 Gigabyte (GB) dedicada, armazenamento do disco rígido de 1 Terabyte (TB) e memória de 4 GB;

- **Avaliação do experimento final dos SRAFs e conclusão do desenvolvimento do PSW:** *notebook* Hewlett-Packard HQ-TER 71025 com Windows 10 *Home* de 64 bits, processador Intel Core i3-5005U de 2 GHz, armazenamento do disco rígido de 1 TB e memória de 8 GB;
- **Gravação da leitura do texto de referência:** fone de ouvido com microfone Multilaser modelo PH043 com sensibilidade de 119 dB, frequência de 20.000 Hz e impedância de 32 Ohms.

5.4.4 Método para o Desenvolvimento do Protótipo de Sistema *Web*

O método adotado para o desenvolvimento do PSW baseou-se na técnica Entrega em Estágio, presente em Engenharia de *Software*. O desenvolvimento do PSW foi realizado levando em consideração a redução da sobrecarga cognitiva do usuário, que ocorre quando a capacidade do cérebro para processar informações é excedida (Segbroeck et al., 2014, Zahabi et al., 2015). Com base nisso, as telas do PSW foram construídas com o menor número possível de elementos visuais.

Com relação ao método Entrega em Estágio, essa técnica contempla a entrega de uma nova funcionalidade, previamente planejada e definida, a cada final de ciclo, conforme ilustrado no diagrama da Figura 5.8.

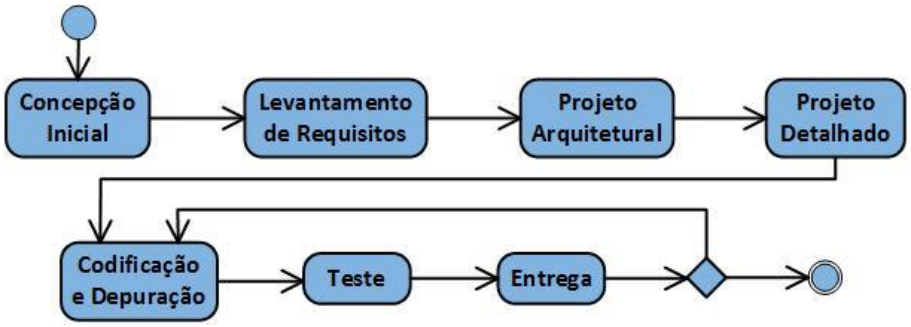

Figura 5.8: Diagrama do modelo de Entrega em Estágio.
Fonte: Adaptado de Wazlawick (2013).

Na primeira etapa, Concepção Inicial, foi definida a visão geral do sistema, a qual consiste em um documento em formato livre, em que foram descritos os aspectos relevantes do sistema. O Levantamento de Requisitos foi realizado para identificar as necessidades do sistema, ou seja, quais funções o sistema deveria executar e quais seriam as possíveis restrições que o sistema deve operar (Wazlawick, 2011; Wazlawick, 2013).

Para o entendimento geral dos requisitos do sistema foi utilizada a engenharia de requisitos, a qual é uma ação de engenharia de *software* para realizar a intermediação entre a atividade de comunicação entre as pessoas envolvidas no projeto e a modelagem do sistema.

Dessa maneira, é necessário adaptá-la de acordo com as necessidades do processo, do projeto, do produto e das pessoas envolvidas (Pressman & Maxim, 2016).

Nessa etapa, Levantamento de Requisitos, foram identificados os Requisitos Funcionais, ou seja, quais eram as funções a serem executadas pelo sistema. Os Requisitos Funcionais podem ser evidentes ou ocultos. Os requisitos evidentes são funcionalidades que são realizadas com o conhecimento do usuário, ao contrário dos requisitos ocultos (Wazlawick, 2011).

Os Requisitos Não-funcionais e Suplementares são requisitos ligados aos Requisitos Funcionais, podendo ser de dois tipos: lógicos ou tecnológicos. As restrições lógicas são as regras de negócio relacionadas às funcionalidades. Já as restrições tecnológicas se relacionam aos meios tecnológicos para realizar determinada funcionalidade do sistema. Os Requisitos Suplementares são restrições que se aplicam a todo o sistema e não exclusivamente a funções específicas. Wazlawick (2011) sugere alguns exemplos de Requisitos Não-funcionais e Suplementares, como:

- **Usabilidade:** qual a maneira de utilização do sistema por parte do usuário, contemplando fatores como ajuda, documentação e manual disponível para consulta;
- **Confiabilidade:** qual medida será implementada para tornar o sistema mais confiável, ou seja, qual tipo de tratamento de falhas o sistema conterá;
- **Desempenho:** qual o nível de eficiência e de precisão que o sistema deverá apresentar;
- **Configurabilidade:** quais as funções do sistema que o usuário poderá configurar;
- **Segurança:** quais são os grupos de usuários e quais funções cada grupo poderá executar;
- **Implementação:** quais são as tecnologias utilizadas, como bibliotecas, BD, linguagem de programação, dentre outras características referentes à tecnologia de desenvolvimento do sistema;
- **Interface:** quais são os elementos relacionados ao desenho das telas do sistema;
- **Empacotamento:** como o sistema será entregue ao usuário final;
- **Legais:** qual meio de assessoramento jurídico o sistema conterá para saber se algum direito autoral ou normas específicas da área para a qual o sistema está sendo desenvolvido está sendo infringida.

Os Requisitos Não-funcionais e Suplementares podem ser classificados como (Wazlawick, 2011):

- **Permanentes e ou Transitórios:** os requisitos considerados permanentes, não se alteram ao longo do ciclo de vida do sistema, ao contrário dos requisitos transitórios que podem sofrer alteração. Uma funcionalidade pode ser permanente e ao mesmo tempo transitória;
- **Obrigatórios ou Desejáveis:** os requisitos obrigatórios devem ser desenvolvidos de qualquer maneira, ao contrário dos requisitos desejáveis, que no caso de

dificuldades que resultem em transtorno durante o desenvolvimento, podem não ser implementados no sistema naquele momento.

Para Pressman & Maxim (2016), o Projeto Arquitetural modela a estrutura global do sistema, na qual é representada a interface com os componentes, as dependências, as relações e as interações. Já para Wazlawick (2013), essa etapa é dividida em Projeto Arquitetural e Projeto Detalhado. O Projeto Arquitetural consiste em mapear o sistema para representar suas diferentes partes (Wazlawick, 2011; Wazlawick, 2013).

A etapa do Projeto Arquitetural pode ser modelada com o padrão de arquitetura Modelo-Visão-Controlador (MVC)[42]. Esse padrão consiste na separação de responsabilidades com relação à gravação de informações no BD (Modelo). A Visão é responsável por exibir os dados contidos no BD e outros elementos. A intermediação entre a Visão e o Modelo é realizada pelo Controlador. Por exemplo, o Controlador pode gerenciar a atualização do estado do Modelo ou enviar um comando para a Visão alterar a apresentação de exibição (Pressman & Maxim, 2016).

Na Figura 5.9, é ilustrado o diagrama da arquitetura MVC, em que o usuário faz uma requisição de alguma informação. Essa informação é gerenciada pelo Controlador que será responsável em buscar a informação no Modelo (BD da aplicação) e repassar para a Visão para que o usuário possa visibilizar a informação solicitada (resposta da informação solicitada ao usuário).

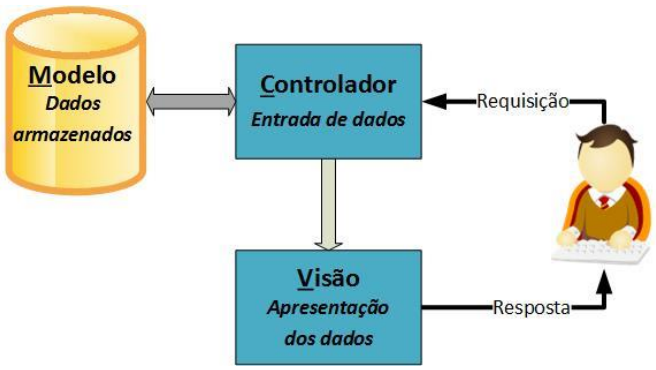

Figura 5.9: Diagrama da arquitetura Modelo-Visão-Controlador (MVC).
Fonte: Adaptado de Alves (2015).

Já o Projeto Detalhado auxilia na especificação da estrutura e do comportamento interno de cada parte do sistema, descrevendo-os em mais detalhes. Na sequência, com base nos requisitos definidos, o protótipo é desenvolvido, testado e entregue nas respectivas etapas: Codificação e Depuração, Teste e Entrega (Wazlawick, 2011; Wazlawick, 2013).

Para a descrição formal do PSW, foi utilizada a Linguagem de Modelagem Unificada – *Unified Modeling Language* (UML). A UML é uma linguagem que pode ser utilizada para descrever determinados cenários para apoiar o desenvolvimento de um *software* (Wazlawick, 2011; Pressman & Maxim, 2016). Alguns exemplos dos diagramas da UML incluem os diagramas estruturais (diagramas de pacote, classes, objetos, estrutura composta, componentes

[42] http://heim.ifi.uio.no/~trygver/themes/mvc/mvc-index.html

e distribuição), os diagramas comportamentais (diagramas de casos de uso, atividades e máquina de estados) e os diagramas de interação (diagramas de comunicação, sequência, tempo e visão geral de integração). No entanto, nem todos os diagramas precisam ser utilizados; usam-se os que, de fato, podem representar alguma informação importante para o desenvolvimento de um determinado *software* (Wazlawick, 2011). Os diagramas utilizados para apoiar o desenvolvimento do PSW foram os diagramas comportamentais, compreendendo os diagramas de caso de uso e de atividade.

Os casos de uso são diagramas que permitem ilustrar as principais atividades de negócio ligadas ao sistema. O caso de uso tem por objetivo ilustrar como ocorre a interação do sistema com o usuário, além de ilustrar quais consultas e transformações de informação são necessárias para completar os processos de interação. Em outras palavras, o caso de uso é uma abordagem para sistematizar e organizar os requisitos. Já o diagrama de atividade pode ser utilizado com o objetivo de representar processos em nível organizacional para que se possa obter uma visão geral de um processo referente a um Requisito Funcional (Wazlawick, 2011; Pressman & Maxim, 2016).

5.5 Considerações Finais

Neste capítulo foram apresentados os materiais e os métodos aplicados para o desenvolvimento deste trabalho.

A aplicação do protocolo da revisão sistemática possibilita encontrar, avaliar e sintetizar as informações de diversos trabalhos na área de reconhecimento automático de fala. Após a seleção de trabalhos relevantes, são selecionados os SRAFs a serem avaliados em um experimento preliminar com dez voluntários. Posteriormente, os SRAFs considerados mais adequados são avaliados com arquivos de áudio de 30 voluntários.

O desenvolvimento do PSW segue o método de Engenharia de *Software* denominado de Entrega em Estágio por permitir entregas parciais a cada módulo ou evolução realizada durante o seu desenvolvimento. Com relação à arquitetura do PSW, optou-se pelo padrão MVC, pois permite a separação de diferentes níveis de camada durante o seu desenvolvimento.

Tendo em vista os materiais e os métodos utilizados, no próximo capítulo é apresentado, a maneira como a técnica de Entrega em Estágio, presentes em Engenharia de *Software*, foi aplicada ao desenvolvimento do PSW.

Capítulo 6

Engenharia de *Software* e Desenvolvimento do Protótipo de Sistema *Web*

Para que um sistema possa ser desenvolvido de maneira eficaz e no menor tempo possível, é necessário o emprego de técnicas que guiem a equipe de programação na condução do seu desenvolvimento. Para isso, o desenvolvimento do PSW seguiu a técnica de Entrega em Estágio.

Cada uma das etapas dessa técnica é descrita neste capítulo, sendo elas: Concepção Inicial (Seção 6.1), Levantamento de Requisitos (Seção 6.2), Projeto Arquitetural (Seção 6.3), Projeto Detalhado (Seção 6.4), Codificação e Depuração (Seção 6.5), Teste e Entrega (Seção 6.6).

6.1 Concepção Inicial

Como mencionado, a técnica Entrega em Estágio foi utilizada para o desenvolvimento do PSW. Diante disso, no Quadro 6.1 é descrita a visão geral do PSW, representando a etapa de Concepção Inicial do projeto.

Quadro 6.1: Visão Geral do Protótipo de Sistema *Web*.

PSW para geração de laudos médicos por meio do reconhecimento automático de fala
Visão Geral do Sistema
O PSW deve gerar laudo médico de maneira automática utilizando tecnologia de reconhecimento automático de fala. O laudo médico é gerado da seguinte maneira: o especialista da área médica irá conectar-se – *login* – no PSW, cadastrar um exame, caso não haja um exame em andamento, e gerar um laudo médico utilizando a sua voz para transcrever o texto em formato digital. Cada laudo médico gerado deve estar associado a um exame referente a um atendimento de um paciente.

6.2 Levantamento de Requisitos

Nos Quadros 6.2, 6.3, 6.4, 6.5, 6.6, 6.7, 6.8, 6.9, 6.10 e 6.11 são apresentadas informações referentes à etapa de Levantamento de Requisito, como o Requisito Funcional (F), o Requisito Não-funcional (NF) e o Requisito Suplementar (S).

No Quadro 6.2 é apresentado o Requisito Funcional referente ao Cadastro do Profissional e os Requisitos Não-funcionais e Suplementares correspondentes.

Quadro 6.2: Requisito Funcional Cadastrar Profissional.

F1: Cadastrar Profissional			Requisito Funcional Evidente	
Descrição: o cadastro de profissional se refere ao cadastro de usuário comum e de usuário administrador. Quando um profissional é cadastrado, automaticamente é atribuído permissão de usuário comum. Um profissional pode ser cadastrado a partir do preenchimento dos campos obrigatórios de Nome, Sexo, Data de Nascimento, número do Cadastro de Pessoa Física, número do Registro Geral, número do Conselho Regional de Medicina, Especialidade, E-mail, *Login*, Senha, Rua, Número da Residência, Bairro, número do Código de Endereçamento Postal, Cidade e Estado. Os campos de preenchimento não obrigatório são referentes aos números de Celular e de Telefone e o Complemento.				
Requisitos Não-funcionais				
Nome	**Restrição**	**Categoria**	**Permanente e/ou Transitório**	**Obrigatório ou Desejável**
NF 1.1: Janela Única	O cadastro do profissional será realizado em uma única tela	Interface	Permanente	Obrigatório
NF 1.2: Informações Relevantes	A tela de cadastro de profissional deverá conter apenas informações relevantes	Interface	Transitório	Desejável
NF 1.3: Controle de Acesso	A função só poderá ser acessada por um usuário autenticado no sistema	Segurança	Permanente	Obrigatório
Requisitos Suplementares				
Nome	**Restrição**	**Categoria**	**Permanente e/ou Transitório**	**Obrigatório ou Desejável**
S 1.1: Armazenamento de Dados	Os dados deverão ser armazenados em um BD do PSW	Segurança	Permanente	Obrigatório
S 1.2: Tipo de Interface	A interface deverá ser acessível por meio de um navegador de *Internet*	Interface	Permanente	Obrigatório

No Quadro 6.3 é apresentado o Requisito Funcional referente ao Gerenciamento do Profissional cadastrado no PSW.

Quadro 6.3: Requisito Funcional Gerenciar Profissional.

F2: Gerenciar Profissional			Requisito Funcional Evidente	
Descrição: os usuários autenticados no sistema, com permissão de administrador, podem gerenciar os profissionais cadastrados. O gerenciamento de profissional consiste em ativar ou desativar um usuário, bem como excluí-lo ou torná-lo administrador.				
Requisitos Não-funcionais				
Nome	**Restrição**	**Categoria**	**Permanente e/ou Transitório**	**Obrigatório ou Desejável**
NF 2.1: Janela Única	O gerenciamento do profissional será realizado em uma única tela	Interface	Permanente	Obrigatório
NF 2.2: Informações Relevantes	A tela de gerenciamento de profissional deverá conter apenas informações relevantes	Interface	Transitório	Desejável
NF 2.3: Controle de Acesso	A função só poderá ser acessada por um usuário autenticado no sistema	Segurança	Permanente	Obrigatório
Requisitos Suplementares				
Nome	**Restrição**	**Categoria**	**Permanente e/ou Transitório**	**Obrigatório ou Desejável**
S 2.1: Armazenamento de Dados	Os dados deverão ser armazenados em um BD do PSW	Segurança	Permanente	Obrigatório
S 2.2: Tipo de Interface	A interface deverá ser acessível por meio de um navegador de *Internet*	Interface	Permanente	Obrigatório

No Quadro 6.4 é apresentado o Requisito Funcional referente ao Cadastro do Paciente.

Quadro 6.4: Requisito Funcional Cadastrar Paciente.

F3: Cadastrar Paciente			Requisito Funcional Evidente	
Descrição: o sistema deverá receber os seguintes campos com preenchimento obrigatório de Nome, Sexo, Data de Nascimento, número do Cadastro de Pessoa Física, número do Registro Geral, Nome da Mãe, Rua, Número da Residência, Bairro, número do Código de Endereçamento Postal, Cidade e Estado e os seguintes campos com preenchimento não obrigatório, como os números de Celular e de Telefone e o Complemento.				
Requisitos Não-funcionais				
Nome	**Restrição**	**Categoria**	**Permanente e/ou Transitório**	**Obrigatório ou Desejável**
NF 3.1: Janela Única	O cadastro do paciente será realizado em uma única tela	Interface	Permanente	Obrigatório

(Continuação)

Requisitos Não-funcionais				
Nome	**Restrição**	**Categoria**	**Permanente e/ou Transitório**	**Obrigatório ou Desejável**
NF 3.2: Informações Relevantes	A tela de cadastro de paciente deverá conter apenas informações relevantes	Interface	Transitório	Desejável
Requisitos Não-funcionais				
Nome	**Restrição**	**Categoria**	**Permanente e/ou Transitório**	**Obrigatório ou Desejável**
NF 3.3: Controle de Acesso	A função só poderá ser acessada por um usuário autenticado no sistema	Segurança	Permanente	Obrigatório
Requisitos Suplementares				
Nome	**Restrição**	**Categoria**	**Permanente e/ou Transitório**	**Obrigatório ou Desejável**
S 3.1: Armazenamento de Dados	Os dados deverão ser armazenados em um BD do PSW	Segurança	Permanente	Obrigatório
S 3.2: Tipo de Interface	A interface deverá ser acessível por meio de um navegador de *Internet*	Interface	Permanente	Obrigatório

No Quadro 6.5 é apresentado o Requisito Funcional Cadastro da Especialidade do Profissional.

Quadro 6.5: Requisito Funcional Cadastrar Especialidade.

F4: Cadastrar Especialidade			Requisito Funcional Evidente	
Descrição: o sistema deverá permitir o cadastro do nome da Especialidade, cujo preenchimento é obrigatório.				
Requisitos Não-funcionais				
Nome	**Restrição**	**Categoria**	**Permanente e/ou Transitório**	**Obrigatório ou Desejável**
NF 4.1: Janela Única	O cadastro da especialidade será realizado em uma única tela	Interface	Permanente	Obrigatório
NF 4.2: Informações Relevantes	A tela de cadastro da especialidade deverá conter apenas informações relevantes	Interface	Transitório	Desejável
NF 4.3: Controle de Acesso	A função só poderá ser acessada por um usuário autenticado no sistema	Segurança	Permanente	Obrigatório

(Continuação)

Requisitos Suplementares				
Nome	**Restrição**	**Categoria**	**Permanente e/ou Transitório**	**Obrigatório ou Desejável**
S 4.1: Armazenamento de Dados	Os dados deverão ser armazenados em um BD do PSW	Segurança	Permanente	Obrigatório
S 4.2: Tipo de Interface	A interface deverá ser acessível por meio de um navegador de *Internet*	Interface	Permanente	Obrigatório

No Quadro 6.6 é apresentado o Requisito Funcional Cadastro do Tipo de Exame, ou seja, qual a categoria do exame sugerido pelo profissional de saúde.

Quadro 6.6: Requisito Funcional Cadastrar Tipo de Exame.

F5: Cadastrar Tipo de Exame			Requisito Funcional Evidente	
Descrição: o sistema deverá permitir o cadastro do nome do Tipo de Exame que é de preenchimento obrigatório.				
Requisitos Não-funcionais				
Nome	**Restrição**	**Categoria**	**Permanente e/ou Transitório**	**Obrigatório ou Desejável**
NF 5.1: Janela Única	O cadastro do tipo de exame será realizado em uma única tela	Interface	Permanente	Obrigatório
NF 5.2: Informações Relevantes	A tela de cadastro de tipo de exame deverá conter apenas informações relevantes	Interface	Transitório	Desejável
NF 5.3: Controle de Acesso	A função só poderá ser acessada por um usuário autenticado no sistema	Segurança	Permanente	Obrigatório
Requisitos Suplementares				
Nome	**Restrição**	**Categoria**	**Permanente e/ou Transitório**	**Obrigatório ou Desejável**
S 5.1: Armazenamento de Dados	Os dados deverão ser armazenados em um BD do PSW	Segurança	Permanente	Obrigatório
S 5.2: Tipo de Interface	A interface deverá ser acessível por meio de um navegador de *Internet*	Interface	Permanente	Obrigatório

No Quadro 6.7 é apresentado o Requisito Funcional Cadastro de Exame.

Quadro 6.7: Requisito Funcional Cadastrar Exame.

F6: Cadastrar Exame			Requisito Funcional Evidente	
Descrição: o sistema deverá permitir o cadastro de Exame para um determinado paciente. Os campos com preenchimento obrigatório são Paciente, Tipo do Exame e Data. Os campos referentes ao Motivo do Exame, Observações do Exame e Encaminhado por – nome do profissional de saúde que encaminhou o paciente para o exame – são de preenchimento não obrigatório.				
Requisitos Não-funcionais				
Nome	**Restrição**	**Categoria**	**Permanente e/ou Transitório**	**Obrigatório ou Desejável**
NF 6.1: Controle de Cadastro	O exame deverá ser vinculado a um paciente	Confiabilidade	Permanente	Obrigatório
NF 6.2: Controle de Usuário	O sistema deverá atribuir ao usuário autenticado o cadastro do exame	Confiabilidade	Permanente	Obrigatório
NF 6.3: Janela Única	O cadastro do exame será realizado em uma única tela	Interface	Permanente	Obrigatório
NF 6.4: Informações Relevantes	A tela de cadastro de exame deverá conter apenas informações relevantes	Interface	Transitório	Desejável
NF 6.5: Controle de Acesso	A função só poderá ser acessada por um usuário autenticado no sistema	Segurança	Permanente	Obrigatório
NF 6.6: Ajuda para Cadastrar Exame	O sistema conterá ajuda para cadastrar um novo exame	Usabilidade	Transitório	Desejável
Requisitos Suplementares				
Nome	**Restrição**	**Categoria**	**Permanente e/ou Transitório**	**Obrigatório ou Desejável**
S 6.1: Armazenamento de Dados	Os dados deverão ser armazenados em um BD do PSW	Segurança	Permanente	Obrigatório
S 6.2: Tipo de Interface	A interface deverá ser acessível por meio de um navegador de *Internet*	Interface	Permanente	Obrigatório

No Quadro 6.8 é apresentado o Requisito Funcional Gerenciar Exame.

Quadro 6.8: Requisito Funcional Gerenciar Exame.

F7: Gerenciar Exame	Requisito Funcional Evidente

(Continuação)

Descrição: o sistema deverá permitir a realização do gerenciamento do exame que consiste em permitir alteração e consulta aos exames cadastrados no sistema. As consultas permitidas são por código do exame, nome do paciente, *login* do profissional e data de realização do exame.				
Requisitos Não-funcionais				
Nome	**Restrição**	**Categoria**	**Permanente e/ou Transitório**	**Obrigatório ou Desejável**
NF 7.1: Janela Única	O gerenciamento do exame será realizado em uma única tela	Interface	Permanente	Obrigatório
NF 7.2: Informações Relevantes	A tela de gerenciamento de exame deverá conter apenas informações relevantes	Interface	Transitório	Desejável
NF 7.3: Controle de Acesso	A função só poderá ser acessada por um usuário autenticado no sistema	Segurança	Permanente	Obrigatório
Requisitos Suplementares				
Nome	**Restrição**	**Categoria**	**Permanente e/ou Transitório**	**Obrigatório ou Desejável**
S 7.1: Busca de Dados	Os dados deverão ser buscados no BD do PSW	Segurança	Permanente	Obrigatório
S 7.2: Tipo de Interface	A interface deverá ser acessível por meio de um navegador de *Internet*	Interface	Permanente	Obrigatório

No Quadro 6.9 é apresentado o Requisito Funcional Cadastro do Laudo Médico.

Quadro 6.9: Requisito Funcional Cadastrar Laudo Médico.

F8: Cadastrar Laudo Médico			Requisito Funcional Evidente	
Descrição: o sistema deverá permitir o cadastro do laudo médico, que será vinculado a um exame, para um determinado paciente. O laudo médico é associado ao profissional autenticado no sistema. Os campos possuem preenchimento não obrigatório.				
Requisitos Não-funcionais				
Nome	**Restrição**	**Categoria**	**Permanente e/ou Transitório**	**Obrigatório ou Desejável**
NF 8.1: Controle de Cadastro	O laudo médico deverá ser vinculado a um exame	Confiabilidade	Permanente	Obrigatório

(Continuação)

Requisitos Não-funcionais				
Nome	**Restrição**	**Categoria**	**Permanente e/ou Transitório**	**Obrigatório ou Desejável**
NF 8.2: Controle de Data e Hora	O sistema deverá armazenar a data e a hora de conclusão do cadastro do laudo médico	Confiabilidade	Permanente	Obrigatório
NF 8.3: Controle de Usuário	O sistema deverá atribuir ao usuário autenticado o cadastro do laudo médico	Confiabilidade	Permanente	Obrigatório
NF 8.4: Janela Única	O cadastro do tipo de exame será realizado em uma única tela	Interface	Permanente	Obrigatório
NF 8.5: Informações Relevantes	A tela de cadastro de laudo médico deverá conter apenas informações relevantes	Interface	Transitório	Desejável
NF 8.6: Controle de Acesso	A função só poderá ser acessada por um usuário autenticado no sistema	Segurança	Permanente	Obrigatório
NF 8.7: Ajuda para Cadastrar Laudo Médico	O sistema conterá ajuda para cadastrar um novo laudo médico	Usabilidade	Transitório	Desejável
NF 8.8: Cadastrar Laudo com o SRAF da Google *Web Speech* API	O PSW executa o SRAF da Google desenvolvido em linguagem de programação JavaScript	Implementação	Transitório	Desejável
NF 8.9: Cadastrar Laudo com o SRAF da Microsoft Bing *Speech* API	O PSW executa o SRAF da Microsoft desenvolvido em linguagem de programação JavaScript	Implementação	Transitório	Desejável
Requisitos Suplementares				
Nome	**Restrição**	**Categoria**	**Permanente e/ou Transitório**	**Obrigatório ou Desejável**
S 8.1: Armazenamento de Dados	Os dados deverão ser armazenados em um BD do PSW	Segurança	Permanente	Obrigatório

(Continuação)

Requisitos Suplementares				
Nome	**Restrição**	**Categoria**	**Permanente e/ou Transitório**	**Obrigatório ou Desejável**
S 8.2: Tipo de Interface	A interface deverá ser acessível por meio de um navegador de *Internet*	Interface	Permanente	Obrigatório

No Quadro 6.10 é apresentado o Requisito Funcional Cadastro de Histórico do Laudo Médico.

Quadro 6.10: Requisito Funcional Cadastrar Histórico do Laudo Médico.

F9: Cadastrar Histórico do Laudo Médico			Requisito Funcional Oculto	
Descrição: quando um laudo médico cadastrado for editado, para cada nova edição, é gerado uma versão de histórico. Por exemplo, um laudo alterado duas vezes possui o laudo com a última alteração realizada e outras duas versões de histórico para esse laudo.				
Requisitos Não-funcionais				
Nome	**Restrição**	**Categoria**	**Permanente e/ou Transitório**	**Obrigatório ou Desejável**
NF 9.1: Controle de Cadastro	O laudo médico deverá ser vinculado a um exame	Confiabilidade	Permanente	Obrigatório
NF 9.2: Controle de Data e Hora	O sistema deverá armazenar a data e a hora de conclusão da alteração do laudo médico	Confiabilidade	Permanente	Obrigatório
NF 9.3: Controle de Usuário	O sistema deverá atribuir ao usuário autenticado a alteração do laudo médico	Confiabilidade	Permanente	Obrigatório
Requisito Suplementar				
Nome	**Restrição**	**Categoria**	**Permanente e/ou Transitório**	**Obrigatório ou Desejável**
S 9.1: Armazenamento de Dados	Os dados deverão ser armazenados em um BD do PSW	Segurança	Permanente	Obrigatório

No Quadro 6.11 é apresentado o Requisito Funcional Gerenciar Laudo Médico.

Quadro 6.11: Requisito Funcional Gerenciar Laudo Médico.

F10: Gerenciar Laudo Médico			Requisito Funcional Evidente	
Descrição: o sistema deverá permitir a realização do gerenciamento do laudo médico que consiste em permitir alteração e consulta aos laudos médicos cadastrados no sistema. As consultas permitidas são por código do laudo médico, código do exame, data de realização do exame, data de realização do laudo e nome do paciente. A exibição do laudo médico possibilita a consulta de todas as versões alteradas por meio do histórico do laudo médico.				
Requisitos Não-funcionais				
Nome	**Restrição**	**Categoria**	**Permanente e/ou Transitório**	**Obrigatório ou Desejável**
NF 10.1: Janela Única	O gerenciamento do laudo médico será realizado em uma única tela	Interface	Permanente	Obrigatório
NF 10.2: Informações Relevantes	A tela de gerenciamento de laudo médico deverá conter apenas informações relevantes	Interface	Transitório	Desejável
NF 10.3: Controle de Acesso	A função só poderá ser acessada por um usuário autenticado no sistema	Segurança	Permanente	Obrigatório
NF 10.4: Editar Laudo com o SRAF da Google *Web Speech* API	O PSW executa o SRAF da Google para continuar a transcrição do laudo	Implementação	Transitório	Desejável
NF 10.5: Editar Laudo com o SRAF da Microsoft Bing *Speech* API	O PSW executa o SRAF da Microsoft para continuar a transcrição do laudo	Implementação	Transitório	Desejável
Requisitos Suplementares				
Nome	**Restrição**	**Categoria**	**Permanente e/ou Transitório**	**Obrigatório ou Desejável**
S 10.1: Busca de Dados	Os dados deverão ser buscados no BD do PSW	Segurança	Permanente	Obrigatório
S 10.2: Tipo de Interface	A interface deverá ser acessível por meio de um navegador de *Internet*	Interface	Permanente	Obrigatório

No Quadro 6.12 são apresentadas as operações relacionadas a cadastrar, alterar, excluir e consultar que são habilitadas para cada Requisito Funcional.

Quadro 6.12: Operações habilitadas para cada Requisito Funcional do Protótipo de Sistema *Web*.

Requisitos	**Cadastrar**	**Alterar**	**Excluir**	**Consultar**
Cadastrar Profissional	X			
Gerenciar Profissional			X	
Cadastrar Paciente	X			
Cadastrar Especialidade	X			
Cadastrar Tipo de Exame	X			
Cadastrar Exame	X			
Gerenciar Exame		X		X
Cadastrar Laudo Médico	X			
Gerenciar Laudo Médico		X		X

A Figura 6.1 ilustra o Caso de Uso do PSW.

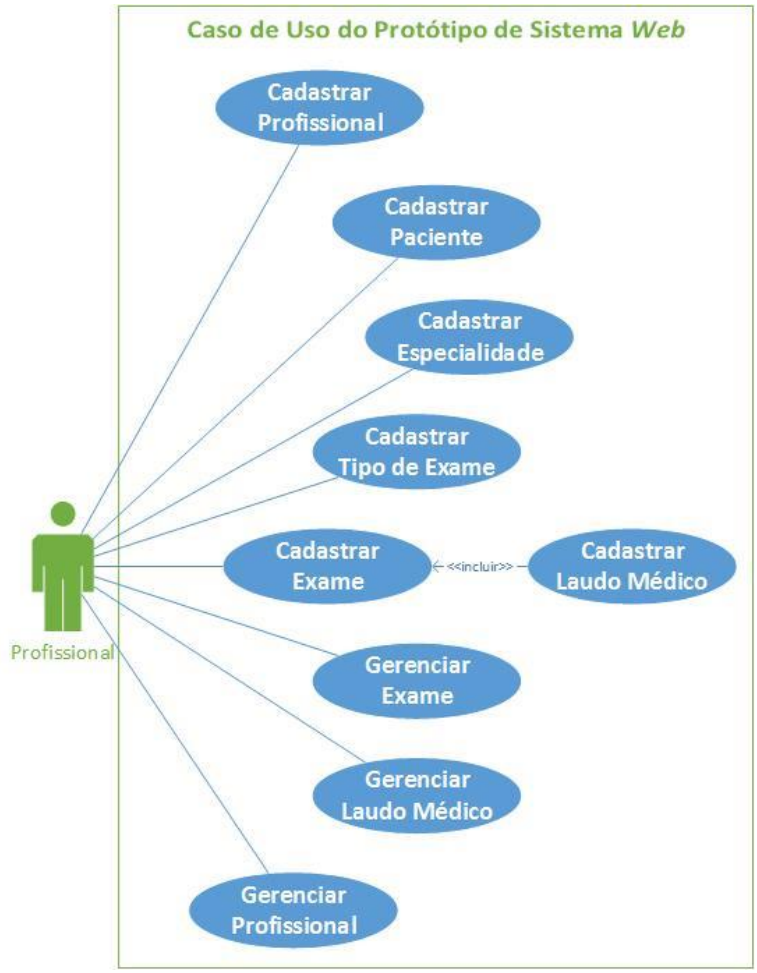

Figura 6.1: Caso de Uso com os principais Requisitos Funcionais do Protótipo de Sistema *Web*.

Na Figura 6.2 é ilustrado o Diagrama de Atividade para realizar o cadastro de um laudo médico. Pode-se perceber que o exame é vinculado a um paciente, e a partir do exame torna-se possível o cadastro de um laudo médico.

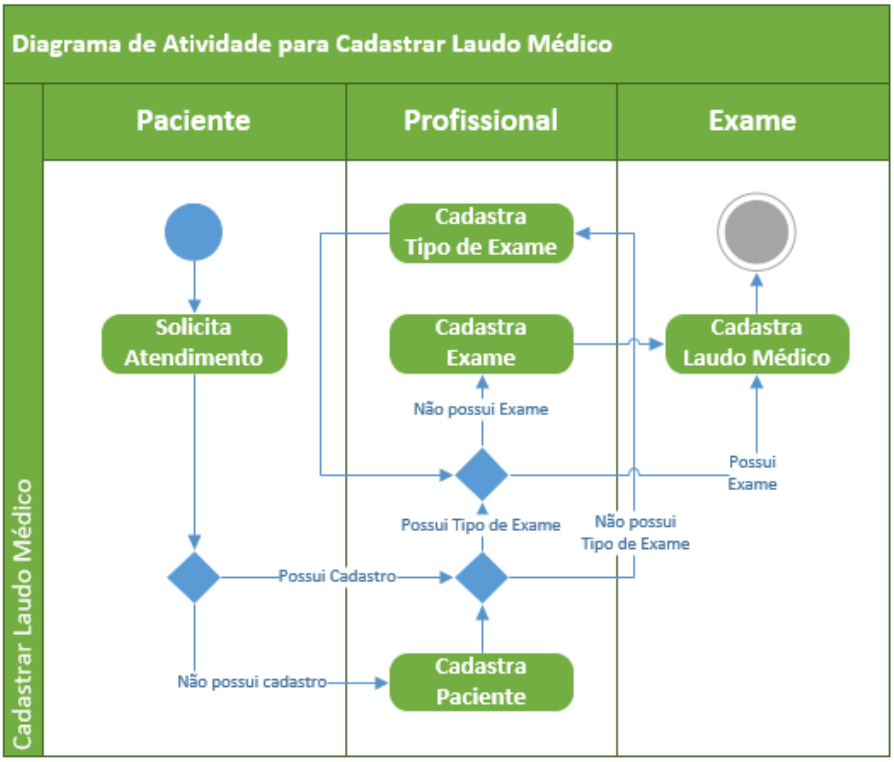

Figura 6.2: Diagrama de Atividade do Requisito Funcional de Cadastro de Laudo Médico.

6.3 Projeto Arquitetural

O diagrama da Figura 6.3 ilustra o Projeto Arquitetural, no qual as principais partes do sistema são apresentadas, tal como os dispositivos que podem acessá-lo. Com relação ao(s) SRAF(s) integrado(s) ao PSW, a sua seleção foi baseada a partir do resultado obtido na avaliação de um experimento preliminar (Seção 7.2).

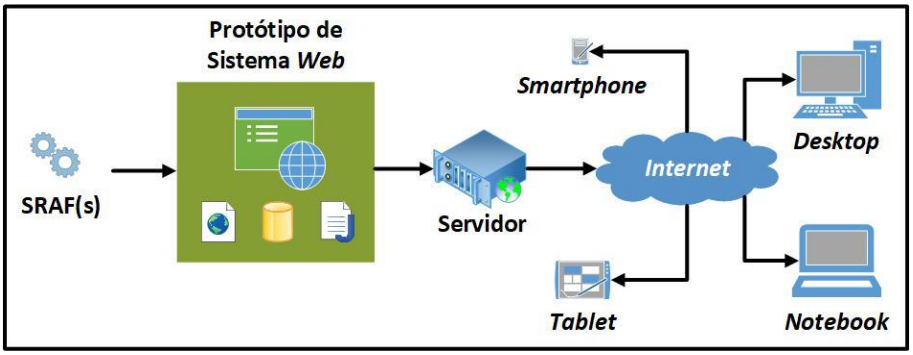

Figura 6.3: Diagrama da arquitetura geral do Protótipo de Sistema *Web*.

O PSW foi hospedado em um servidor Apache Tomcat e pode ser acessado via *Internet*, pelos seguintes dispositivos: microcomputadores (*desktops*), *notebooks*, *tablets* e *smartphones*.

6.4 Projeto Detalhado

Na etapa do Projeto Detalhado foi definida toda a estrutura do PSW incluindo, por exemplo, os subsistemas da tecnologia de reconhecimento automático de fala que são incluídas. Além dos subsistemas, o PSW conta com dois módulos: o módulo principal e o módulo administrador.

O módulo principal tem como finalidade realizar os cadastros do sistema e, sobretudo, realizar o cadastro de exames vinculando os laudos médicos gerados.

Já o módulo administrador possui a funcionalidade de gerenciar os usuários que podem acessar o sistema e atribuir permissão de administrador a outros usuários, habilitando-os a gerenciar todos os usuários do sistema.

Na Figura 6.4 é ilustrado o Modelo Conceitual do PSW.

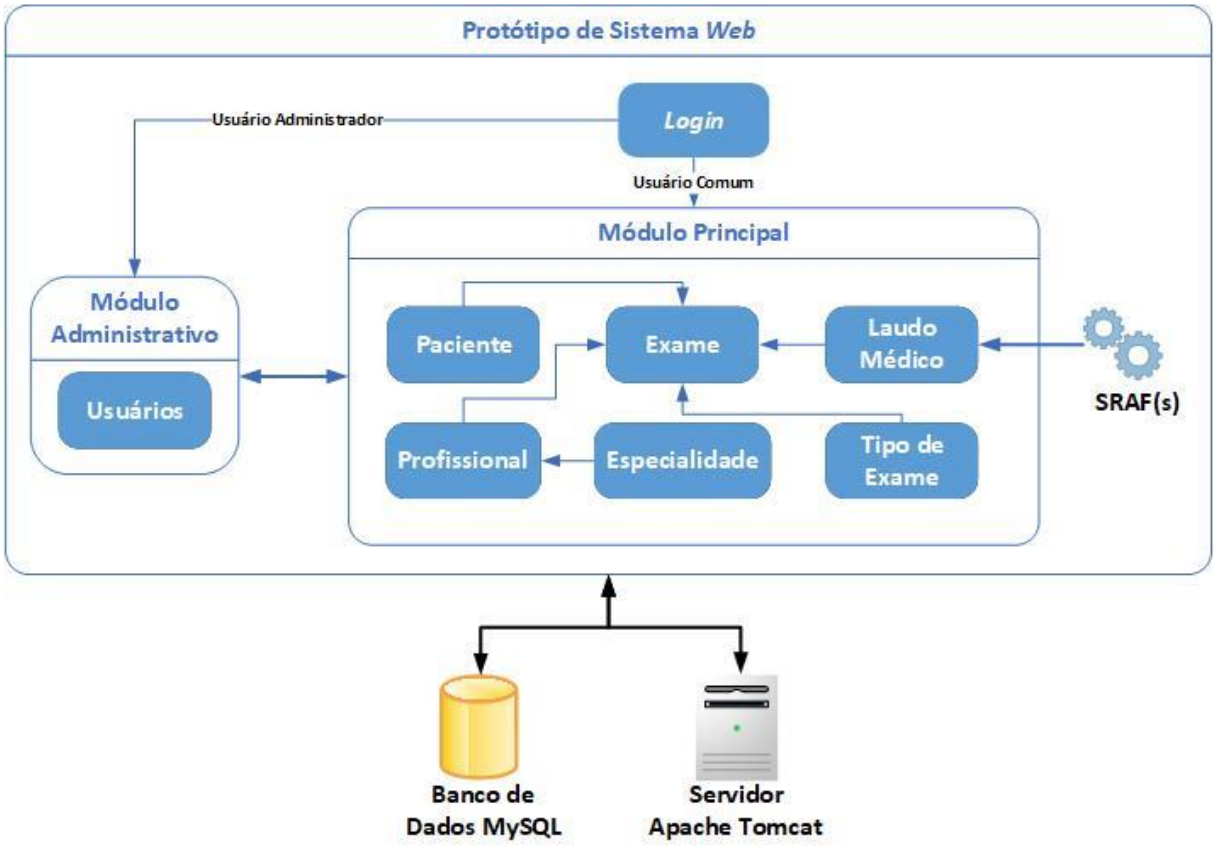

Figura 6.4: Modelo Conceitual do Protótipo de Sistema *Web*.

6.5 Codificação e Depuração

A etapa de construção do PSW corresponde à etapa de Codificação e Depuração. Nessa etapa, o Maven foi utilizado para prover uma estrutura de diretórios para o PSW, cujo intuito foi o de organizar os arquivos-fontes (Tabela 6.1) e gerenciar as dependências do projeto, ou seja, os arquivos JARs são gerenciados por meio do arquivo de Modelo de Objeto do Projeto – *Project Object Model* (pom.xml).

O arquivo pom.xml possibilita a vinculação de dependências externas do projeto. O Maven foi configurado para utilizar a versão do Java 1.8.

Tabela 6.1: Organização dos arquivos-fonte do projeto.

Diretório	**Descrição**
/	Raiz do projeto.
/pom.xml	Arquivo que contém as informações para o funcionamento do Maven, tais como: a sua versão, a distribuição WAR do PSW e as dependências necessárias para o funcionamento do projeto.
/*src*/*main*/java	Arquivos-fontes do Java e o arquivo de configuração do Hibernate hibernate.cfg.xml.
/*src*/*main*/*webapp*	Arquivos-fontes das páginas *Webs*, como os arquivos em xHTML, CSS e Javascipt.
/*src*/*main*/*webapp/admin*	Arquivo-fonte xHTML da página administradora.
/*src*/*main*/*webapp/META-INF*	Arquivo-fonte XML com as configurações de acesso ao BD.
/*src*/*main*/*webapp/*publico	Arquivos-fontes xHTMLs de acesso público, como as páginas de *login* e do rodapé do PSW.
/*src*/*main*/*webapp/resource*	Diretórios css, imagens e js que contém o arquivo-fonte CSS, as imagens do PSW e os arquivos-fontes em JavaScripts, respectivamente.
/*src*/*main*/*webapp/restrito*	Arquivos-fontes xHTMLs com as páginas de acesso restrito para os usuários autenticados.
/*src*/*main*/*webapp/template*	Arquivo-fonte xHTML com o visual comum a todas as páginas do PSW.
/*src*/*main*/*webapp/WEB-INF*	Arquivos-fontes XMLs com configurações e carregamento do Spring *Security*, Hibernate e Primefaces para quando o servidor do PSW for iniciado.
/*target*	Diretório com os arquivos gerados pelo Maven e o arquivo WAR.

6.5.1 Persistência dos Dados

Para a persistência dos dados, ou seja, armazenar de maneira permanente todos os registros salvos no PSW para serem recuperados posteriormente, é necessário um BD e uma

maneira de se comunicar a ele. A persistência dos dados foi realizada por meio do Hibernate que, de maneira automatizada, gera as tabelas do BD relacional.

A comunicação com o BD foi realizada por meio do *driver* Java *Database Connectivity* (JDBC[43]). O JDBC é uma especificação de comunicação entre a linguagem Java com o BD. O BD do PSW é o *MySQL*, portanto, foi utilizado o conector *MySQL Connector*[44] versão 5.1.30.

6.5.2 Separação em Camadas de Responsabilidades

O PSW é desenvolvido em camadas de responsabilidades com o intuito de facilitar manutenções futuras do sistema. As camadas de responsabilidade são: camada de acesso aos dados para realizar a gravação dos registros no BD, camada de regra de negócio para a tomada de decisão das funcionalidades do sistema e a camada de apresentação para o desenvolvimento do visual do sistema para a exibição dos dados.

A camada de acesso aos dados possui como objetivo realizar a conexão com o BD e a leitura ou a gravação dos dados. Essa capacidade de separar a responsabilidade do acesso aos dados em classes específicas é baseada pelo Padrão de Projeto *Data Access Object* (DAO), o qual possui a finalidade de prover o acesso aos dados de maneira transparente para as classes da camada de regra de negócio.

A camada de regra de negócio (RN) é responsável por decidir o que deve ser gravado ou recuperado do BD. Cada classe da RN representa uma tabela do BD, por exemplo, a tabela de usuários, que possui como regra de negócio a possibilidade de cadastrar novos usuários, desde que seja usuário administrador. Após o cadastro do usuário, a camada de RN repassa as informações à camada DAO para que o comando solicitado pelo usuário seja executado.

Já a camada de apresentação é responsável pela exibição e coleta de dados do usuário. Por exemplo, o usuário solicita a listagem de pacientes cadastrados no sistema, e então, essa solicitação é enviada para a RN que repassará para o DAO, que possui a implementação do método para recuperar no BD os pacientes cadastrados. A listagem é retornada realizando o caminho inverso até ser exibida na tela do usuário. Isso é necessário, pois, a camada de apresentação tem acesso direto apenas à camada de RN, porém, não possui acesso direto ao DAO.

Cada classe da camada de apresentação contém um arquivo-fonte em xHTML. Esse arquivo contém a codificação da tela de exibição para o usuário e uma classe correspondente em Java, a qual possui a função de chamar o método adequado da RN e retornar à solicitação para ser exibida ao usuário.

Essas três camadas compõem o *Plain and Old* Java *Object* (POJO), podendo ser traduzido como: “singelo clássico objeto Java” ou “bom e velho objeto Java”. A Figura 6.5 ilustra um exemplo de cadastro de paciente utilizando a separação por camadas.

[43] http://www.oracle.com/technetwork/java/javase/jdbc/index.html
[44] https://www.mysql.com/products/connector/

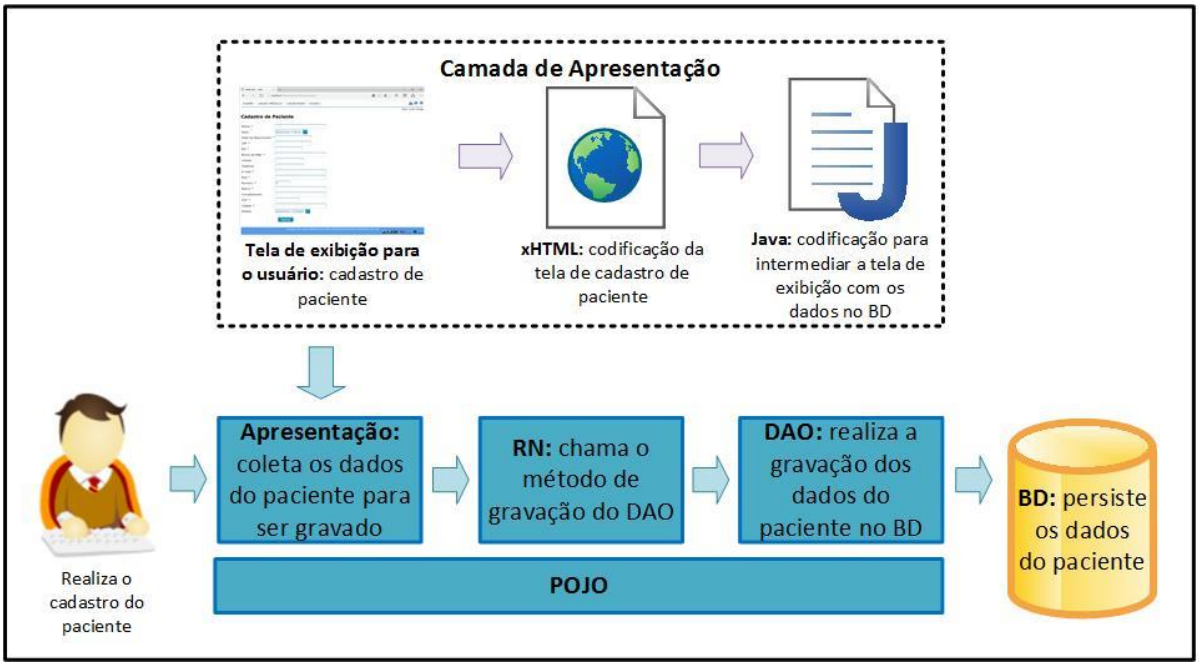

Figura 6.5: Separação de responsabilidade por três camadas.

Como mencionado, para adequar a separação de responsabilidades em camadas, foi utilizado o padrão de arquitetura MVC. Na Figura 6.6 é ilustrada a maneira de como esse padrão foi utilizado para modelar o PSW em paralelo com a codificação separada em camadas de responsabilidades.

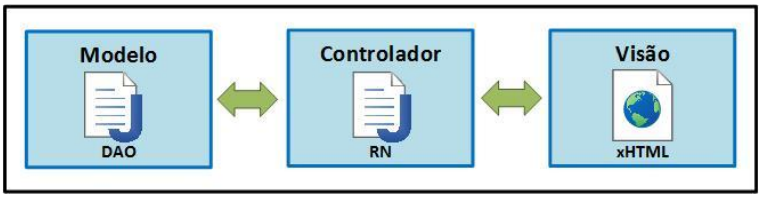

Figura 6.6: Paralelo entre o padrão de arquitetura Modelo-Visão-Controlador (MVC) com a separação de responsabilidade em camadas.

6.5.3 Proteção do Protótipo de Sistema *Web* com Spring *Security*

O Spring *Security*[45] é um *framework* para fornecer segurança de acesso para as páginas *Web* (Alves, 2015) e foi utilizado para garantir que apenas usuários autenticados possam acessar as funcionalidades do PSW.

Como o acesso a todas as páginas do PSW requerem autenticação do usuário, o primeiro usuário administrador do sistema é criado de maneira manual diretamente no BD. Quando o usuário administrador se autenticar no PSW, ele pode cadastrar novos usuários. Na Figura 6.7 é ilustrado o fluxograma do funcionamento geral do *framework* Spring *Security*.

[45] https://projects.spring.io/spring-security/

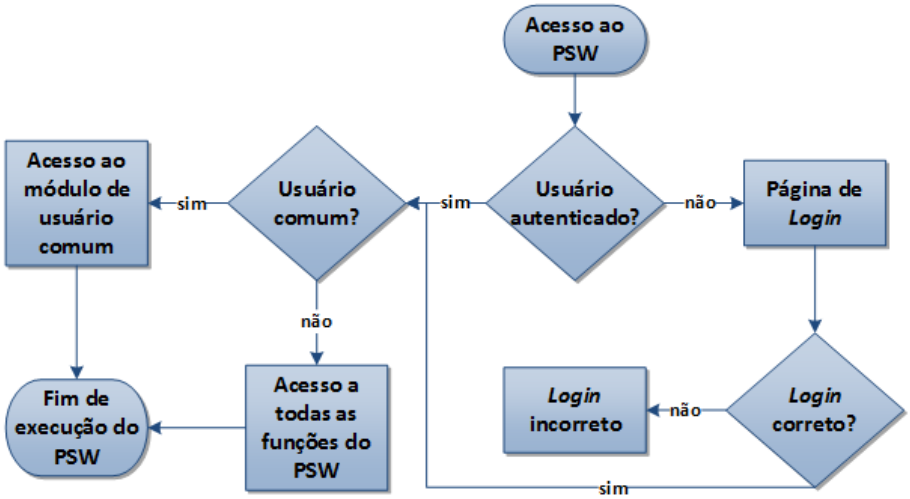

Figura 6.7: Fluxograma do funcionamento geral de autenticação do Protótipo de Sistema *Web*.

O PSW possui dois níveis de usuários: usuário comum, com acesso ao módulo de cadastro, e o usuário administrador, que possui acesso ao módulo de cadastro e ao módulo administrador, que pode cadastrar novos usuários e bloquear o acesso de qualquer usuário.

6.5.4 Requisitos para a Utilização do Protótipo de Sistema *Web*

Para a instalação do PSW em um servidor é necessário possuir os sistemas operacionais Windows XP ou Linux com o Java versão 1.8.

Com relação aos requisitos para os clientes desktops e notebooks do PSW, são necessárias as seguintes versões para os navegadores de *Internet*: Google Chrome versão 59.0.3071.115, *Edge* versão 38.14393.1066.0 e Mozilla Firefox versão 54.0.1.

Para os clientes acessando por meio de dispositivos portáteis, como os smartphones e os tablets, é necessário possuir os seguintes sistemas operacionais: Android versão 4.1.2 ou o WindowsPhone versão 10.

6.6 Teste e Entrega

As etapas de Codificação e Depuração, Teste e Entrega foram realizadas quatro vezes. Na primeira entrega foram apresentados o módulo de cadastro e administrativo e a integração do subsistema de reconhecimento automático de fala da Google *Web Speech* API. Ao final dessa entrega foi definido o seguinte ajuste: criar histórico de laudos alterados.

Ao final da primeira entrega, seguindo o método de Entrega em Estágio, voltou à etapa de Codificação e Depuração. Quando a implementação foi concluída, uma nova entrega foi apresentada. Nessa entrega, os seguintes ajustes foram sugeridos para melhorar a aparência do PSW:

- Alterar as imagens do rodapé;

- Padronizar o tamanho dos campos de texto para as telas de cadastro de exame e laudo médico;
- Exibir os 30 primeiros exames e laudos médicos cadastrados em suas respectivas telas;
- Permitir que os laudos médicos pudessem ser editados por meio do SRAF;
- Exibir o nome de qual SRAF foi utilizado para gerar o laudo médico.

Após as novas sugestões, retornou-se à etapa de Codificação e Depuração em que foram realizados os ajustes sugeridos, bem como, integrado o segundo SRAF da Microsoft Bing *Speech* API. Na sequência, uma nova entrega foi proposta, sendo sugeridas novas correções:

- Resolver a demora em que o PSW exibe a transcrição do reconhecimento na geração de laudo médico com o SRAF Bing *Speech* API;
- Permitir que o usuário selecione a data no momento do cadastro de um exame em substituição ao PSW definir automaticamente com a data do dia de cadastro.

Na última entrega, foram sugeridas implementações de duas novas funcionalidades: permitir a transcrição dos laudos médicos a partir de arquivos de áudio e gerar laudos médicos no formato *Portable Document Format* (PDF).

Com relação ao Requisito de Empacotamento, ou seja, de qual maneira o PSW será distribuído, foi utilizado o *Web application ARchive* (WAR), que é um JAR empregado para distribuir a aplicação desenvolvida com a especificação JSF. A etapa de Entrega será detalhada na Seção 7.4, na qual são apresentadas todas as características e telas do PSW.

6.7 Considerações Finais

Durante a Concepção Inicial do projeto, foi definido o principal objetivo do PSW que é a geração de laudos médicos utilizando tecnologia de reconhecimento automático de fala.

Na sequência foram definidos os requisitos do sistema, como os cadastros de profissional, paciente, especialidade, tipo de exame, exame e laudo médico e o gerenciamento dos exames, dos laudos médicos e dos usuários cadastrados.

Na etapa do Projeto Arquitetural, a arquitetura do PSW foi definida, considerando a integração dos SRAFs e os dispositivos eletrônicos capazes de acessar o PSW. Em seguida, no Projeto Detalhado, foram definidos os dois módulos do sistema, contando com o módulo de cadastro e o módulo administrativo e o BD utilizado, bem como, o servidor para hospedar a aplicação – Apache Tomcat.

Na fase de Codificação e Depuração foram detalhados os principais pontos do desenvolvimento do PSW, como a maneira de persistir os dados no BD; a separação em camadas, com cada uma representando uma responsabilidade; a proteção do sistema ao solicitar senha de acesso; e os requisitos mínimos para que um servidor possa hospedar o PSW, e também os requisitos dos dispositivos para poder acessá-los.

Durante o desenvolvimento do PSW, os Testes basicamente foram realizados concomitantemente com o seu desenvolvimento, no qual foram efetuados em quatro entregas, com cada uma delas sendo definidas novas funcionalidades ou correções a serem implementadas.

Com relação a quarta entrega, as funcionalidades sugeridas de gerar laudo médico no formato PDF e permitir a sua transcrição por arquivo de áudio não foram implementadas.

No próximo capítulo serão apresentados os resultados e a discussão referentes a RS, a avaliação dos SRAFs e, ao final, o PSW será apresentado.

Capítulo 7

Resultados e Discussão

Neste capítulo são relatados os resultados da revisão sistemática, como a extração de informações dos trabalhos e suas respectivas descrições.

A avaliação preliminar dos sete SRAFs para a Língua Portuguesa do Brasil é apresentada, e na sequência, são discutidos os sistemas selecionados para serem acoplados ao PSW. Também são apresentadas as principais características do PSW.

Dessa maneira, na Seção 7.1, é abordada uma visão geral com as fases da revisão sistemática[46]. Na Seção 7.2, são apresentados os resultados e a discussão dos SRAFs avaliados no experimento preliminar. O experimento final com os SRAFs selecionados é detalhado na Seção 7.3. E por fim, o PSW é apresentado na Seção 7.4.

7.1 Revisão Sistemática

Inicialmente foi verificada a real necessidade de se realizar a RS. Para essa tarefa, foi utilizado o fluxograma ilustrado na Figura 5.1, no qual foi constatado a inexistência de RS para o tema proposto, referente a SRAF. Dessa maneira, iniciou-se a RS. Ao final da pesquisa por trabalhos que atendiam aos requisitos iniciais nas bases de dados, foram retornados 9.728 artigos compreendendo os anos de 2011 a 2017.

Com relação aos anos de 2015 e 2017, não foi possível realizar a atualização da RS na base de busca do CiteSeerX, pois a sequência de palavra, sem adição de filtro de intervalo de datas de 2011 a 2017 utilizada na pesquisa, retornou 4.379.324 publicações.

Com relação à realização de busca avançada utilizando a sequência de palavras para procurar palavras correspondentes no resumo das publicações e adicionando o intervalo de data para os anos de 2015 a 2017, foram retornadas 3.528.464 publicações. A mesma pesquisa foi realizada para encontrar palavras correspondentes no título dos artigos, sendo retornados os mesmos 3.528.464 artigos. A pesquisa por título sem aplicação de intervalo de datas também retornou 4.379.324 publicações.

Foram realizadas tentativas de busca nos meses de agosto, setembro e final de outubro de 2017, porém, o número de publicações retornadas não se alteraram, inviabilizando a pesquisa na base de dados do CiteSeerX.

46 A tabela completa da Revisão Sistemática pode ser verificada por meio do link: https://goo.gl/EgpCZw

Para verificar uma possível solução para esse problema foi tentado dois contatos através da página de ajuda do CiteSeerX[47]. Com a primeira tentativa sendo realizada em agosto de 2017 e a segunda tentativa em setembro de 2017. Até o final de outubro de 2017 não houve retorno do CiteSeerX.

Após a aplicação do critério de exclusão por títulos duplicados foram selecionados 6.588 trabalhos, conforme é ilustrado na Figura 7.1.

Figura 7.1: Quantidade de publicações por ano encontradas nas bases de busca.

Depois da aplicação dos demais critérios de exclusão, foram selecionadas 513 publicações, das quais foram extraídas informações. Na terceira fase, foi avaliada a qualidade desses trabalhos por meio da atribuição de pontuação.

Em seguida, na quarta fase, os trabalhos foram ordenados por SRAFs. Para cada sistema encontrado (SRAF único), foram selecionados todos os trabalhos que utilizaram esse sistema, formando grupos de publicações para cada SRAF. Dessa maneira, os que obtiveram maiores pontuações foram selecionados para a análise completa. Essa fase resultou na seleção de 136 publicações.

Na Figura 7.2 é ilustrado o fluxograma com as quatro fases específicas da Revisão Sistemática (RS) com a sua respectiva quantidade de trabalhos selecionados. Essas fases foram: (1) seleção por critérios de exclusão, (2) extração de informações, (3) avaliação de qualidade das publicações, e (4) seleção de publicações com SRAF único e por critério de qualidade.

[47] http://csxstatic.ist.psu.edu/contact

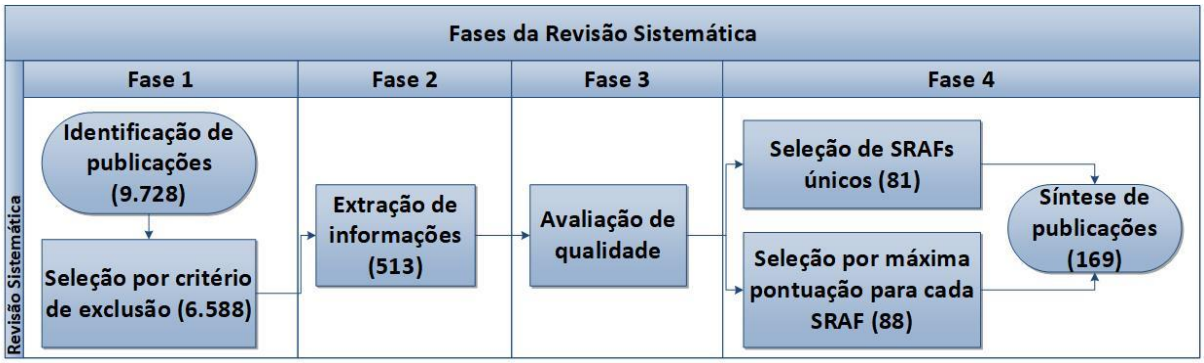

Figura 7.2: Fluxograma com a quantidade de publicações
para as fases da Revisão Sistemática.

Os resultados dos SRAFs encontrados são apresentados na Subseção 7.1.1, incluindo as tecnologias utilizadas por esses sistemas, como modelagem acústica, extração de características e *corpus*; e nível de precisão. Vale ressaltar que o número total desses atributos levantados não totaliza o número total dos trabalhos da terceira fase (513), pois algumas publicações apresentaram mais de um atributo e outros não apresentaram algumas dessas informações.

7.1.1 Resultados da Extração de Informações

Para essa etapa foram considerados 513 trabalhos, que foram pré-selecionados na segunda fase da RS. Na Figura 7.3, são ilustrados os SRAFs que foram mencionados em duas ou mais publicações.

O *Hidden Markov Model Toolkit*[48] (HTK), o CMU *Sphinx toolkit*[49], o Kaldi[50] e o PocketSphinx[51] são ferramentas de código aberto para o desenvolvimento de SRAFs.

O HTK é uma ferramenta para o desenvolvimento e a manipulação de Modelos Ocultos de Markov, que contém um conjunto de módulos de bibliotecas e ferramentas disponíveis em linguagem de programação C (HTK, 2016).

O CMU *Sphinx*, desenvolvido na linguagem de programação Java, é composto por modelos acústicos e aplicações de amostras, além de um *software* adicional para o treinamento do Modelo Acústico, compilação do Modelo de Linguagem e dicionário de pronúncia de domínio público (CMUSphinx, 2017).

Já o PocketSphinx é um motor de reconhecimento de fala, desenvolvido na linguagem de programação C, que exige pouco recurso de processamento computacional, tornando-o ideal para dispositivos portáteis e móveis (PocketSphinx, 2017).

A ferramenta Kaldi foi desenvolvida, na Universidade Johns Hopkins em 2009, nos Estados Unidos. O seu desenvolvimento é baseado na linguagem de programação C++, cuja

[48] http://htk.eng.cam.ac.uk/
[49] https://cmusphinx.github.io/
[50] http://kaldi-asr.org/
[51] http://cmusphinx.sourceforge.net/wiki/versions

proposta é a de disponibilizar algoritmos mais genéricos possíveis para viabilizar o seu uso em aplicações diversas (Kaldi, 2017).

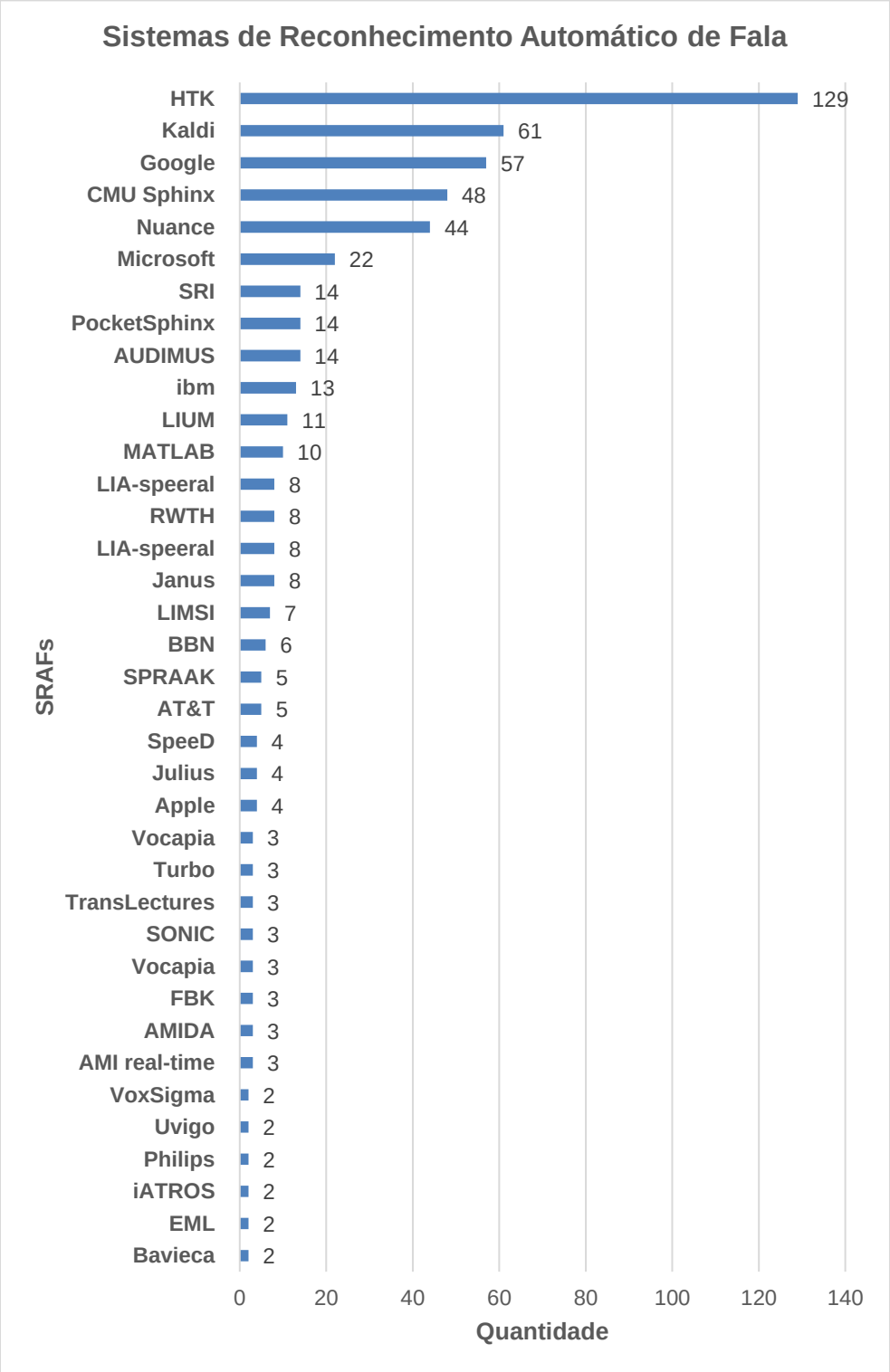

Figura 7.3: Quantidade de Sistemas de Reconhecimento Automático de Fala apresentados em pelo menos duas publicações.

Com relação às principais empresas que produzem tecnologia para o reconhecimento de fala, destaca-se, a Nuance[52], a VoiceInteraction[53], a SRI *International*[54] e a IBM Watson[55], cujo foco é o de desenvolvimento de SRAFs de uso comercial. A Nuance possui soluções corporativas, para usuários comuns e também para o âmbito hospitalar. Dentre algumas das suas soluções, destacam-se o seu uso em tarefas para realizar pesquisas na *Internet*, criar relatórios e planilhas, entre outros (Nuance, 2017).

A VoiceInteraction é uma empresa especializada em produtos do segmento de processamento de fala. Alguns outros serviços oferecidos pela empresa são: (a) *closed caption* automático, que é a legendagem para programas de televisão e de rádio com o intuito de transcrever para texto, (b) síntese de fala para um ou vários falantes, (c) transcrição de fala para texto, de um ou mais falantes, a partir de um arquivo de áudio, e (d) quiosque interativo, em que um agente virtual é capaz de responder a perguntas sobre conteúdos associados em um determinado tema (VoiceInteraction, 2017).

A SRI desenvolve SRAFs para aplicações de aprendizagem e treinamento de informática, como ensino de línguas estrangeiras, desenvolvimento de leitura e tutorial interativo e treinamento corporativo e simulação. Além de fornecer um motor de reconhecimento de fala para ser acoplado em outras aplicações, também possui um conjunto de ferramentas para a construção e a aplicação de MLs probabilísticos, marcação e segmentação estatística e tradução automática (SRI, 2017).

A IBM desenvolveu o Watson, uma tecnologia cognitiva cujo objetivo é o de analisar e interpretar dados, realizar recomendações de acordo com a personalidade, tom da fala e emoções do usuário, criar *bots* para *chats* (aplicações de *softwares* para simular ações humanas), entre outros (IBM, 2017).

Em relação a Google, as tecnologias de reconhecimento de fala desenvolvidas por esta empresa incluem: *Web Speech* API e *Cloud Speech* API. O *Web Speech* API permite incorporar o reconhecimento e a síntese de fala em páginas da *Web* (Shires & Wennborg, 2014). Já o *Cloud Speech* API conta com serviço de aprendizado de máquina, que incluem reconhecimento de fala, tradução de fala e texto, classificação de imagens, dentre outros (Google, 2017).

A Microsoft também disponibiliza algumas ferramentas que permitem o desenvolvimento de aplicações que utilizam o reconhecimento de fala, como, o *Speech* API (SAPI)[56], o *Software Development Kit* (SDK)[57] e *Cognitive Services* APIs. O Microsoft *Cognitive Services* é um serviços para os desenvolvedores adicionarem recursos às suas aplicações, por exemplo, algoritmos de processamento de imagens para o reconhecimento facial e de emoção e compreensão da fala e da linguagem (Microsoft, 2017e).

Os demais sistemas mencionados nos trabalhos foram: 800-FREE-411, 800-CALL-411, 800-555-TELL, AT&T Navigator, ChaCha, Speak4It, Speak4itSM, Vlingo, Yahoo! One

[52] http://www.nuance.com/index.htm
[53] http://www.voiceinteraction.com.br
[54] http://www.speech.sri.com/
[55] https://www.ibm.com/watson/
[56] https://msdn.microsoft.com/en-us/library/ee125663(v=vs.85).aspx
[57] https://msdn.microsoft.com/en-us/library/hh362943(v=office.14).aspx

Search e YPMobile (Feng et al., 2011); A-STAR (Sakti et al., 2013); Barista (Can et al., 2014); Bonilla et al., (2016); BUT (Karafiát et al., 2015); Butko & Nadeu (2011); Chalegre-Paula & Neto, (2016); Cut et al. (2013); Dictation PRO e E-Speaking (Shanmugapriya & RajaMohammed, 2014); FLaVoR (Yilmaz et al., 2014); Fluentiala (Soller et al., 2012); Fraunhofer IAIS (Stadtschnitzer et al., 2014); Gallardo-Antolín et al. (2013); García-Moral et al. (2011); GE Healthcare (Prevedello et al., 2014); Hearch (Varona et al., 2011); ICSI Parallex (Sheffield et al., 2013); InproTK (Chao & Thomaz, 2016); JaCHMM (Ultes et al., 2013); JSAPI (Balaji & Sadashivappa, 2015); Li et al., (2017); LIMSI-Vocapia (Clavel et al., 2013); LIUM-CRIM (Gupta et al., 2015); L&H (Johnson et al., 2014); Mengistu et al. (2011); OpenDial (Lison, 2015); OpenFST (Ajmera et al., 2012); Philips-RWTH (Riemann et al., 2016); Revuelta-Martínez et al. (2012); Saarland University (Helmke et al., 2015); SAVAS (Álvarez et al., 2016); SHOUT (Sinclair et al., 2014); SOLON (Delcroix et al., 2013); SpeaKING (Ahlgrim et al., 2016); SpeeD (Cucu et al., 2014); SUMMIT (Hazen, 2011); SVoG (Adde, 2013); Tazti (Shanmugapriya & RajaMohammed, 2014); tri4b (Benton & Dredze, 2015); USC SAIL (Misu et al., 2012); Verbio (Griol et al., 2014); VoiceTra (Matsuda et al., 2017); WebASR (Hain et al., 2016); Zhang et al., (2017).

Na Figura 7.4 é ilustrada a distribuição anual com as quantidades das publicações selecionadas na RS.

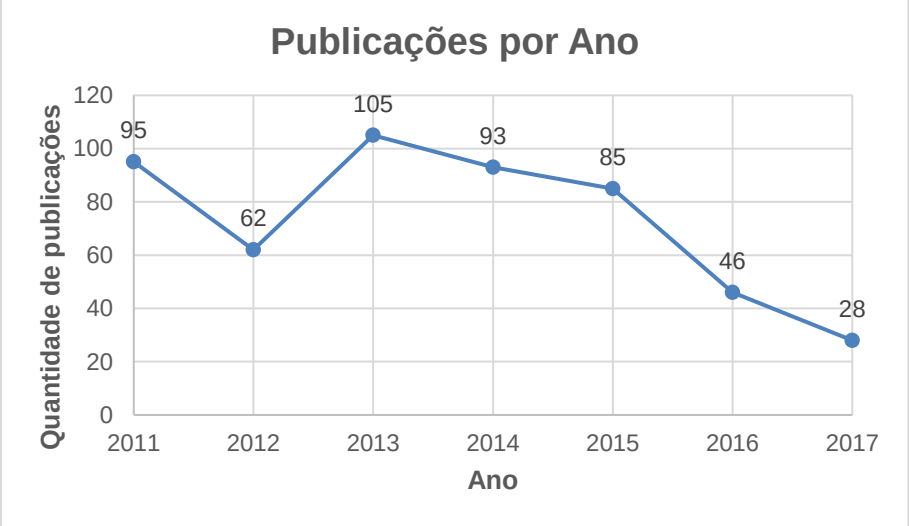

Figura 7.4: Quantidade de publicações por ano.

Nas Figuras 7.5, 7.6, 7.7 e 7.8 são apresentados os gráficos com as tecnologias dos SRAFs, como modelagem acústica, extração de características, métrica de avaliação da taxa de precisão e *corpus* utilizado para o seu treinamento. Nesses gráficos, o rótulo "Não informado" se refere à porcentagem sobre o número total de trabalhos considerados: 513.

Nos trabalhos encontrados, a principal tecnologia para a modelagem acústica (Figura 7.5) foi o Modelo Oculto de Markov – *Hidden Markov Model* (HMM) –, seguido por Modelo de Mistura Gaussiana – *Gaussian Mixture Models* (GMM) e Rede Neural Profunda – *Deep Neural Network* (DNN). Para os SRAFs que utilizaram tecnologia híbrida, o HMM continua sendo bastante utilizado em conjunto com GMM, Perceptron Multicamadas – *Multi-layer*

Perceptron (MLP) – e DNN. A categoria "Não Informado" compreende a tecnologia para a modelagem acústica que não tenha sido informada no artigo ou quando tal tecnologia utilizada pelo SRAF não tenha sido encontrada em outras publicações. As demais tecnologias encontradas estão no Apêndice C.

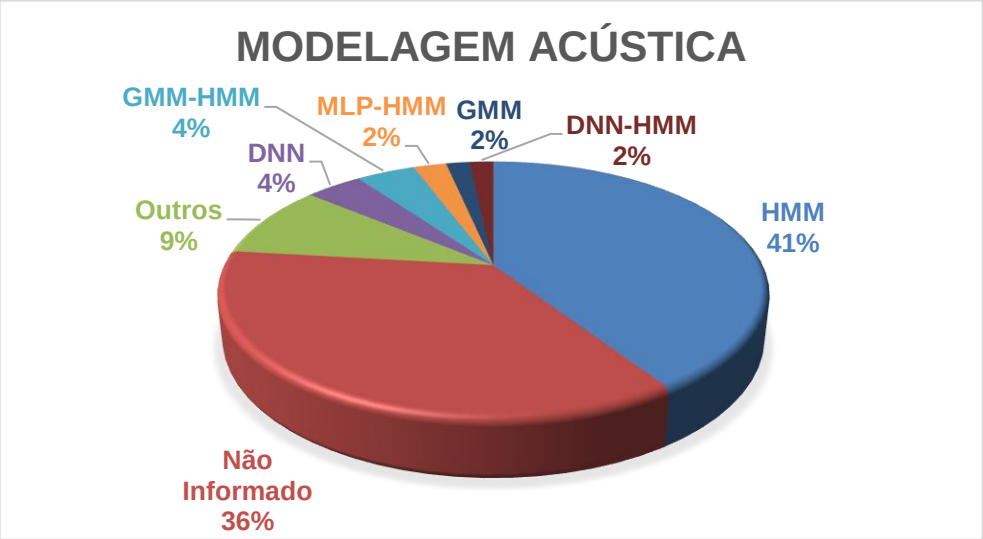

Figura 7.5: Tecnologia para a modelagem acústica utilizadas nos Sistemas de Reconhecimento Automático de Fala.

Com relação às técnicas de extração de características, a Figura 7.6 ilustra as principais tecnologias utilizadas nos SRAFs. Conforme pode ser observado, o MFCC é a técnica predominante, seguido por Percepção Linear Preditiva – *Perceptual Linear Prediction* (PLP). A categoria "Não Informado" compreende a tecnologia de extração de características que não tenha sido informada no artigo ou quando tal tecnologia utilizada pelo SRAF não tenha sido encontrada em outras publicações. As demais tecnologias encontradas estão no Apêndice D.

Figura 7.6: Técnicas de Extração de Características utilizadas nos Sistemas de Reconhecimento Automático de Fala.

Após a utilização de métodos para realizar o processo do reconhecimento de fala, é importante avaliar o nível de precisão de um SRAF. As principais técnicas utilizadas podem ser verificadas na Figura 7.7. A técnica mais utilizada é a taxa de erro de palavra – *Word Error Rate* (WER) –, seguida pela Taxa de Reconhecimento de Palavra – *Word Recognition Rate* (WRR). A categoria "Não Informado" compreende a quantidade de publicações que não realizaram nenhum tipo de avaliação de precisão. As demais métricas de avaliação encontradas estão no Apêndice E.

Figura 7.7: Técnicas para avaliação da taxa de precisão para os Sistemas de Reconhecimento Automático de Fala.

Com relação aos principais *corpora* utilizados para o treinamento de SRAFs, destacam-se o *Wall Street Journal* (WSJ[58]), o Aurora[59] e o TIMIT[60]. Na Figura 7.8, são ilustrados os principais *corpora* utilizados. A categoria "Outros" contém os trabalhos que construíram o seu próprio *corpus*, utilizaram vídeos ou áudios disponíveis na *Internet*, programação de canais de televisão ou de rádio ou quando o *corpus* foi utilizado no máximo em 11 trabalhos. A categoria "Não informado" representa a quantidade de publicações que não informaram o *corpus* utilizado ou utilizaram um SRAF que não permite ser treinado. Os *corpora* utilizados no treinamento de SRAFs para a Língua Portuguesa do Brasil estão no Apêndice F.

[58] https://catalog.ldc.upenn.edu/ldc93s6a
[59] http://aurora.hsnr.de/index-2.html
[60] https://catalog.ldc.upenn.edu/ldc93s1

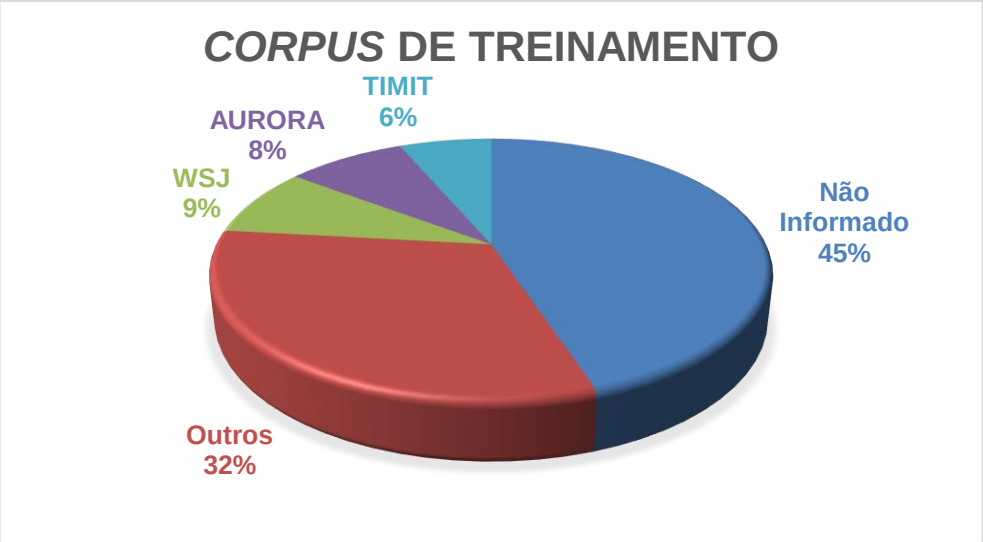

Figura 7.8: *Corpora* para realizar o treinamento dos Sistemas de Reconhecimento Automático de Fala.

O WSJ0 é um *corpus* para realizar o reconhecimento contínuo de fala com grande vocabulário, sendo desenvolvido com a leitura de textos de notícias do *Wall Street Journal*. Em versões posteriores, foram incorporadas outras fontes de notícias de negócios norte-americanas e de outros domínios. Os textos selecionados para leitura consistem em um subconjunto de 5.000 palavras ou 20.000 palavras do *corpus* de texto WSJ. Além da leitura dos textos, também foram incluídos alguns ditados espontâneos, coletados com a leitura de jornalistas usando artigos de notícias hipotéticas (LDC, 2017a). Já o WSJ1 é constituído por cerca de 78.000 pronúncias (73 horas de fala), das quais 4.000 são resultados do ditado espontâneo de jornalistas com diferentes níveis de experiência em ditado (LDC, 2017b).

O primeiro *corpus* do Aurora recebeu o nome de Aurora-2. O seu desenvolvimento se baseou no corpus TIDigits[61], contendo sequências de dígitos em inglês. Os ruídos foram adicionados artificialmente. A segunda base de dados, Aurora-3, é composta por sequência de dígitos gravadas em ambientes com ruído de carro, disponível nas Línguas Finlandesa, Italiana, Alemã, Espanhola e Dinamarquesa (Aurora, 2017).

Até o Aurora-3, era possível apenas o reconhecimento de dígitos. Para se tornar possível o treinamento de sistemas para o reconhecimento em tarefa de grande vocabulário, foi desenvolvido o Aurora-4, baseado nos dados contidos no WSJ. Diferentes sinais de ruído foram adicionados artificialmente. Para realizar o reconhecimento de fala em ambientes fechados, como quartos e interiores de veículos, foi desenvolvido o Aurora-5, no qual foi investigado o efeito da distorção em combinação com o ruído de fundo aditivo (Aurora, 2017).

Em relação ao TIMIT, ele é composto por gravações de 630 falantes com os oito principais dialetos do inglês americano, no qual, cada locutor gravou dez frases. O *corpus* TIMIT também inclui transcrições ortográficas e fonéticas, cujas transcrições foram verificadas manualmente (LDC, 2017c).

[61] https://catalog.ldc.upenn.edu/ldc93s10

Na Figura 7.9, é ilustrado o país de origem no qual foi desenvolvida a pesquisa das publicações. Para os trabalhos que foram realizados por pesquisadores de dois ou mais países, foi considerado, como país de desenvolvimento, o país do primeiro autor.

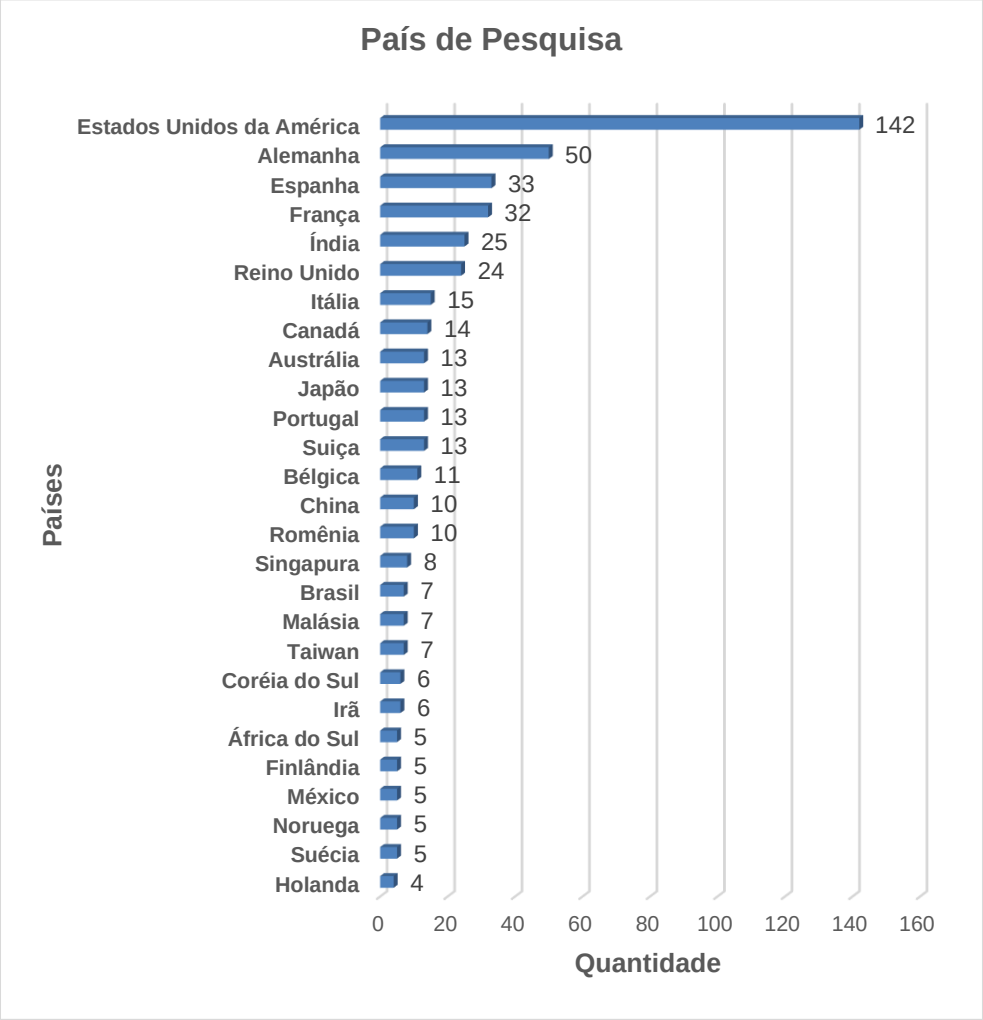

Figura 7.9: Quantidade de publicações por países.

Pode-se observar que os Estados Unidos da América foram responsáveis por 27,68% das produções de trabalhos nessa área. Países como Alemanha, Espanha, França, Índia e Reino Unido também apresentaram quantidades consideráveis de trabalhos.

A Argentina, Argélia, Bulgária, Egito, Eslováquia, Eslovênia, Irlanda, Letônia, Líbano, Nova Zelândia, Paraguai, Tailândia, Turquia e Vietnã publicaram um trabalho cada. A Arábia Saudita, Colômbia, Israel e República Tcheca contribuíram com duas publicações cada. Já a Tunísia contém uma publicação.

Na Figura 7.10 são apresentados os idiomas nos quais os SRAFs foram treinados, ou seja, qual o idioma a ser reconhecido pelo sistema. A Língua Inglesa se destaca, representando 64,52% dos SRAFs. Para os sistemas multilinguismo, 85,71% também contemplam o reconhecimento da Língua Inglesa. Outros idiomas de treinamento, que também se destacam, são as Línguas Espanhola, Francesa e Alemã. Os SRAFs treinados para a Língua Portuguesa do Brasil compreendem cinco publicações.

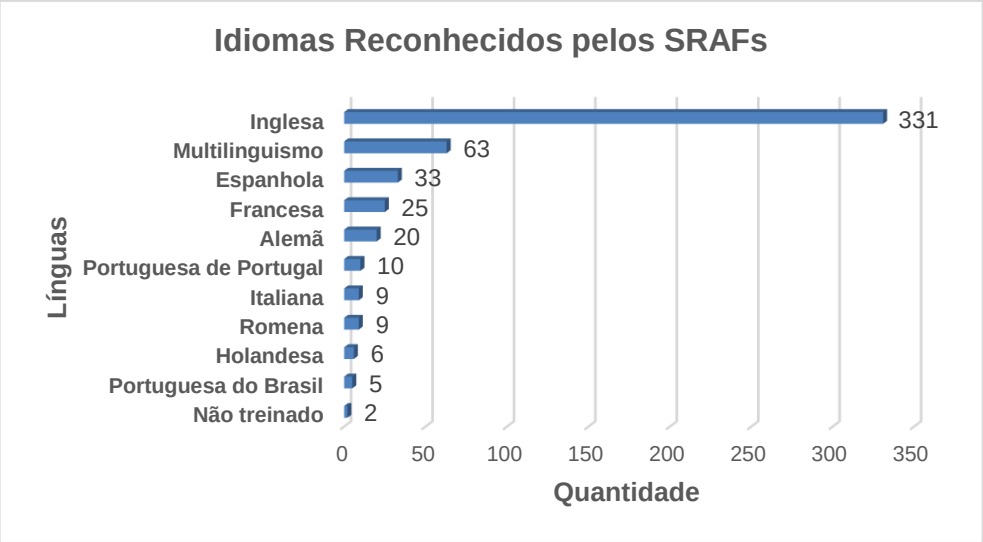

Figura 7.10: Quantidade de idiomas utilizados para o treinamento dos Sistemas de Reconhecimento Automático de Fala.

Com relação às duas publicações que não treinaram os sistemas, por exemplo, o Bavieca (Bolaños et al., 2013) desenvolveu um SRAF, no entanto, não o treinou. E no trabalho de Tsontzos & Orglmeister, (2011), foi desenvolvido um framework para que os pesquisadores de reconhecimento automático de fala possam usufruir da Arquitetura Orientada a Serviço – *Service-Oriented Architecture* (SOA) – em seus sistemas. O SRAF-base desse framework é o CMU Sphinx-4.

7.1.2 Discussão dos Trabalhos Selecionados da Revisão Sistemática

Os trabalhos selecionados foram divididos em grupos e subgrupos. Os subgrupos foram criados para englobar os diferentes métodos desenvolvidos relacionados à tecnologia de reconhecimento de fala.

Na Figura 7.11 é ilustrada a quantidade de trabalhos por grupo, cujas descrições são:

- **Grupo 1:** trabalhos que desenvolveram método para SRAFs;
- **Grupo 2:** sistemas para legendagem ou transcrição; pesquisa de documentos, áudios ou vídeos por fala; tradução da fala; controle residencial ou em central de atendimento; e sistemas de diálogo falado, cujo foco é a interação do sistema com o usuário;
- **Grupo 3:** sistemas para avaliação automática da qualidade da fala; estimar a idade do falante; estudar o sotaque; detectar algum tipo de doença da fala, como a disartria; detectar sentimentos ou avaliar ou reduzir a carga cognitiva durante a fala; e sistemas para uso hospitalar;
- **Grupo 4:** *framework* para desenvolvimento de SRAFs.

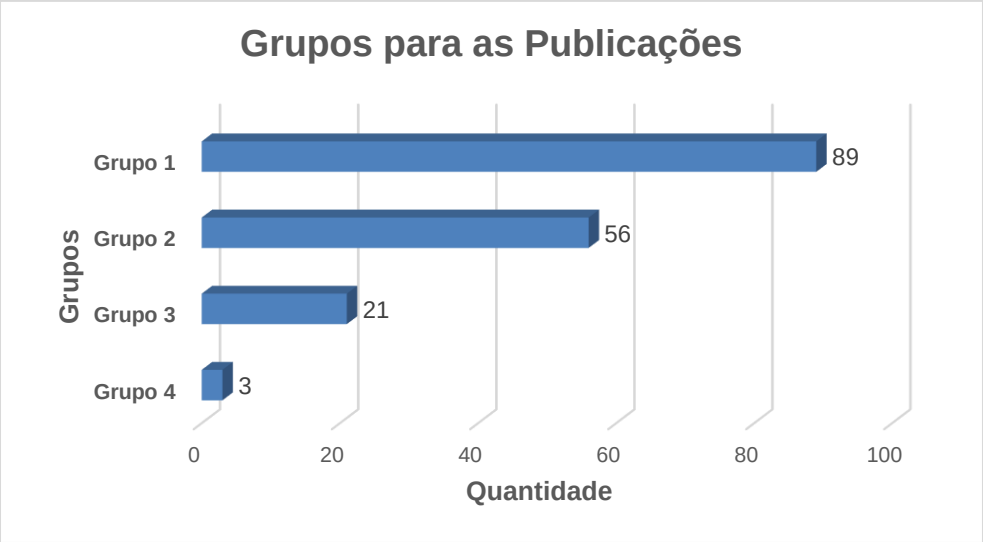

Figura 7.11: Grupos com as categorias das publicações.

O Grupo 1 correspondeu a 52,66% dos trabalhos selecionados, englobando a maioria das publicações. Esse grupo foi criado para se obter uma visão geral sobre algumas pesquisas realizadas na área.

O Grupo 2 (33,14%) compreende os trabalhos que utilizaram a tecnologia de reconhecimento automático de fala para propósito geral. Os SRAFs também são frequentemente utilizados para realizar transcrição e legendagem, como em programações de televisão ou de vídeos da *Internet* e também para apoiar pesquisa de dados digitais.

Com relação à interação entre os seres humanos com dispositivos eletrônicos, os SRAFs podem ser utilizados para agregar funcionalidades a aplicações de uso diverso e também para o controle residencial ou em central de atendimento. Com relação à tradução entre diferentes idiomas, são contemplados os sistemas com enfoque para auxiliar na comunicação entre diversas culturas. Esse grupo também compreende as publicações cujo enfoque é o desenvolvimento de aplicações de interação entre seres humanos com sistemas computacionais, em que o SRAF reconhece a fala do usuário e responde, por meio da síntese de voz, a algum questionamento realizado por ele.

Já o Grupo 3 corresponde a SRAFs para a avaliação da qualidade ou de alguma característica da fala. Esse grupo compreende um total de 12,43% das publicações selecionadas.

Por fim, o Grupo 4 compreendeu a aproximadamente 0,02% das publicações selecionadas. O objetivo desse grupo foi o de identificar o desenvolvimento de novos *frameworks* para apoiar a construção de SRAFs.

O Grupo 1 foi dividido em sete subgrupos, com cada um representando um método desenvolvido referente a alguma tecnologia do SRAF (Figura 7.12). Os subgrupos são:

- **Subgrupo 1:** método para melhoria da Extração de Características, por exemplo, suprimir o ruído, aprimorar o reconhecimento em ambientes com reverberação ou diminuir a distorção da fala durante a transmissão pelo canal acústico;
- **Subgrupo 2:** treinamento do Modelo Acústico (MA) para o SRAF;

- **Subgrupo 3:** treinamento do Modelo de Linguagem (ML) para o SRAF;
- **Subgrupo 4:** técnica para otimizar a construção de um Léxico;
- **Subgrupo 5:** método para melhoria do Decodificador de um SRAF;
- **Subgrupo 6:** palavra fora do vocabulário de treinamento de um SRAF;
- **Subgrupo 7:** pós-processamento para melhorar a saída de transcrição gerada por um SRAF, ou seja, para resolver o problema referente aos erros ortográficos.

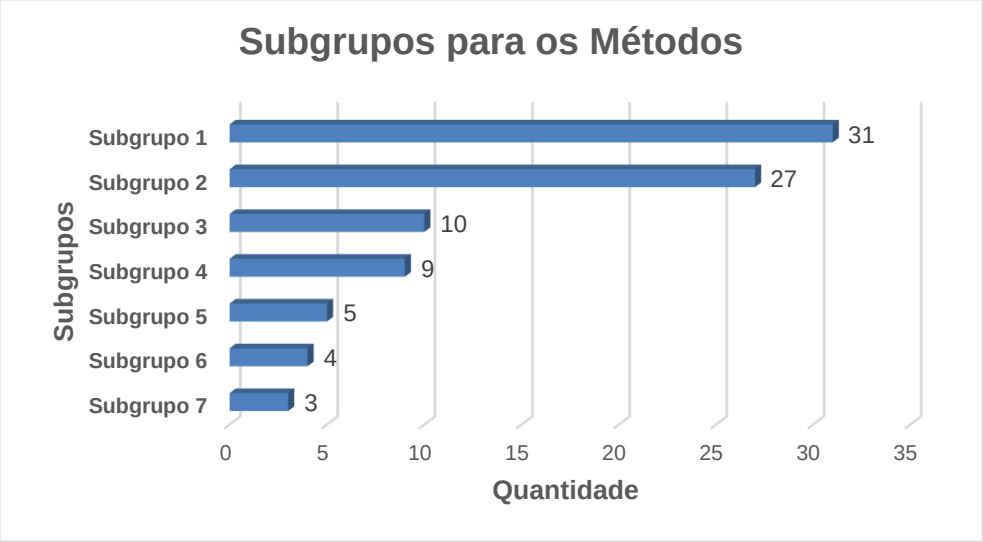

Figura 7.12: Subgrupos do Grupo 9, referente ao desenvolvimento de métodos para os Sistemas de Reconhecimento Automático de Fala.

Analisando o gráfico da Figura 7.12, pode-se perceber que a grande maioria dos trabalhos teve como objetivo o desenvolvimento de métodos para aperfeiçoar a Extração de Características ou o MA.

A melhoria na técnica de Extração de Características, Subgrupo 1, correspondendo a 34,83%. Já o aperfeiçoamento do MA (Subgrupo 2) compreendeu a 30,34% das publicações.

Com relação aos subgrupos para o aperfeiçoamento do ML (Subgrupo 3), de técnica para aprimorar a construção de um Léxico (Subgrupo 4) e para melhorar o algoritmo do Decodificador (Subgrupo 5), corresponderam a 11,23%, 10,11% e 5,62%, respectivamente.

Outros trabalhos apresentaram métodos relacionados à melhoria do reconhecimento de palavras fora do vocabulário (Subgrupo 6) compreendeu a 4,49% e pós-processamento (Subgrupo 7) equivale a 3,37% das publicações.

7.2 Sistemas de Reconhecimento Automático de Fala Avaliados em um Experimento Preliminar

Com o intuito de realizar uma avaliação prévia sobre os SRAFs disponíveis para a Língua Portuguesa do Brasil foram selecionados sete sistemas. Esses sistemas foram

avaliados por meio da taxa WER. Na Figura 7.13 é ilustrado o gráfico combinado com a média e o desvio padrão das taxas WERs dos dez voluntários para cada SRAF.

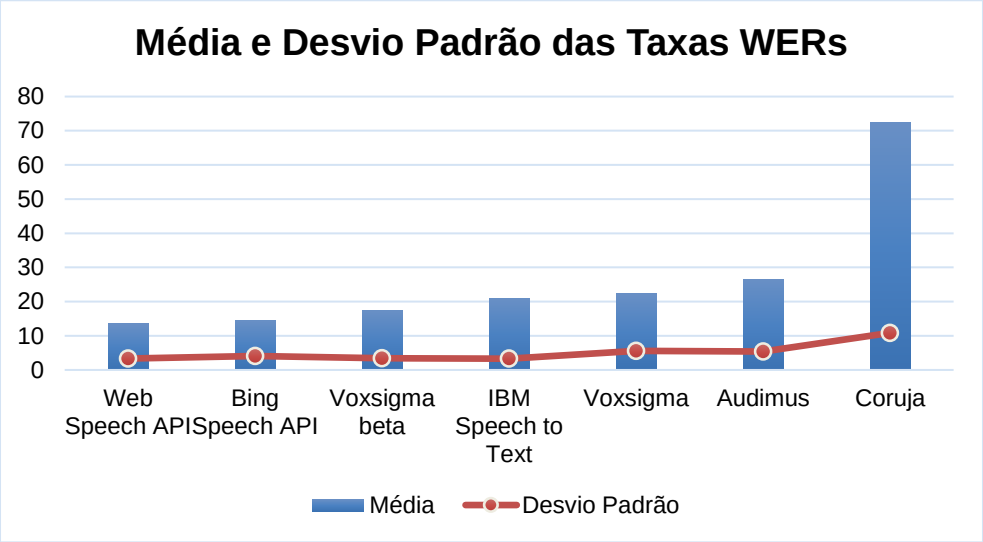

Figura 7.13: Gráfico combinado com a média e o desvio padrão das taxas de erro de palavra – *Word Error Rate* (WER) – dos dez voluntários para cada SRAF.

Na Tabela 7.1 são apresentados os resultados da WER para cada voluntário, cuja idade média foi de 26,2 anos com desvio padrão (DP) de 7,98. Nas três últimas linhas, em itálico, dessa tabela são mostrados a mediana, a média e o respectivo DP para cada SRAF.

Tabela 7.1: Taxas de erro de palavra – *Word Error Rate* (WER) – (%) com os resultados individuais dos voluntários para os Sistemas de Reconhecimento Automático de Fala avaliados (Masc. = Masculino e Fem. = Feminino).

Voluntários	***Web Speech* API**	**Bing *Speech* API**	**Voxsigma beta**	**IBM *Speech to Text***	**Voxsigma**	**Audimus**	**Coruja**
Masc. 1	11,34	11,18	15,88	21,39	17,83	27,07	64,67
Masc. 2	15,40	16,37	19,12	26,58	23,01	27,39	83,79
Masc. 3	14,10	14,75	18,48	21,07	24,80	30,31	66,45
Masc. 4	7,94	12,80	17,50	17,99	21,56	26,09	66,29
Masc. 5	8,10	6,81	13,78	15,40	14,59	15,56	63,53
Fem. 1	17,99	20,91	18,64	19,93	22,37	28,20	82,17
Fem. 2	15,88	15,23	18,96	22,37	28,36	33,55	80,71
Fem. 3	14,59	12,48	18,48	19,93	27,07	31,12	85,25
Fem. 4	15,23	19,45	22,69	25,44	29,66	27,39	76,82
Fem. 5	14,75	15,56	10,21	19,61	13,29	19,45	53,16
Mediana	*14,67*	*14,99*	*18,48*	*20,50*	*22,69*	*27,39*	*71,63*
Média	*13,53*	*14,55*	*17,37*	*20,97*	*22,25*	*26,61*	*72,28*
DP	*3,34*	*4,05*	*3,40*	*3,29*	*5,59*	*5,37*	*10,85*

Os dados da Tabela 7.1, são quantitativos, contínuos e pareados, pois, a partir de uma amostra de cada indivíduo obteve-se diversos resultados para diferentes tratamentos. Na Figura 7.14 é ilustrado o diagrama de caixa – *boxplot* – dos SRAFs avaliados. Cada caixa representa um sistema. O "x" dentro da caixa representa a média e a linha horizontal a mediana.

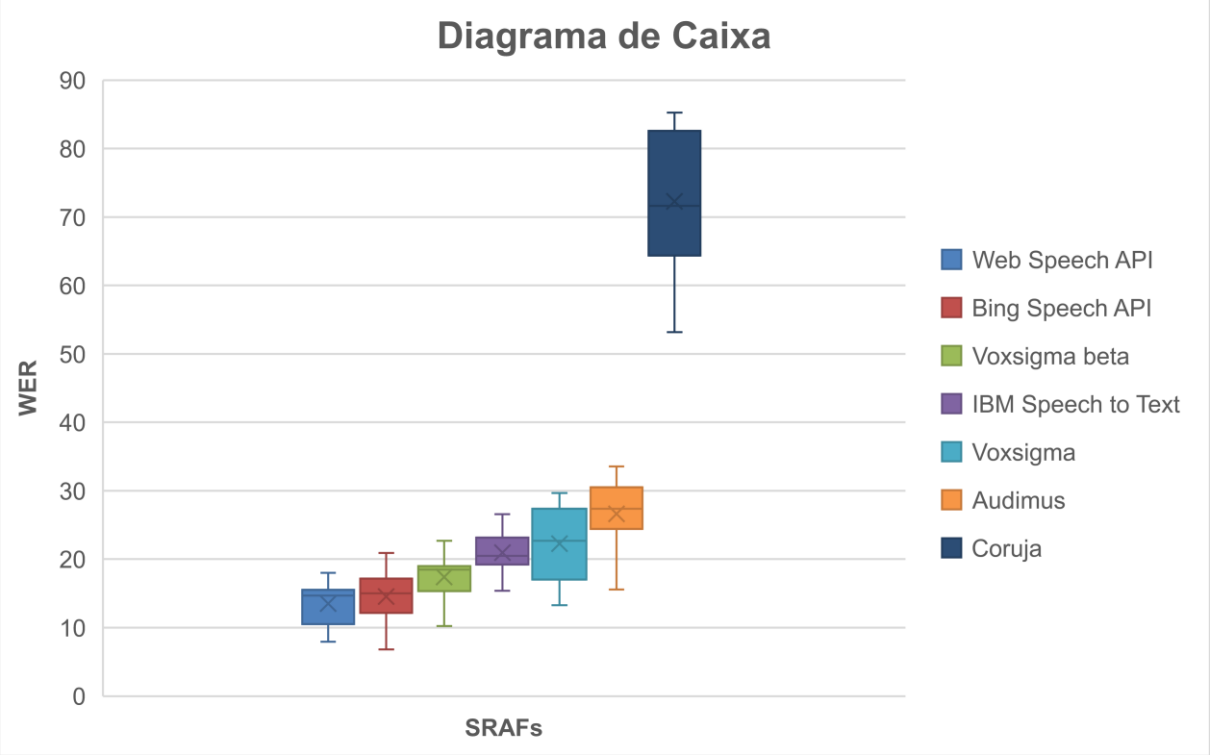

Figura 7.14: Diagrama de Caixa dos Sistemas de Reconhecimento Automático de Fala avaliados no experimento preliminar.

Conforme ilustrado na Figura 7.14, pode-se verificar que o Voxsigma beta apresentou a menor variabilidade. Já o Coruja apresentou a maior variabilidade, seguido pelo Voxsigma. Os limites inferiores do Coruja e do Voxsigma apresentaram a maior distância até o primeiro quartil.

Com relação à análise estatística, o resultado do teste de normalidade dos dados, constatou que o grupo do *Web Speech* API não apresenta uma distribuição normal, pois o *p*-valor foi de 0,041, não sendo significativo para ser considerado uma distribuição normal.

Como a distribuição dos dados não é normal, logo os dados são não-paramétricos. Dessa maneira, foi aplicado o teste de Friedman para a comparação entre os grupos, em que cada grupo corresponde a um SRAF. O *p*-valor desse teste foi de 0,0001, portanto existe diferença em comparação entre pelo menos dois grupos.

Para a verificação de quais grupos se diferem entre si, foi aplicado o pós-teste de Dunn, cujas comparações são apresentadas na Tabela 7.2. As linhas em itálico e em negrito dessa tabela representam os grupos em que houve diferença significativa.

Tabela 7.2: Comparação entre os grupos dos Sistemas de Reconhecimento Automático de Fala avaliados (vs. = *versus*).

Comparação	***p*-valor**
Web Speech API vs. Bing *Speech* API	$p > 0,05$
Web Speech API vs. Voxsigma beta	$p > 0,05$
Web *Speech* API vs. IBM *Speech to Text*	$p > 0,05$
Web *Speech* API vs. Voxsigma	$p > 0,05$
Web Speech API vs. Audimus	***p < 0,001***
Web Speech API vs. Coruja	***p < 0,001***
Bing *Speech* API vs. Voxsigma beta	$p > 0,05$
Bing *Speech* API vs. IBM *Speech to Text*	$p > 0,05$
Bing *Speech* API vs. Voxsigma	$p > 0,05$
Bing Speech API beta vs. Audimus	***p < 0,01***
Bing Speech API vs. Coruja	***p < 0,001***
Voxsigma beta vs. IBM *Speech to Text*	$p > 0,05$
Voxsigma beta vs. Voxsigma	$p > 0,05$
Voxsigma beta vs. Audimus	***p < 0,05***
Voxsigma beta vs. Coruja	***p < 0,001***
IBM *Speech to Text* vs. Voxsigma	$p > 0,05$
IBM *Speech to Text* vs. Audimus	$p > 0,05$
IBM *Speech to Text* vs. Coruja	$p > 0,05$
Voxsigma vs. Audimus	$p > 0,05$
Voxsigma vs. Coruja	$p > 0,05$
Audimus vs. Coruja	$p > 0,05$

Conforme pode ser verificado na Tabela 7.2, o *Web Speech* API, o Bing *Speech* API e o Voxsigma beta apresentaram desempenho estatisticamente superior em relação aos SRAFs Audimus e Coruja. As demais comparações entre os grupos não apresentaram diferença estatisticamente significativa. Ou seja, os SRAFs *Web Speech* API, Bing *Speech* API, Voxsigma beta, IBM *Speech to Text* e Voxsigma são estatisticamente iguais, não podendo concluir que um apresenta desempenho superior em relação ao outro.

Neste trabalho, o Coruja apresentou uma alta taxa WER, muito superior aos valores relatados na dissertação de Silva (2010), que foi de 32,87%, sem menção ao DP. Em outro trabalho, o Coruja apresentou WER de 29% (Neto et al., 2011). Já no trabalho de Moura et al. (2010) foi relatado taxa de acerto igual a 71% para o teste independente de locutor e de 60,42% de taxa de palavras corretas também independente de locutor (Silva et al., 2010b).

O Coruja foi treinado com os *corpora* de áudio Spoltech e LapsStory e avaliado com o *corpus* LapsBenchmark, disponíveis na página do grupo FalaBrasil[62], da Universidade Federal do Pará. O Spoltech conta com aproximadamente 4 horas de voz e o LapsStory é composto por cinco livros falados – *audiobooks* – com cerca de 1 hora de duração cada e mais 15 horas e 42 minutos de áudios. Com relação ao ML, o *corpus* de texto do Coruja foi treinado com aproximadamente 2 milhões e 346 mil frases. Todas essas frases correspondem a um total de cerca de 41 milhões e 348 mil palavras.

[62] http://www.laps.ufpa.br/falabrasil/downloads.php

Segundo o autor, os testes do Coruja foram realizados com o *corpus* LapsBenchmark, composto por 35 locutores, sendo 25 homens e 10 mulheres que pronunciaram 20 frases, correspondendo a aproximadamente 54 minutos de áudio. O Coruja também foi utilizado na dissertação de Batista (2013), no qual foi reportado taxa WER de 39,57%, sem menção ao DP.

Para tentar identificar a possível causa dessa disparidade, foram feitos contatos com os autores, cujas respostas obtidas foram a de que os testes conduzidos foram descritos em seus trabalhos, ou seja, utilizando o *corpus* LapsBenchmark.

Essa disparidade entre a WER relatada nos trabalhos de Silva (2010) e de Batista (2013) com os resultados obtidos neste experimento pode ser devido à complexidade do texto utilizado, o qual é composto por palavras específicas do domínio médico, e que pode não ter sido contemplado durante o treinamento do MA e do ML do Coruja.

O escasso treinamento do ML do Coruja pode ser constatado quando comparado com o ML utilizado para a busca de voz da Google, cujo sistema foi treinado com cerca de 230 bilhões de palavras, utilizando 1 milhão de palavras únicas (Schalkwyk et al., 2010).

Além do escasso treinamento do ML, o treinamento do MA também pode ser uma deficiência do Coruja, já que em comparação com o trabalho de Jaitly et al. (2012), conduzido por pesquisadores da Google em parceria com a Universidade de Toronto, o MA foi treinado com cerca de 5.870 horas de pesquisa de voz e mais 1.400 horas de dados de áudio do YouTube.

Com base nos resultados dessa avaliação preliminar, optou-se por selecionar os SRAFs *Web Speech* API e Bing *Speech* API para realizar experimentos mais detalhados, pois esses sistemas apresentaram os melhores desempenhos, bem como, permitem atender a grande maioria de diferentes dispositivos móveis e sistemas operacionais. A não inclusão das duas versões do Voxsigma e do Audimus se deve ao fato de se tratarem de sistemas que requerem licença de utilização. O IBM *Speech to Text* não foi considerado, pois o Bing *Speech* API atende a necessidade de o PSW funcionar em dispositivos móveis e navegadores de *Internet* não desenvolvidos pela Google. Com relação ao Coruja, não foi considerado devido ao seu desempenho inferior em relação ao*s* demais SRAs avaliados.

7.3 Avaliação dos Sistemas de Reconhecimento Automático de Fala da Google *Web Speech* API e da Microsoft Bing *Speech* API

Para a avaliação do experimento final, foram considerados 30 voluntários (Seção 5.2), com idade média de 26,7 anos e DP de 7,6 anos. Nesse conjunto, também foram considerados os dez voluntários do experimento preliminar que compreendem as dez primeiras linhas dos voluntários da Tabela 7.3. Nessa tabela é apresentada a taxa WER para cada voluntário

referente à avaliação dos SRAFs *Web Speech* API e Bing *Speech* API. As linhas em itálico representam a mediana, a média e o respectivo DP.

Tabela 7.3: Taxas de erro de palavra – *Word Error Rate* (WER) – (%) com os resultados individuais dos voluntários para os Sistemas de Reconhecimento Automático de Fala avaliados (Masc. = Masculino e Fem. = Feminino).

Voluntários	***Web Speech* API**	**Bing *Speech* API**
Masc. 1	11,34	16,70
Masc. 2	15,40	23,34
Masc. 3	14,10	19,29
Masc. 4	7,94	18,64
Masc. 5	8,10	12,15
Fem. 1	17,99	19,29
Fem. 2	15,88	24,15
Fem. 3	14,59	19,29
Fem. 4	15,23	18,48
Fem. 5	14,75	18,64
Masc. 6	13,94	22,53
Masc. 7	5,83	10,21
Masc. 8	14,75	21,55
Masc. 9	11,02	13,61
Masc. 10	15,40	21,55
Masc. 11	16,37	20,42
Masc. 12	9,72	17,02
Masc. 13	12,15	20,42
Masc. 14	9,24	13,78
Masc. 15	17,82	23,01
Fem. 6	10,05	12,15
Fem. 7	11,51	16,04
Fem. 8	5,02	12,64
Fem. 9	12,80	18,31
Fem. 10	10,37	14,42
Fem. 11	10,21	13,29
Fem. 12	9,40	16,53
Fem. 13	21,07	23,18
Fem. 14	10,37	16,04
Fem. 15	6,81	13,61
Mediana	*11,83*	*18,395*
Média	*12,30*	*17,68*
DP	*3,83*	*3,90*

Os dados da Tabela 7.3 são quantitativos, contínuos e pareados, cujo *p*-valor para verificar a normalidade dos dados para ambos os grupos, foi superior a 0,10, portanto, constatou-se que os dados dos grupos apresentam distribuição normal.

Para a comparação entre os dois grupos, Google *Web Speech* API e Microsoft Bing *Speech* API, foi aplicado o teste *t* pareado, que obteve *p*-valor < 0,0001. Portanto, existe

diferença estatisticamente significativa entre a taxa WER do *Web Speech* API em comparação com o Bing *Speech* API. Essa diferença pode ser verificada na Figura 7.15, em que é apresentado o gráfico comparativo entre a taxa WER de ambos os SRAFs.

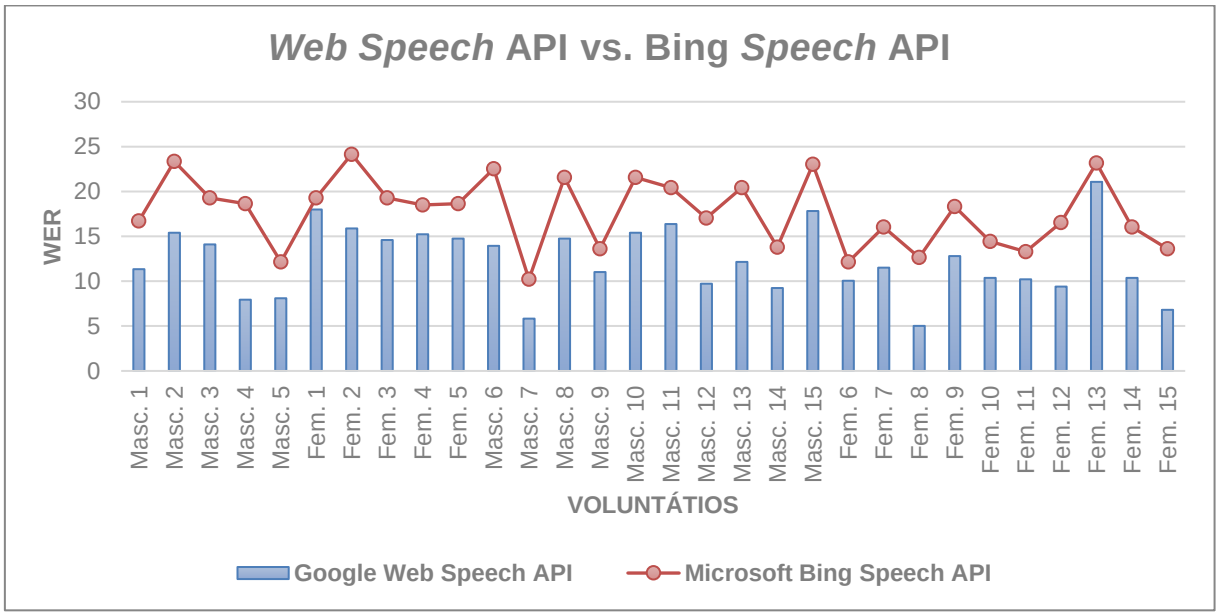

Figura 7.15: Gráfico combinado com as taxas de erro de palavra – *Word Error Rate* (WER) – do Google *Web Speech* API e do Microsoft Bing *Speec* API (vs. = versus).

Conforme pode ser verificado na Figura 7.15, nota-se que o Google *Web Speech* API obteve taxas WERs inferiores às taxas do Microsoft Bing *Speech* API. Apenas as taxas WERs das voluntárias Fem. 1 e Fem. 13 foram próximas, no entanto, as taxas do Google continuaram inferiores.

As menores taxas WERs para o grupo masculino do *Web Speech* API foram de 5,83% para o voluntário Masc. 7 e 7,94% para o Masc. 4. Já para o Bing *Speech* API foram de 10,21% e 13,61% para os voluntários Masc. 7 e Fem. 9, respectivamente. Para o grupo feminino, as menores taxas WER do *Web Speech* API foram das voluntárias Fem. 8 e Fem. 15 que obtiveram 5,02% e 6,81%, respectivamente. Já para o Bing *Speech* API, as menores taxas WERs foram de 11,34% (Fem. 8) e 12,15% (Fem. 6).

O voluntário Masc. 7 apresentou a menor taxa WER para o SRAF da Microsoft e a segunda menor WER para o SRAF da Google. Por outro lado, a voluntária Fem. 13 apresentou a pior precisão no SRAF da Google e a terceira pior precisão no SRAF da Microsoft. Para esses dois voluntários é possível que uma ou mais características das suas vozes contribuíram para influenciar diretamente na precisão dos SRAFs. Como o foco deste trabalho não é o de encontrar possíveis fatores que possam influenciar na precisão desses sistemas, futuramente podem ser estudadas possíveis características da voz que mais impactam na taxa WER.

Na Figura 7.16 é ilustrado o histograma com as taxas WERs do Google *Web Speech* API.

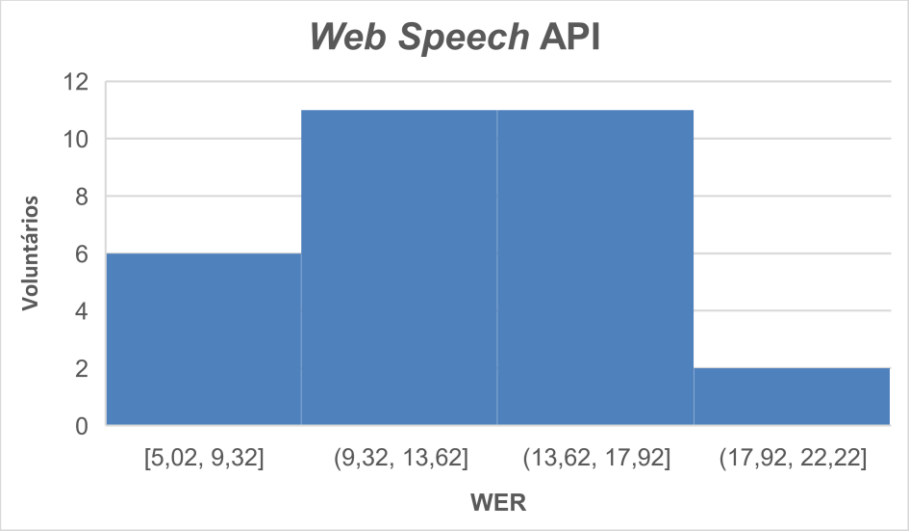

Figura 7.16: Histograma com as taxas de erro de palavra – *Word Error Rate* (WER) – do Google *Web Speech* API.

Analisando o histograma da Figura 7.16, pode-se perceber que o Google *Web Speech* API apresentou seis taxas WERs inferiores a 9,40%. Com relação às maiores taxas WERs, obteve-se duas médias, sendo de 17,99% para o Fem. 1 e de 21,07% para o Fem. 13. No geral, 11 médias ficaram no intervalo entre 9,40% a 12,80% e outras 11 médias entre 13,94% e 17,82%.

Na Figura 7.17 é ilustrado o histograma com as taxas WERs do Microsoft Bing *Speech* API.

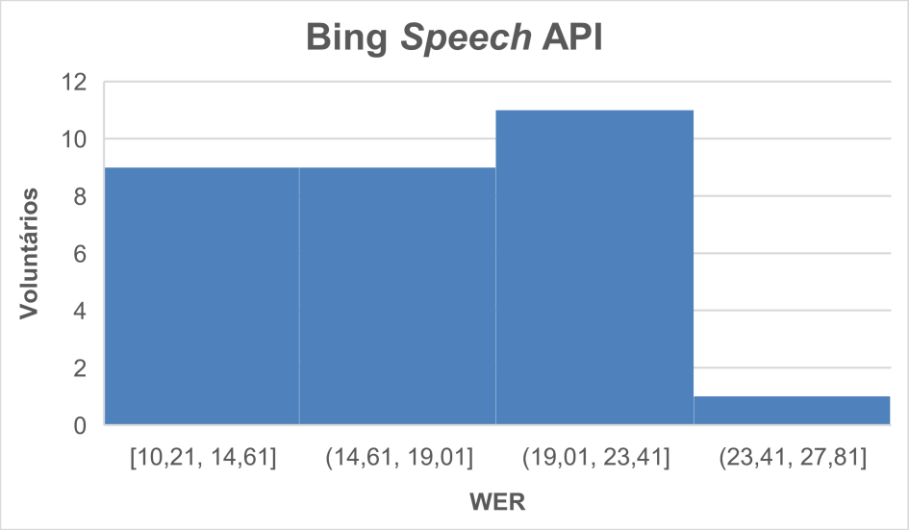

Figura 7.17: Histograma com as taxas de erro de palavra – *Word Error Rate* (WER) – do Microsoft Bing *Speech* API.

Para o Bing *Speech* API, nove taxas WERs ficaram entre 10,21% e 13,78% e outras oito médias entre 14,42% e 18,48%. A maioria das médias, correspondendo a 11 taxas WERs, ficou entre 19,29% e 23,34% e uma taxa superior a 24%.

Pode-se perceber que o *Web Speech* API apresentou oito taxas WERs inferiores a 10% (Fem. 8, Masc. 7, Fem. 15, Masc. 4, Masc. 5, Masc. 14, Fem. 12 e Masc. 12) e uma taxa WER superior a 20% (Fem. 13). O Bing *Speech* API não apresentou nenhuma taxa WER inferior a 10% e nove taxas WERs superiores a 20% (Masc. 11, Masc. 13, Masc. 8, Masc. 10, Masc. 6, Masc. 15, Fem. 13, Masc. 2 e Fem. 2).

Ao considerar a menor taxa WER do *Web Speech* API (5,02%) e a maior (21,07%), a variação foi de 16,05%. Já para o Bing *Speech* API, a variação foi menor, sendo de 13,94%, com a menor taxa WER de 10,21% e a maior WER de 24,15%.

A taxa média do *Web Speech* API foi de 12,30% correspondendo a 16 voluntários com taxa WER inferior à sua média, sendo oito voluntários (Masc. 7, Masc. 4, Masc.5, Masc. 14, Masc. 12, Masc. 9, Masc.1 e Masc. 13) e oito voluntárias (Fem. 8, Fem. 15, Fem. 12, Fem. 6, Fem. 11, Fem. 10, Fem. 14 e Fem. 7).

O Bing *Speech* API obteve taxa média de 17,68%, que totalizaram 15 voluntários com taxa WER inferior à sua média. Desse número, nove são voluntárias (Fem. 3, Fem. 6, Fem. 8, Fem. 11, Fem. 15, Fem. 10, Fem. 14, Fem. 7 e Fem. 12) e seis voluntários (Masc. 1, Masc. 5, Masc. 7, Masc. 9, Masc. 14 e Masc. 12).

7.4 Apresentação do Protótipo de Sistema *Web*

Nesta seção é apresentado o PSW para a geração de laudos médicos por meio do reconhecimento da fala. Os SRAFs integrados compreendem o Google *Web Speech* API e o Microsoft Bing *Speech* API.

O SRAF *Web Speech* API funciona exclusivamente nos serviços oferecidos pela Google, como em aparelhos portáteis (*smartphones* e *tablets*) com sistema operacional Android, em *desktops* ou *notebooks* com sistema operacional Chrome OS e no navegador de *Internet* Chrome.

Sendo assim, para que o PSW possa ser utilizado em dispositivos portáteis com sistemas operacionais WindowsPhone e em navegadores como *Edge* e Mozilla Firefox, o Bing *Speech* API se torna uma opção adequada.

O Bing *Speech* API foi integrado em sua versão *Websocket* com o modo interativo. A sua escolha foi baseada devido a esse modo permitir o retorno de resultados parciais durante o reconhecimento. Para que o reconhecimento seja contínuo, a API foi modificada para ser reiniciada a cada final de seguimento de áudio enviado ao servidor da Microsoft. Essa adaptação foi necessária, pois, até o momento da escrita desta dissertação, o modo ditado não retornava resultados parciais das transcrições, e, portanto, não permitia a exibição do texto dos resultados parciais de reconhecimento da fala.

De acordo com dados publicados pela W3Schools, os sistemas operacionais mais utilizados em agosto de 2017 foram: Windows 10 (36,90%), Windows 8 (9,30%), Windows 7 (29,80%), Windows Vista (0,10%), Windows XP (0,70%), Linux (6%), Mac (9,90%), Chrome OS (0,20%) (W3Schools, 2017a) e os dispositivos portáteis, correspondendo a 7,20%

do total dos sistemas operacionais, estando distribuídos da seguinte maneira: iOS (1,30%), Android (5,64%), Windows (0,19%) e outros (0,07%) (W3Schools, 2017b).

Com relação ao uso de navegadores de *Internet*, a utilização estava dividida da seguinte maneira: Chrome (76,90%), Mozilla Firefox (13,10%), *Edge* e *Internet Explorer* (4,30%), Safari (3%) e o Opera (1,20%) (W3Schools, 2017c). Assim, a porcentagem de usuários que o PSW atende atualmente pode ser estimada em aproximadamente 87%, considerando os acessos proveniente dos navegadores de *Internet* Chrome e *Edge* e dos dispositivos móveis com sistemas operacionais Android e Windows.

Na Figura 7.18 é ilustrada a tela inicial do PSW, em que é necessário a autenticação para ter acesso às funcionalidades do sistema.

Figura 7.18: Tela de autenticação.

Após a autenticação bem-sucedida do usuário, de acordo com o nível de acesso, podendo ser usuário administrador ou usuário comum, o sistema permite o acesso, conforme é ilustrado na Figura 7.19, em que o usuário é direcionado para a página principal com a exibição dos exames cadastrados no sistema.

O menu gerenciar (a) contém as opções para acessar o gerenciamento dos laudos médicos e do gerenciamento dos exames; o menu cadastrar (b) contém os cadastros de exame, paciente, tipo de exame, profissional e especialidade; e o menu ajuda (c) contém um manual de auxílio para a utilização do PSW.

Quando o usuário autenticado não for pertencente ao grupo de administradores, o ícone para acessar o gerenciamento de profissionais (d) não é exibido, o ícone (e) permite acesso à tela inicial do PSW e o ícone (f) permite ao usuário sair da aplicação.

Figura 7.19: Tela inicial do Protótipo de Sistema *Web*.

Na Figura 7.20 é ilustrada a tela de gerenciamento de exames, em que é possível realizar buscas dos exames cadastrados por código do exame (a), nome do paciente (b), nome do profissional (c), data de realização do exame (d). Nessa tela também é possível editar o exame (e), exibir o exame (f), gerar um laudo médico com o SRAF da Google (g) ou da Microsoft (h).

Figura 7.20: Tela de gerenciamento de exames.

Na Figura 7.21 é ilustrada a tela para cadastrar um exame.

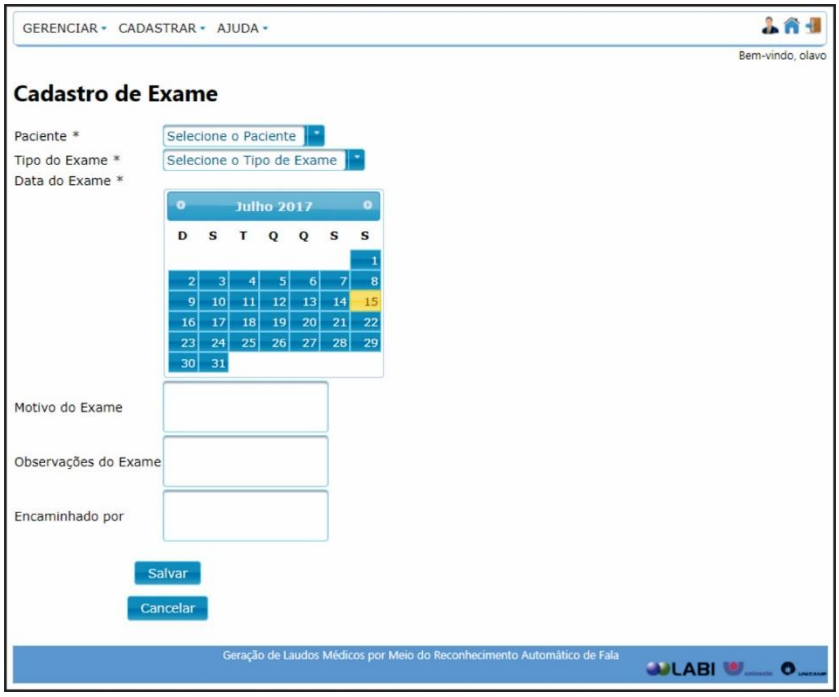

Figura 7.21: Tela de cadastro de exame.

Na Figura 7.22 é ilustrada a tela para cadastrar um paciente. O campo para preenchimento do Cadastro de Pessoa Física (CPF) possui um validador para verificar se o CPF inserido é válido.

Todos os campos das telas de cadastros com asterisco (*) são de preenchimento obrigatório.

Figura 7.22: Tela de cadastro de paciente.

Na Figura 7.23 é ilustrada a tela para cadastrar um profissional. O campo CPF também possui validação. Os campos que diferem do cadastro do paciente são: número do Conselho Regional de Medicina (CRM), especialidade, *e-mail*, *login* e senha de acesso ao PSW.

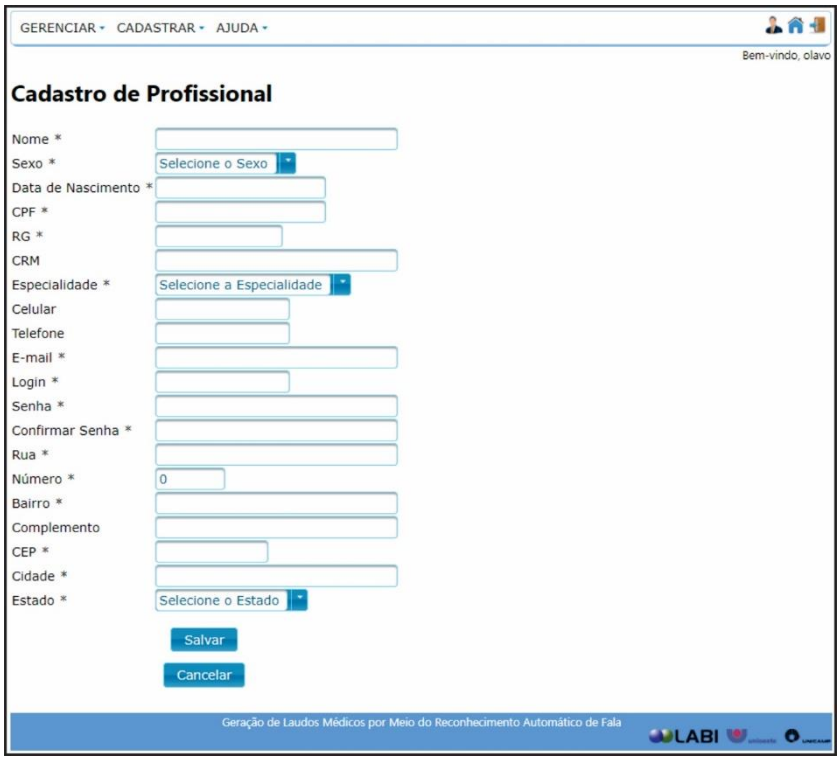

Figura 7.23: Tela de cadastro de profissional.

Na Figura 7.24A é ilustrada a tela de cadastro da especialidade médica e na Figura 7.24B, o cadastro do tipo de exame.

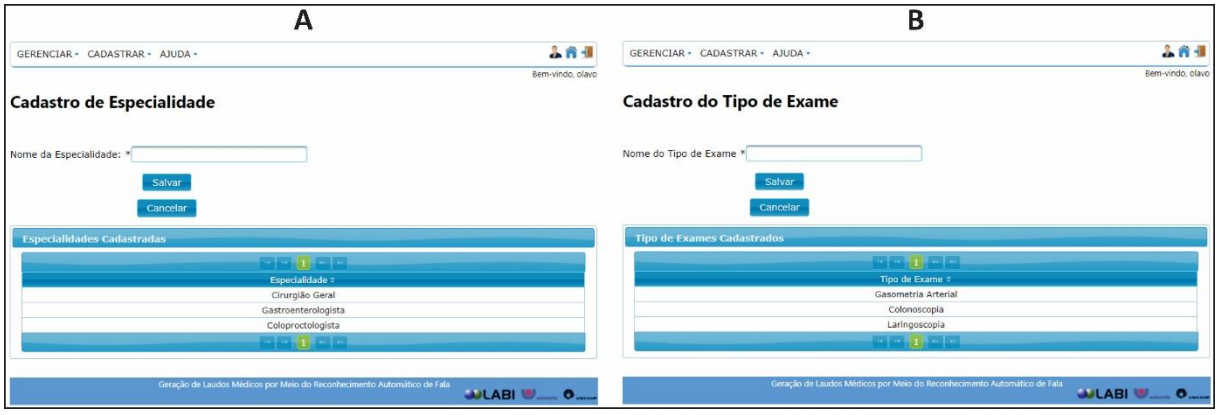

Figura 7.24: (A) Tela de cadastro de especialidade; e (B) Tela de cadastro do tipo de exame.

Na Figura 7.25 é ilustrada a tela de exibição do exame, no qual são apresentadas as informações do exame (a) e todos os laudos médicos gerados para esse exame (b). Nesse exemplo, foram gerados dois laudos médicos, com cada laudo podendo ser acessado pelas guias “1” e “2” na região (b).

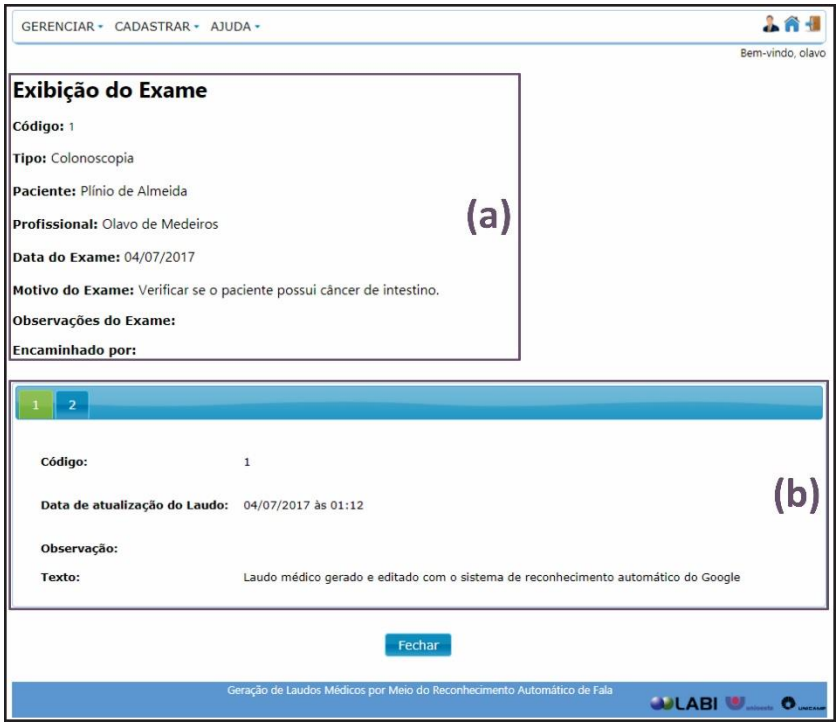

Figura 7.25: Tela de exibição do exame.

Na Figura 7.26 é ilustrada a tela para gerar um laudo médico com o SRAF da Google. A exibição desse laudo médico atualizado é ilustrada na região (a) da Figura 7.29. Quando o usuário pressionar o botão “Iniciar” (a) é possível iniciar o ditado do laudo médico. Quando o ditado é iniciado, a imagem do microfone é alterada para a imagem do microfone sem o “X” (b). No momento do ditado, a hipótese de transcrição preliminar é exibida em (c).

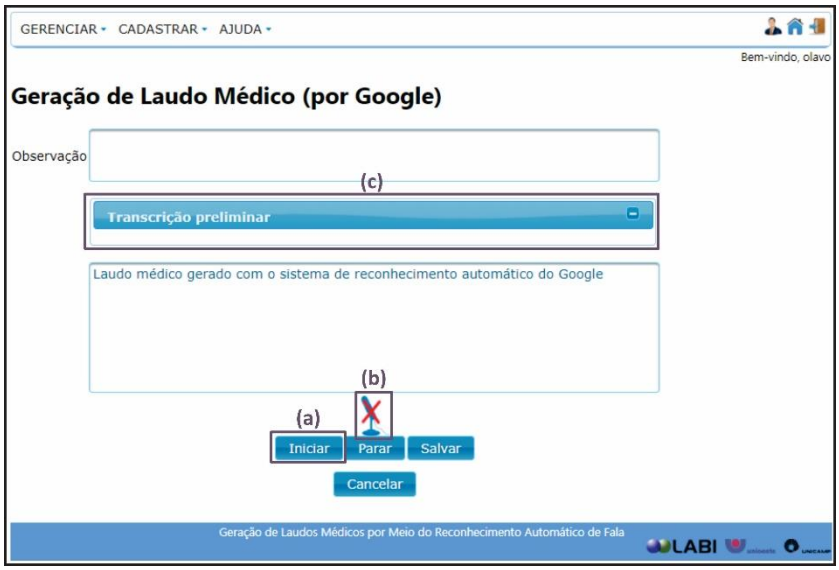

Figura 7.26: Tela de geração de laudo médico com o Sistema de Reconhecimento Automático de Fala da Google.

Na Figura 7.27 é ilustrada a tela para gerar um laudo médico com o SRAF da Microsoft. Para iniciar o ditado é necessário pressionar o botão “Ativar” para carregar a API correspondente. Na sequência, a sua operação é similar ao SRAF da Google (Figura 7.26).

Figura 7.27: Tela de geração de laudo médico com o Sistema de Reconhecimento Automático de Fala da Microsoft.

Na Figura 7.28 é ilustrada a tela para exibir um laudo médico, na qual é apresentada a última versão, ou seja, o último laudo médico editado pelo usuário. A tela para a exibição do exame pode ser acessada através do botão Exibir das Figuras 7.19 ou 7.20.

Figura 7.28: Tela de exibição de um laudo médico.

Na Figura 7.29 é ilustrada a tela para exibir o histórico de um laudo médico. Na guia “1” (a) é exibido o primeiro laudo cadastrado e na guia “2” (b) é apresentada a segunda versão desse laudo médico editado, cujo conteúdo pode ser verificado na Figura 7.28. Nesse exemplo, foram geradas duas versões desse laudo médico.

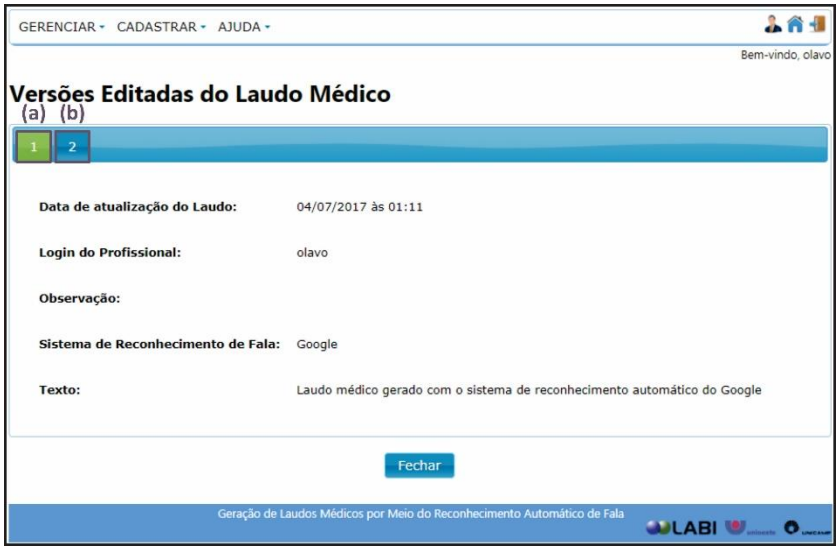

Figura 7.29: Tela de exibição do histórico de um laudo médico.

Na Figura 7.30 é ilustrada a tela de gerenciamento de laudos médicos, em que é possível realizar buscas dos laudos cadastrados por código do laudo médico (a), código do exame (b), data de realização do exame (c), data de geração do laudo médico (d) e nome do paciente (e). Nessa tela também é possível exibir a última edição do laudo médico (f), e o seu respectivo histórico (g), editá-lo com o SRAF da Google (h) ou editá-lo com o SRAF da Microsoft (i). Também é possível a edição dos laudos médicos de maneira manual utilizando o teclado.

Figura 7.30: Tela de gerenciamento dos laudos médicos.

Na Figura 7.31 é ilustrado o módulo administrativo do PSW. Nesse módulo é possível realizar o gerenciamento dos profissionais cadastrados que contempla as opções de ativar um profissional (a), torná-lo administrador (b) e excluí-lo (c).

Figura 7.31: Tela de gerenciamento de profissionais do Protótipo de Sistema *Web*.

7.5 Considerações Finais

A RS resultou em um total de 169 trabalhos relacionados a tecnologias de reconhecimento automático de fala, dentre eles, cinco utilizaram SRAF para a Língua Portuguesa do Brasil. Ao todo, foram encontrados 81 SRAFs.

A avaliação dos sete sistemas para a Língua Portuguesa do Brasil, resultou na seleção do *Web Speech* API e do Bing *Speech* API, para serem acoplados ao PSW. A seleção de dois sistemas se deve ao fato de que o *Web Speech* API funciona apenas em sistemas operacionais e navegadores de *Internet* desenvolvidos pela Google. Portanto, o Bing *Speech* API foi acoplado para possibilitar o seu uso em dispositivos móveis com Windows e navegador de *Internet Edge*, além de ser uma alternativa de utilização, caso a Google interrompa o seu serviço.

No capítulo seguinte é apresentada a conclusão alcançada por este trabalho, onde é relatado as principais contribuições, bem como as limitações e as sugestões de trabalhos futuros.

Capítulo 8

Conclusão

Neste trabalho foi investigado o uso da tecnologia de reconhecimento automático de fala para ser utilizada no âmbito médico. A fundamentação deste trabalho foi apoiada por uma revisão sistemática, a qual contribuiu para se obter um panorama geral sobre o uso dessa tecnologia em pesquisas científicas e as técnicas desenvolvidas para melhorar a sua precisão.

Foi realizado um estudo sobre as principais características dos sons e como ocorre o processo da fala humana e a maneira de como é compreendida. A partir desse estudo, foi apresentado um paralelo de funcionamento entre a tecnologia de um SRAF com a capacidade humana de compreender a fala.

Também foram discutidas as principais técnicas de aprendizado de máquina utilizadas em SRAFs, como o HMM, que é a tecnologia mais utilizada, sendo amplamente empregada em diversos sistemas e as RNNs, que estão presentes nos SRAFs estado-da-arte, cujos sistemas apresentam a maior taxa de precisão de reconhecimento.

A partir da avaliação dos SRAFs para a Língua Portuguesa do Brasil, constatou-se que os sistemas que operam com a tecnologia RNN-LSTM apresentam os melhores níveis de reconhecimento para a Língua Portuguesa do Brasil, como o Google *Web Speech* API e o Microsoft Bing *Speech* API.

De maneira geral, os SRAFs apresentam deterioração no reconhecimento quando utilizados em ambientes ruidosos provenientes, por exemplo, de dispositivos elétricos ou mecânicos e até mesmo da fala concorrente de outros oradores.

Dentre as dificuldades para se treinar um SRAF, tem-se o tamanho do vocabulário, em que dependendo do idioma podem existir centenas de milhares de palavras, como a Língua Portuguesa do Brasil que conta com um universo de 381.000 palavras (VOLP, 2009) e variações de dialeto, por exemplo, da fala carioca, gaúcha, mineira, nordestina e paulista. Além disso, esses sistemas devem ser capazes de lidar com as variações das diferentes características da fala particular de cada indivíduo, como ritmo de fala, qualidade vocal, idade, gênero e sotaque.

Desse modo, além do grande número de conteúdo textual para gerar a transcrição dos SRAFs, também é necessário o seu treinamento com dados de áudio, que deve contar com um grande número de pronúncias de palavras para tratar da variação de características da fala.

Com relação às tecnologias de código aberto para os SRAFs que operam em Língua Portuguesa do Brasil, um grande empecilho para a construção desses sistemas é a escassez de *corpora* para realizar o seu treinamento adequado.

Mesmo diante das limitações enfrentadas pelos SRAFs, como a adição de ruído no sinal da fala e as palavras fora do vocabulário de treinamento, considerando o contexto deste

trabalho, concluímos que o seu uso pode ser recomendado para testes iniciais no âmbito médico.

A recomendação da utilização de SRAF no âmbito médico é devido ao nível satisfatório de precisão de reconhecimento da fala dos sistemas integrados ao Protótipo de Sistema *Web*. Além disso, conforme relatado nos trabalhos de Johnson et al., (2014); Prevedello et al. (2014); Ahlgrim et al., (2016); Hodgson & Coiera, (2016), a utilização de SRAF reduziu, de maneira significativa, o tempo para confecção de documentos médicos.

A hipótese deste trabalho foi validada com a construção do Protótipo de Sistema *Web* para a geração de laudos médicos para a Língua Portuguesa do Brasil por meio da tecnologia de sistemas computacionais de reconhecimento automático de fala.

É interessante mencionar que a utilização de SRAFs no âmbito médico pode ser viável não apenas para o contexto de laudos médicos, mas também pode prover suporte como uma memória de trabalho, por exemplo, durante procedimentos médicos.

8.1 Principais Contribuições

Com a conclusão deste trabalho, como principais resultados, pode-se citar:

- Realização de uma revisão sistemática na área de SRAFs;
- Avaliação do nível de precisão de SRAFs para a Língua Portuguesa do Brasil;
- Construção de um Protótipo de Sistema *Web* para geração de laudos médicos utilizando SRAFs;
- Validação da hipótese de que o uso da tecnologia de reconhecimento automático de fala pode ser utilizada no âmbito médico.

8.2 Limitações

As principais limitações deste trabalho são:

- A utilização dos SRAFs integrados ao Protótipo de Sistema *Web* requer acesso à *Internet*;
- O Protótipo de Sistema *Web* não integrou tecnologia de reconhecimento automática de fala de código aberto;
- A geração de laudos médicos não foi avaliada em ambiente hospitalar.

8.3 Trabalhos Futuros

Com a conclusão deste trabalho foram identificados alguns trabalhos futuros, que incluem (a) o desenvolvimento de novas funcionalidades ao Protótipo de Sistema *Web*, (b) sua avaliação, (c) o emprego da tecnologia de reconhecimento automático de fala em outros

projetos desenvolvidos no LABI/UNIOESTE, (d) o desenvolvimento de uma solução própria de um SRAF, e (e) também a investigação para melhorar a taxa de precisão dos SRAFs.

Com relação à inclusão de novas funcionalidades ao Protótipo de Sistema *Web*, tem-se:

- Desenvolver um módulo que permita transcrever laudos médicos a partir de arquivos de áudio;
- Exportar o laudo médico para o formato PDF;
- Integrar comandos verbais durante o ditado de um laudo médico, como criar um novo parágrafo, interromper ou iniciar o reconhecimento;
- Realizar testes no PSW com profissionais da saúde;
- Quantificar a possível redução do tempo para a confecção de laudos médicos utilizando a tecnologia de reconhecimento automático de fala.

A utilização da tecnologia de reconhecimento automático de fala pode ser integrada a outros projetos desenvolvidos no LABI/UNIOESTE, como:

- Integrar a tecnologia de reconhecimento automático de fala para a geração dos laudos médicos para realizar o mapeamento em uma representação estruturada (Oliva et al., 2016);
- Integrar a tecnologia de reconhecimento automático de fala para disponibilizar legendas durante um acompanhamento, com interação remota em tempo real, em um exame de colonoscopia (Machado et al., 2012; Takaki, 2015).

Com relação ao desenvolvimento de uma tecnologia própria de SRAF, as sugestões incluem:

- Treinar um SRAF para a Língua Portuguesa do Brasil utilizando alguma ferramenta de código aberto;
- Desenvolver uma solução própria de um SRAF.

Por fim, investigar a tecnologia de SRAF para melhorá-la, por exemplo, quais características da voz podem influenciar na precisão de um SRAF.

Referências Bibliográficas

Abad, A., Meinedo, H., Trancoso, I. & Neto, J. (2012). Transcription of multi-variety portuguese media contents, *Computational Processing of the Portuguese Language: 10th International Conference (PROPOR)*, 2012, Coimbra, Portugal, pp. 409–420.

Adde, L. (2013). *A Discriminative Approach to Pronunciation Variation Modeling in Speech Recognition*, Tese de doutorado, Norwegian University of Science and Technology, Trondheim, Noruega.

Adell, J., Bonafonte, A., Cardenal, A., Fonollosa, M. R. J. A. R., Moreno, A., Navas, E. & Banga, E. R. (2012). The buceador multi-language search engine for digital libraries, *8th International Conference on Language Resources and Evaluation (LREC)*, 2012, Istambul, Turquia, pp. 1705–1709.

Ahlgrim, C., Maenner, O. & Baumstark, M.-W. (2016). Introduction of digital speech recognition in a specialised outpatient department: a case study, *BMC Med. Inf. & Decision Making* **16**: 132.

Ahn, T. Y. & Lee, S.-M. (2016). User experience of a mobile speaking application with automatic speech recognition for efl learning, *British Journal of Educational Technology* **47**(4): 778–786.

Ajmera, J., Deshmukh, O. D., Jain, A., Nanavati, A. A., Rajput, N. & Srivastava, S. (2012). Audio cloud: Creation and rendering, *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces (IUI)*, 2012, Lisboa, Portugal, pp. 277–280.

Álvarez, A., Mendes, C., Raffaelli, M., Luís, T., Paulo, S., Piccinini, N., Arzelus, H., Neto, J., Aliprandi, C. & del Pozo, A. (2016). *Automating live and batch subtitling of multimedia contents for several european languages, Multimedia Tools and Applications* **75**(18): 10823–10853.

Alves, W. P. (2015). *Java para Web - Desenvolvimento de Aplicações*, Érica, São Paulo.

Anusuya, M. A. & Katti, S. K. (2011). Front end analysis of speech recognition: A review, *International Journal of Speech Technology* **14**(2): 99–145.

Arısoy, E., Kurimo, M., Saraçlar, M., Hirsimäki, T., Pylkkönen, J., Alumäe, T. & Sak, H. (2008). *Statistical Language Modeling for Automatic Speech Recognition of Agglutinative Languages*, Routledge, capítulo Speech Recognition: Technologies and Applications, pp. 193–204.

Aurora (2017). Aurora, Aurora speech recognition experimental framework. Acesso em: junho/17.
Disponível em: *http://aurora.hsnr.de/index-2.html*

Baker, J. (1975). The dragon system - an overview, *IEEE Transactions on Acoustics, Speech, and Signal Processing* **23**(1): 24–29.

Balaji, V. & Sadashivappa, G. (2015). Speech disabilities in adults and the suitable speech recognition software tools - a review, *International Conference on Computing and Network Communications (CoCoNet)*, 2015, Trivandrum, Índia, pp. 559–564.

Batista, P. S. (2013). *Avanços em Reconhecimento de Fala para Português Brasileiro e Aplicações: Ditado no LibreOffice e Unidade de Resposta Audível com Asterisk*, Dissertação de mestrado, Universidade Federal do Pará, Belém, Brasil.

Bauer, W., Westfall, G. D. & Dias, H. (2013). *Física para Universitários: Relatividade, Oscilações, Ondas e Calor*, AMGH, São Paulo.

Beekes, R. S. P. (2011). *Comparative Indo-European Linguistics*, 2 edn, John Benjamins Publishing Company, Holanda.

Benton, A. & Dredze, M. (2015). Entity linking for spoken language, *The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, 2015, Denver, Estados Unidos da América, pp. 225–230.

Bolaños, D., Cole, R. A., Ward, W. H., Tindal, G. A., Schwanenflugel, P. J. & Kuhn, M. R. (2013). Automatic assessment of expressive oral reading, *Speech Communication* **55**(2): 221–236.

Bonilla, D. A., Nedjah, N. & de Macedo Mourelle, L. (2016). Online pattern recognition for portuguese phonemes using multi-layer perceptron combined with recurrent non-linear autoregressive neural networks with exogenous inputs, *IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, 2016, Cartagena, Colômbia, pp. 1–6.

Braga, A. P., Carvalho, A. P. L. F. & Ludemir, T. B. (2000). *Redes Neurais Artificiais*: Teoria e Aplicações, 11 edn, LTC, Rio de Janeiro.

Butko, T. & Nadeu, C. (2011). Audio segmentation of broadcast news in the albayzin-2010 evaluation: overview, results, and discussion, *EURASIP Journal on Audio, Speech, and Music Processing* (1): 1–10.

Can, D., Gibson, J., Vaz, C., Georgiou, P. G. & Narayanan, S. S. (2014). Barista: A framework for concurrent speech processing by usc-sail, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, Florença, Itália, pp. 3306–3310.

Cavus, N. & Ibrahim, D. (2016). Learning english using children's stories in mobile devices, *British Journal of Educational Technology* **4**(2): 625–641.

Chalegre-Paula, M. C. & Neto, F. B. L. (2016). Expanded abstract: An adaptive support system for phonological treatment of phonetic disorders using visual feedbacks and

speech recognition, *IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, 2016, Cartagena, Colômbia, pp. 1–2.

Chao, C. & Thomaz, A. (2016). Timed petri nets for fluent turn-taking over multimodal interaction resources in human-robot collaboration, *I. J. Robotics Res.* **35**(11): 1330–1353.

Clark, A., Fox, C. & Lappin, S. (eds) (2010). *Computational Linguistica and Natural Language Processing*, Wiley-Blackwell, Singapura.

Clavel, C., Adda, G., Cailliau, F., Garnier-Rizet, M., Cavet, A., Chapuis, G., Courcinous, S., Danesi, C., Daquo, A.-L., Deldossi, M., Guillemin-Lanne, S., Seizou, M. & Suignard, P. (2013). Spontaneous speech and opinion detection: mining call-centre transcripts, *Language Resources and Evaluation* **47**(4): 1089–1125.

CMUSphinx (2017). CMUSphinx, Frequenty Asked Questions (FAQ). Acesso em: julho/17. Disponível em: *https://cmusphinx.github.io/wiki/faq/*

Cristófaro-Silva, T. (2003). *Fonética e Fonologia do Português - Roteiro de Estudos e Guia de Exercícios*, 7 edn, Contexto, São Paulo.

CRM-PR (2008). Conselho Regional De Medicina Do Paraná, Parecer no 1936/2008 CRMPR, Diferença entre atestado e laudo médico. Acesso em: junho/17. Disponível em: *http://www.portalmedico.org.br/pareceres/CRMPR/pareceres/2008/1936_2008.htm*

Cucu, H., Buzo, A., Petrică, L., Burileanu, D. & Burileanu, C. (2014). Recent improvements of the speed romanian lvcsr system, *10th International Conference on Communications (COMM)*, 2014, Bucareste, Romênia, pp. 1–4.

Cutnell, J. D. & Johnson, K. W. (2016). *Física - Volume 1*, 9 edn, LTC, Rio de Janeiro.

Dahl, G. E., Yu, D., Deng, L. & Acero, A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition, *IEEE Transactions on Audio, Speech, and Language Processing* **20**(1): 30–42.

Davis, M. H. & Scharenborg, O. (2017). *Speech Perception and Spoken Word Recognition*, Routledge, capítulo Speech Perception by Humans and Machines, pp. 181-203.

Davis, S. & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, *IEEE Transactions on Acoustics, Speech, and Signal Processing* **28**(4): 129–135.

Delcroix, M., Kinoshita, K., Nakatani, T., Araki, S., Ogawa, A., Hori, T., Watanabe, S., Fujimoto, M., Yoshioka, T., Oba, T., Kubo, Y., Souden, M., Hahm, S.-J. & Nakamura, A. (2013). Speech recognition in living rooms: Integrated speech enhancement and recognition system based on spatial, spectral and temporal modeling of sounds, *Computer Speech & Language* **27**(3): 851–873.

Domingues, L. (2013). Grupos do Google, Tutorial de instalação do Julius + modelos acústicos do Coruja no Ubuntu x86. Acesso em: julho/17. Disponível em: https://groups.google.com/forum/#!searchin/coruja-users/ubuntu$20instala%C3%A7%C3%A3o%7Csort:relevance/coruja-users/Sjl8InZ11B8/EAThYSloI0J

Donaj, G. & Kačič, Z. (eds) (2017). *Language Modeling for Automatic Speech Recognition of Inflective Languages - An Applications-Oriented Approach Using Lexical Data*, Springer International Publishing, Suiça.

Felix, V. G., L. J. Mena, R. O. & Maestre, G. E. (2017). A pilot study of the use of emerging computer technologies to improve the effectiveness of reading and writing therapies in children with down syndrome, *British Journal of Educational Technology* **48**(2): 611–624.

Feng, J., Johnston, M. & Bangalore, S. (2011). Speech and multimodal interaction in mobile search, *IEEE Signal Processing Magazine* **28**(4): 40–49.

FLAC (2014). Free Lossless Audio Codec, Introduction. Acesso em: junho/17. Disponível em: *https://xiph.org/flac/features.html*

Fuller, D., Pimentel, J. T. & Peregoy, B. M. (2014). *Anatomia e Fisiologia Aplicadas à Fonoaudiologia*, Manole, Barueri.

Ganapathy, S., Thomas, S., Dimitriadis, D. & S.Rennie (2015). Investigating factor analysis features for deep neural networks in noisy speech recognition, *16th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2015, Dresden, Alemanha, pp. 1898–1902.

García-Moral, A. I., Solera-Urena, R., Pelaez-Moreno, C. & de Maria, F. D. (2011). Data balancing for efficient training of hybrid ann/hmm automatic speech recognition systems, *IEEE Transactions on Audio, Speech, and Language Processing* **19**(3): 468–481.

Gavat, I., Militaru, D. M. & Dumitru, C. O. (2008). *Speech Recognition Technologies and Applications*, I-Tech, capítulo Knowledge Resources in Automatic Speech Recognition and Understanding for Romanian Language, Áustria, pp. 241–260.

Gold, B., Morgan, N. & Ellis, D. (eds) (2011). *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*, 2 edn, John Wiley & Sons, Estados Unidos da América.

Google (2017). Google Cloud Platform, Cloud Speech API. Acesso em: junho/17. Disponível em: *https://cloud.google.com/speech/*

Gopi, A., P, S. D., T, S., Stephen, J. & VK, B. (2015). Multilingual speech to speech mt based chat system, *International Conference on Computing and Network Communications (CoCoNet)*, 2015, Trivandrum, Índia, pp. 771–776.

Graaff, V. (2003). *Anatomia Humana*, 6 edn, Manole, Barueri.

Graves, A. (ed.) (2012). *Supervised Sequence Labelling with Recurrent Neural Networks*, Springer-Verlag Berlin Heidelberg, Alemanha.

Griol, D., Callejas, Z., López-Cózar, R. & Riccardi, G. (2014). A domain-independent statistical methodology for dialog management in spoken dialog systems, *Computer Speech & Language* **28**(3): 743–768.

Gupta, V., Deléglise, P., Boulianne, G., Estève, Y., Meignier, S. & Rousseau, A. (2015). Crim and lium approaches for multi-genre broadcast media transcription, *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, Scottsdale, Estados Unidos da América, pp. 681–686.

Gür, B. (2012). *Improving Speech Recognition Accuracy for Clinical Conversations*, Dissertação de mestrado, Massachusetts Institute of Technology, Cambridge, Estados Unidos da América.

Hain, T., Christian, J., Saz, O., Deena, S., Hasan, M., Ng, R.-W.-M., Milner, R., Doulaty, M. & Liu, Y. (2016). webasr 2 - improved cloud based speech technology, *International Speech Communication Association (INTERSPEECH)*, 2016, São Francisco, Estados Unidos da América, pp. 1613–1617.

Hakkani-Tür, D., Tur, G., Celikyilmaz, A., Chen, Y.-N. V., Gao, J. & L. Deng, Y.-Y. W. (2016). Multi-domain joint semantic frame parsing using bi-directional rnn-lstm, *Proceedings of The 17th Annual Meeting of the International Speech Communication Association (INTERSPEECH)*, São Francisco, Estados Unidos da América, 2016. Acesso em: julho/2017.
Disponível em: *https://www.microsoft.com/en-us/research/publication/multijoint/*

Hall, J. E. & Guyton, A. (2011). *Tratado de Fisiologia Médica*, 12 edn, Elsevier, Rio de Janeiro.

Haykin, S. (2008). *Neural Networks and Learning Machines*, 3 edn, Pearson, Estados Unidos da América.

Hazen, T. J. (2011). Mce training techniques for topic identification of spoken audio documents, *IEEE Transactions on Audio, Speech, and Language Processing* **19**(8): 2451–2460.

Helmke, H., Rataj, J., Mühlhausen, T., Ohneiser, O., Ehr, H., Matthias, K., Oualil, Y. & Schulder, M. (2015). Assistant-based speech recognition for atm applications, *11th USA/EUROPE Air Traffic Management R&D Seminar*, 2015, Lisboa, Portugal, pp. 1–10.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech, *J. Acoust. Soc. Am.* **87**(4): 1738–1752.

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory, *Neural Comput.* **9**(8): 1735–1780.

Hodgson, T. & Coiera, E. (2016). Risks and benefits of speech recognition for clinical documentation: a systematic review, *Journal of the American Medical Informatics Association* **23**(e1): e169–e179.

HTK (2016). HTK, What is HTK?. Acesso em: junho/17.
Disponível em: *http://htk.eng.cam.ac.uk/*

Huang, X., Aceo, A. & Hon, H.-W. (eds) (2001). *Spoken Language Processing*, Prentice Hall PTR, Estados Unidos da América.

IBM (2017). IBM, About Speech to Text. Acesso em: junho/17.
Disponível em: *https://www.ibm.com/watson/developercloud/doc/speech-to-text/index.html*

IBM-Microsoft (1991). *Multimedia Programming Interface and Data Specifications 1.0*, IBM Corporation e Microsoft Corporation. Acesso em: junho/17.
Disponível em: *https://research.google.com/pubs/pub38130.html*

ICSI (2015). International Computer Science Institute (ICSI), The alignment process. Acesso em: junho/17.
Disponível em: http://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm

Ili (2017). Ili, Software optimized for traveling. Acesso em julho/17.
Disponível em: *https://iamili.com/travel.html*

Jaitly, N., Nguyen, P., Senior, A. & Vanhoucke, V. (2012). Application of pretrained deep neural networks to large vocabulary speech recognition, *13th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2012, Portland, Estados Unidos da América, pp. 1–4. Acesso em julho/17.
Disponível em: *https://research.google.com/pubs/pub38130.html*

Johnson, M., Lapkin, S., Long, V., Sanchez, P., Suominen, H., Basilakis, J. & Dawson, L. (2014). A systematic review of speech recognition technology in health care, *BMC Medical Informatics and Decision Making* **14**(1): 94.

Jurafsky, D. & Martin, J. H. (2016). Stanford University, Speech and Language Processing, 3 edn. Acesso em julho/17.
Disponível em: https://web.stanford.edu/~jurafsky/slp3/9.pdf

Kaldi (2017). Kaldi, About the Kaldi project. Acesso em: junho/17.
Disponível em: *http://kaldiasr.org/doc/about.html*

Karafiát, M., Grézl, F., Burget, L., Szöke, I. & Cernocký, J. (2015). Three ways to adapt a CTS recognizer to unseen reverberated speech in BUT system for the aspire challenge, *16th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2015, Dresden, Alemanha, pp. 2454–2458.

Kaushik, L., Sangwan, A. & Hansen, J. H. L. (2013). Automatic sentiment extraction from youtube videos, *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, Olomouc, República Checa, pp. 239–244.

Kaushik, L., Sangwan, A. & Hansen, J. H. L. (2017). Automatic sentiment detection in naturalistic audio, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **PP**(99): 1–1.

Kesten, P. R. & Tauck, D. L. (2015). *Física na Universidade para as Ciências Físicas e da Vida - Volume 2*, LTC, Rio de Janeiro.

Kim, P. (2017). *MATLAB Deep Learning: With Machine Learning, Neural Networks and Artificial Intelligence*, Apress, Coréia do Sul.

Kitchenham, B. & Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering, *Relatório Técnico EBSE 2007-001*, University of Durham, Inglaterra.

Klatt, D. H. (1977). Review of the arpa speech understanding project, *The Journal of the Acoustical Society of America* **62**(6).

Law-To, J. & Grefenstette, G. (2011). Voxalead: A scalable video search engine based on content, *Proceedings of the 19th ACM International Conference on Multimedia*, 2011, Scottsdale, Estados Unidos da América, pp. 747–748.

LDC (2017a). Linguistic Data Consortium, CSR-I (WSJ0) Complete. Acesso em: junho/17. Disponível em: *https://catalog.ldc.upenn.edu/ldc93s6a*

LDC (2017b). Linguistic Data Consortium, CSR-II (WSJ1) Complete. Acesso em: junho/17. Disponível em: *https://catalog.ldc.upenn.edu/ldc94s13a*

LDC (2017c). Linguistic Data Consortium, TIMIT Acoustic-Phonetic Continuous Speech Corpus. Acesso em: junho/17.
Disponível em: *https://catalog.ldc.upenn.edu/ldc93s1*

Lee, K. F., Hon, H. W. & Reddy, R. (1990). An overview of the sphinx speech recognition system, *IEEE Transactions on Acoustics, Speech, and Signal Processing* **38**(1): 35–45.

Lesser, V., Fennell, R., Erman, L. & Reddy, D. (1975). Organization of the hearsay ii speech understanding system, *IEEE Transactions on Acoustics, Speech, and Signal Processing* **23**(1): 11–24.

Leuski, A., Gowrisankar, R., Richmond, T., Shapiro, A., Xu, Y. & Feng, A. (2014). Mobile personal healthcare mediated by virtual humans, *Proceedings of the Companion Publication of the 19th International Conference on Intelligent User Interfaces*, 2014, Haifa, Israel, pp. 21–24.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals, *Soviet Physics Doklady* **10**: 707.

Li, J., Deng, L., Haeb-Umbach, R. & Gong, Y. (eds) (2016). *Robust Automatic Speech Recognition - A Bridge to Practical Applications*, Elsevier, Estados Unidos da América.

Li, K., Qian, X. & Meng, H. (2017). Mispronunciation detection and diagnosis in l2 english speech using multidistribution deep neural networks, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **25**(1): 193–207.

Lison, P. (2015). A hybrid approach to dialogue management based on probabilistic rules, *Computer Speech & Language* **34**(1): 232–255.

Liu, F., Tur, G., Hakkani-Tür, D. & Yu, H. (2011). Towards spoken clinical-question answering: evaluating and adapting automatic speech-recognition systems for spoken clinical questions, *Computer Speech Language* **18**(5): 625–630.

Lopes-Filho, O. (2013). Novo Tratado de Fonoaudiologia, 3 edn, Manole, Barueri.

Lowerre, B. T. (1976). *The Harpy Speech Recognition System*, Tese de doutorado, Carnegie Mellon University, Pittsburgh, Estados Unidos da América.

Lyons, J. (1987). *Linguagem e Linguística Uma introdução*, LTC, Rio de Janeiro.

Machado, R. B., Lee, H. D., Ayrizono, M. L. S., Leal, R. F., Coy, C. S. R., Fagundes, J. J. & Chung, W. F. (2012). Prototype of a computer system for managing data and video colonoscopy exams, *Journal of Coloproctology (Rio de Janeiro)* **32**(1): 50–59.

Makino, S., Kawabata, T. & Kido, K. (1983). Recognition of consonant based on the perceptron model, *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP*, 1983, Boston, Estados Unidos da América, pp. 738–741 vol. 8.

Matsuda, S., Hayashi, T., Ashikari, Y., Shiga, Y., Kashioka, H., Yasuda, K., Okuma, H., Uchiyama, M., Sumita, E., Kawai, H. & Nakamura, S. (2017). Development of the "voicetra" multi-lingual speech translation system, *IEICE Transactions on Information and Systems* **E100.D**(4): 621–632.

Mengistu, K. T., Rudzicz, F. & Falk, T. H. (2011). Using acoustic measures to predict automatic speech recognition performance for dysarthric speakers, *7th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, Florença, Itália, pp. 75–78.

Microsoft (2017a). Microsoft Translator, Quebre a barreira do idioma. Acesso em: julho/17. Disponível em: *https://translator.microsoft.com/*

Microsoft (2017b). Microsoft Azurre, Bing speech API overview. Acesso em: junho/17. Disponível em: *https://docs.microsoft.com/pt-br/azure/cognitive-services/speech/home*

Microsoft (2017c). Microsoft .NET, Download .NET Framework. Acesso em: julho/17. Disponível em: *https://www.microsoft.com/net/download/framework*

Microsoft (2017d). Microsoft Visual Studio, O Que Há de Novo no Visual Studio 2017. Acesso em: julho/17.
Disponível em: *https://www.visualstudio.com/pt-br/vs/whatsnew/*

Microsoft (2017e). Microsoft Cognitive Services, Put intelligence APIs to work. Acesso em: julho/17.
Disponível em: *https://azure.microsoft.com/en-us/services/cognitive-services/?v=17.25c*

Misu, T., Georgila, K., Leuski, A. & Traum, D. (2012). Reinforcement learning of questionanswering dialogue policies for virtual museum guides, *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, Association for Computational Linguistics, Seul, Coréia do Sul, pp. 84–93.

Mitchell, T. M. (ed.) (1997). *Machine Learning*, McGraw-Hill Science, Estados Unidos da América.

Moura, R., Neto, N., Patrick, C., Batista, P. & Klautau, A. (2010). Fftranscriber: Software para transcrição otimizado para aplicações forenses, *VIII Seminário Nacional de Fonética Forense*, 2010, Tocantins, Brasil, pp. 1–5.

Moreno, J., Garrote, M., Martínez, P. & J. L. Martínez-Fernández, J. L. (2011). *Some Experiments in Evaluating ASR Systems Applied to Multimedia Retrieval*, Springer Berlin Heidelberg, Alemanha, pp. 12–23.

Neto, N., Patrick, C., Klautau, A. & Trancoso, I. (2011). Free tools and resources for brazilian portuguese speech recognition, *Journal of the Brazilian Computer Society* **17**(1): 53–68.

Nuance (2017). Nuance, Improve clinical documentation across the continuum of care. Acesso em: julho/17.
Disponível em: *https://www.nuance.com/*

Oliva, J. T., Lee, H. D., Spolaôr, N., Coy, C. S. R. & Chung, W. F. (2016). Prototype system for feature extraction, classification and study of medical images, *Expert Systems with Applications* **63**: 267–283.

Oliveira-Junior, H. A., Caldeira, A. M. & Machado, M. A. S. (eds) (2007). *Inteligência Computacional Aplicada à Administração, Economia e Engenharia em Matlab*, Thomson Learning, São Paulo.

Pallett, D. S., Fiscus, J. G., Fisher, W. M., Garofolo, J. S., Lund, B. A. & Przybocki, M. A. (1994). 1993 benchmark tests for the arpa spoken language program, *Proceedings of the Workshop on Human Language Technology*, 1994, Plainsboro, Estados Unidos da América, pp. 49–74.

Papangelis, A., Gatchel, R., Metsis, V. & Makedon, F. (2013). An adaptive dialogue system for assessing post traumatic stress disorder, *Proceedings of the 6th International Conference on PErvasive Technologies Related to Assistive Environments*, 2013, Rodes, Grécia, pp. 49:1–49:4.

PocketSphinx (2017). Carnegie Mellon University, PocketSphinx - Sphinx for handhelds. Acesso em: julho/17.
Disponível em: *http://www.speech.cs.cmu.edu/pocketsphinx/*

Pressman, R. S. & Maxim, B. R. (2016). *Engenharia de Software: uma abordagem profissional*, 8 edn, AMGH, Porto Alegre.

Prevedello, L. M., Ledbetter, S., Farkas, C. & Khorasani, R. (2014). Implementation of speech recognition in a community-based radiology practice: Effect on report turnaround times, *Journal of the American College of Radiology* **11**(4): 402–406.

Price, P., Fisher, W. M., Bernstein, J. & Pallett, D. S. (1988). The darpa 1000-word resource management database for continuous speech recognition, *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1988, Nova York, Estados Unidos da América, pp. 651–654 vol.1.

Quintanilha, I. M. (2017). *End-to-end speech recognition applied to brazilian portuguese using deep learning*, Dissertação de mestrado, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil.

Rabiner, L. & Juang, B.-H. (eds) (1993). *Fundamentals of Speech Recognition*, Prentice-Hall International, Estados Unidos da América.

Revuelta-Martínez, A., Rodríguez, L. & García-Varea, I. (2012). A computer assisted speech transcription system, *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 2012, Avignon, França, pp. 41–45.

Riemann, M., Knipfer, C., Rohde, M., Adler,W., Schuster, M., Noeth, E., Oetter, N., Shams, N., Neukam, F.-W. & Stelzle, F. (2016). Oral squamous cell carcinoma of the tongue: Prospective and objective speech evaluation of patients undergoing surgical therapy, *Journal of the Sciences and Specialities of the Head and Neck* **38**(7): 993–1001.

Roberto-Douglas, C. (2009). Tratado de Fisiologia Aplicada às Ciências Médicas, 6 edn, Guanabara Koogan, Rio de Janeiro.

Sak, H., Senior, A. W. & Beaufays, F. (2014). Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition, *CoRR* **abs/1402.1128**. Acesso em: julho/2017.
Disponível em: *http://arxiv.org/abs/1402.1128*

Sakti, S., Paul, M., Finch, A., Sakai, S., Vu, T. T., Kimura, N., Hori, C., Sumita, E., Nakamura, S., Park, J., Wutiwiwatchai, C., Xu, B., Riza, H., Arora, K., Luong, C. M. & Li, H. (2013). A-star: Toward translating asian spoken languages, *Comput. Speech Lang.* **27**(2): 509–527.

Santos-Perez, M., Gonzalez-Parada, E. & Cano-garcia, J. M. (2013). Mobile embodied conversational agent for task specific applications, *IEEE Transactions on Consumer Electronics* **59**(3): 61o–614.

Schalkwyk, J., Beeferman, D., Beaufays, F., Byrne, B., Chelba, C., Cohen, M., Kamvar, M. & Strope, B. (2010). *Advances in Speech Recognition - Mobile Environments, Call*

Centers and Clinics, Springer, capítulo "Your Word is my Command": Google Search by Voice: A Case Study, pp. 61–90.

Segbroeck, M. V., Travadi, R., Vaz, C., Kim, J., Black, M. P., Potamianos, A. & Narayanan, S. S. (2014). Classification of Cognitive Load from Speech using an i-vector Framework, Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association (INTERSPEECH), 2014, Singapura, pp. 751–755.

Sheffield, D., Anderson, M., Lee, Y. & Keutzer, K. (2013). Hardware/software codesign for mobile speech recognition, *14th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2013, Lion, França, pp. 627–631.

Shires, G & Wennborg, H. (2014). W3C, Web Speech API Specification. Acesso em: junho/17.
Disponível em: *https://dvcs.w3.org/hg/speech-api/raw-file/tip/webspeechapi.html*

Siddique, N. & Adeli, H. (eds) (2013). *Computational Intelligence Synergies of Fuzzy Logic, Neural Networks and Evolutionary Computing*, John Wiley & Sons, Índia.

Silva, C. P. A. (2010). *Um software de Reconhecimento de Voz para o Pt-Br*, Dissertação de mestrado, Universidade Federal do Pará, Belém, Brasil.

Silva, E., Baptista, L., Fernandes, H. & Klautau, A. (2005). Desenvolvimento de um sistema de reconhecimento automático de voz contínua com grande vocabulário para o português brasileiro, *XXV Congresso da Sociedade Brasileira de Computação*, 2005, São Leopoldo, Brasil, pp. 2258-2267. Acesso em: julho/2017.
Disponível em: *http://www.nilc.icmc.usp.br/til/til2005/arq0069.pdf*

Silva, P., Batista, P., Neto, N. & Klautau, A. (2010b). An open-source speech recognizer for brazilian portuguese with a windows programming interface, *The International Conference on Computational Processing of Portuguese (PROPOR), 2010, Porto Alegre*, pp. 1–4.

Sinclair, M., Bell, P., Birch, A. & McInnes, F. (2014). A semi-markov model for speech segmentation with an utterance-break prior, *15th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2014, Singapura, pp. 2351–2355.

Smaragdis, P., Radhakrishnan, R. & Wilson, W. (2009). *Multimedia Content Analysis: Theory and Applications*, Springer, capítulo Context Extraction Through Audio Signal Analysis, pp. 1–34.

Soller, R. W., Chan, P. & Higa, A. (2012). Performance of a new speech translation device in translating verbal recommendations of medication action plans for patients with diabetes, *Journal of Diabetes Science and Technology* **6**(4): 927–937.

SRI (2017). SRI International, Software Development Kits. Acesso em: junho/17. Disponível em: *http://www.speechatsri.com/products/sdk.shtml*

Stadtschnitzer, M., Schwenninger, J., Stein, D. & Koehler, J. (2014). Exploiting the large-scale german broadcast corpus to boost the fraunhofer iais speech recognition system, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC), European Language Resources Association (ELRA)*, 2014, Reiquiavique, Islândia, pp. 3887–3890.

Takaki, W. S. R. (2015). *Proposta de um Sistema Embarcado para Transmissão de Vídeos em Tempo Real com Aplicação em Telemedicina*, Dissertação de mestrado, Universidade Estadual do Oeste do Paraná, Foz do Iguaçu, Brasil.

Tsontzos, G. & Orglmeister, R. (2011). Cmu sphinx4 speech recognizer in a service-oriented computing style, *IEEE International Conference on Service-Oriented Computing and Applications (SOCA)*, 2011, Irvine, Estados Unidos da América, pp. 1–4.

Ultes, S., ElChabb, R., Schmitt, A. & Minker, W. (2013). Jachmm: A java-based conditioned hidden markov model library, *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3213–3217.

Varella, D. (2011). Drauzio Varella, Colonoscopia. Acesso em: junho/17.
Disponível em: *https://drauziovarella.com.br/cancer/colonoscopia/*

Varona, A., Nieto, S., Rodriguez-Fuentes, L. J., Penagarikano, M., Bordel, G. & Díez, M. (2011). A spoken document retrieval system for tv broadcast news in spanish and basque, *Procesamiento del Lenguaje Natural* (47): 75–83.

Vasquez-Correa, J. C., Orozco-Arroyave, J. R. & Nöth, E. (2016). Word accuracy and dynamic time warping to assess intelligibility deficits in patients with parkinsons disease, *XXI Symposium on Signal Processing, Images and Artificial Vision (STSIVA)*, 2016, Bucaramanga, Colômbia, pp. 1–5.

Virtanen, T., Singh, R. & Raj, B. (eds) (2012). *Techniques for Noise Robustness in Automatic Speech Recognition*, John Wiley & Sons, Reino Unido

Vocapia (2017). Vocapia, VoxSigma® Speech to Text Software Suite. Acesso em: junho/17.
Disponível em: *http://www.vocapia.com/voxsigma-speech-to-text.html*

Voiceinteraction (2017). Voiceinteraction, Produtos. Acesso em: junho/17.
Disponível em: *http://www.voiceinteraction.pt/?page_id=423*

Vogel, M., Kaisers, W., Wassmuth, R. & Mayatepek, E. (2015). Analysis of documentation speed using web-based medical speech recognition technology: Randomized controlled trial, *J Med Internet Res* **17**(11): e247.

VOLP (2009). Vocabulário Ortográfico da Língua Portuguesa, Busca no Vocabulário. Acesso em: julho/2017.
Disponível em: *http://www.academia.org.br/nossa-lingua/busca-no-vocabulario?sid=19*

W3Schools (2017a). W3Schools, OS Platform Statistics. Acesso em: junho/17.
Disponível em: *https://www.w3schools.com/browsers/browsers_os.asp*

W3Schools (2017b). W3Schools, Mobile Devices Statistics. Acesso em: junho/17. Disponível em: *https://www.w3schools.com/browsers/browsers_mobile.asp*

W3Schools (2017c). W3Schools, The Most Popular Browsers. Acesso em: junho/17, Disponível em: *https://www.w3schools.com/browsers/default.asp*

Ward, J. P. T. & Linden, R. W. A. (2014). Fisiologia Básica: Guia Ilustrado de Conceitos Fundamentais, 2 edn, Manole, Barueri.

Waverly (2017). Waverly Labs, A World Without Language Barriers. Acesso em julho/17. Disponível em: *http://www.waverlylabs.com/*

Wazlawick, R. S. (2011). *Análise e projeto de sistemas de informação orientados a objetos*, 2 edn, Elsevier, Rio de Janeiro.

Wazlawick, R. S. (ed.) (2013). *Engenharia de Software: Conceitos e Práticas*, Elsevier, Rio de Janeiro.

Winder, R. & Graham, R. (2009). Desenvolvendo Software em Java, 3 edn, LTC, Barueri.

Wolf, J. & Woods, W. (1977). The hwim speech understanding system, *Acoustics, Speech and Signal Processing, IEEE International Conference on ICASSP*, 1977, Hartford, Estados Unidos da América, pp. 784–787 vol. 2.

Wöllmer, M., Eyben, F., Schuller, B. & Rigoll, G. (2010). Recognition of spontaneous conversational speech using long short-term memory phoneme predictions, *International Speech Communication Association (INTERSPEECH)*, 2010, Makuhari, Japão, pp. 1946–1949

Yilmaz, E., Pelemans, J. & hamme, H. V. (2014). Automatic assessment of children's reading with the flavor decoding using a phone confusion model, *15th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2014, Singapura, pp. 969–972.

Yu, D. & Deng, L. (eds) (2015). Automatic Speech Recognition: A Deep Learning Approach, Springer, Inglaterra.

Yu, D., Deng, L. & Dahl, G. E. (2010). Roles of pre-training and fine-tuning in contextdependent dbn-hmms for real-world speech recognition. Acesso em: julho/2017. Disponível em: *https://www.microsoft.com/en-us/research/publication/roles-of-pre-training-and-fine-tuning-in-context-dependent-dbn-hmms-for-real-world-speech-recognition/*

Zahabi, M., Kaber, D. B. & Swangnetr, M. (2015). Usability and safety in electronic medical records interface design: A review of recent literature and guideline formulation, *Human Factors: The Journal of the Human Factors and Ergonomics Society* **57**(5): 805–834.

Zhang, Z., Weninger, F., Wöllmer, M., Han, J. & Schuller, B. (2017). Towards intoxicated speech recognition, *International Joint Conference on Neural Networks (IJCNN)*, 2017, Anchorage, Estados Unidos da América, pp. 1555–1559.

Apêndice A

Texto de Referência para Avaliar os Sistemas de Reconhecimento Automático de Fala

O texto de referência utilizado foi publicado na página de *internet* Drauzio Varella (Varella, 2011):

No futuro ninguém morrerá de câncer do intestino primeiro porque tumores malignos localizados no intestino delgado duodeno jejuno e íleo são raríssimos depois porque as lesões que se desenvolvem no intestino grosso cólon sigmoide e reto podem ser retiradas pela colonoscopia ainda na fase pré maligna na maioria dos casos o câncer de intestino se instala em lesões precursoras que adquirem a forma de pequenos cogumelos com pedículos de comprimento variável os pólipos em sua apresentação inicial eles são formados por arranjos glandulares de arquitetura muito semelhante a da mucosa normal são os pólipos adenomatosos ou adenomas a medida que o processo de transformação progride esses adenomas podem crescer e suas células se tornarem cada vez mais alteradas para constituir os pólipos hiperplásicos ao longo desse processo de multiplicação e de transformação celular eventualmente ocorrem as mutações de dna características das células malignas só então surge o câncer de intestino que mais tarde se disseminará para outros órgãos nos últimos 30 anos os avanços na tecnologia das fibras ópticas permitiram obter fibroscópios flexíveis capazes de visualizar a mucosa que vai do reto a válvula ileocecal local em que o íleo desemboca no intestino grosso os fibroscópios modernos não possibilitam apenas acesso visual as lesões da mucosa estão acoplados a pinças cortantes que permitem retirá-las a colonoscopia portanto não é simples exame diagnóstico é procedimento cirúrgico capaz de evitar o aparecimento do câncer de cólon embora seguro o exame não é desprovido de riscos em cada 1000 procedimentos um a dois pacientes sofrem complicações que vão de sangramentos a perfurações da parede intestinal além disso a a questão dos custos a necessidade de anestesia e de pessoal treinado e o desconforto do preparo com laxantes para esvaziar completamente o conteúdo intestinal por essas limitações a prevenção por meio da colonoscopia não deve ser indicada aleatoriamente mas ater-se as situações em que existe risco maior de desenvolver câncer de cólon por exemplo os que sofrem de doenças inflamatórias intestinais doença de crohn retocolite ulcerativa e outras devem submeter-se ao exame com maior frequência famílias que apresentam vários membros com múltiplos pólipos intestinais polipose familiar também além desses e de outros grupos de risco menos comuns os parentes de primeiro grau de mulheres e homens que tiveram câncer de cólon precisam ser acompanhados com mais

cuidado a partir de uma idade mais precoce embora não haja unanimidade a maioria dos especialistas aconselha que o primeiro exame nesses casos seja realizado 5 a 10 anos antes da idade em que o parente mais jovem recebeu o diagnóstico aqueles que não pertencem a nenhum grupo de risco nem tiveram parentes com câncer de cólon devem fazer a primeira colonoscopia entre os 50 e os 55 anos idade em que o risco se torna significativo se nesse exame forem encontrados e retirados um ou mais pólipos a colonoscopia deverá ser repetida no ano seguinte a discordância no entanto quanto ao intervalo ideal para a repetição nos casos em que o exame anterior foi normal pesquisadores da universidade de indiana acabam de publicar um estudo no o jornal de medicina da nova inglaterra sobre essa questão em mulheres e homens com idade média de 56,7 anos os autores realizaram 2436 colonoscopias em que nenhum pólipo foi encontrado num período médio de 5,34 anos o exame foi repetido nesse segundo exame um ou mais adenomas pequenos foram encontrados em 16 % dos casos e adenomas maiores do que 1 cm em 1,5 % não houve um caso sequer de câncer de cólon conclusão repetir colonoscopias de resultado normal antes de 5 anos é exagero de indicação a sociedade americana de câncer vai mais longe sugere que nessa eventualidade elas sejam repetidas 10 anos mais tarde

Apêndice B

Código-fonte para Avaliar o Google *Web Speech* API

A seguir é apresentado o código-fonte, em C#, para gerar as transcrições dos experimentos preliminar e final do SRAF da Google *Web Speech* API:

```
using System;
using System.Net;
using System.IO;

namespace GoogleSpeechAPI
{
    class Program
    {
        static void Main(string[] args)
        {
            try
            {
                FileStream fileStream = File.OpenRead("audio.flac");
                MemoryStream memoryStream = new MemoryStream();
                memoryStream.SetLength(fileStream.Length);
                fileStream.Read(memoryStream.GetBuffer(), 0, (int)fileStream.Length);
                byte[] BA_AudioFile = memoryStream.GetBuffer();
                HttpWebRequest _HWR_SpeechToText = null;
                _HWR_SpeechToText =
                            (HttpWebRequest)HttpWebRequest.Create(
                                "https://www.google.com/speech-
api/v2/recognize?output=" +
                                "json&lang=pt-
BR&key=AIzaSyBJ6VJ326Rpb23msih2wGhXENEwU1TF1PA&client=" +
                                "chromium&maxresults=1&pfilter=2");
                _HWR_SpeechToText.Credentials = CredentialCache.DefaultCredentials;
                _HWR_SpeechToText.Method = "POST";
                _HWR_SpeechToText.ContentType = "audio/x-flac; rate=44100";
                _HWR_SpeechToText.ContentLength = BA_AudioFile.Length;
                Stream stream = _HWR_SpeechToText.GetRequestStream();
                stream.Write(BA_AudioFile, 0, BA_AudioFile.Length);
                stream.Close();

                HttpWebResponse HWR_Response =
(HttpWebResponse)_HWR_SpeechToText.GetResponse();
                if (HWR_Response.StatusCode == HttpStatusCode.OK)
                {
                    StreamReader SR_Response = new
StreamReader(HWR_Response.GetResponseStream());
                    Console.WriteLine(SR_Response.ReadToEnd());
                }
            }
```

```
            catch (Exception ex)
            {
                Console.WriteLine(ex.ToString());
            }
            Console.ReadLine();
        }
    }
}
```

Apêndice C

Técnicas de Modelagem Acústica

A Tabela C.1 contém os classificadores utilizados para a modelagem acústica. Na coluna da direita pode-se verificar a quantidade de cada tecnologia utilizada.

Tabela C.1: Quantidade de classificadores utilizados para a modelagem acústica.

Modelagem Acústica	**Quantidade**
Modelo Oculto de Markov – *Hidden Markov Model* (HMM)	240
Modelo de Mistura Gaussiana – *Gaussian Mixture Model* (GMM)-HMM	25
Rede Neural Profunda – *Deep neural Network* (DNN)	24
Perceptron Multicamadas – *Multi-layer Perceptron* (MLP)-HMM	14
DNN-HMM	10
GMM	10
MLP	5
Modelos Oculto de Markov de Densidade Contínua – *Continuous Density Hidden Markov Models* (CDHMM)	5
Modelos de Mistura Gaussiana Subespacial – *Subspace Gaussian Mixture Models* (SGMM)	5
Rede de Neural Convolucional – *Convolutional Neural Network* (CNN)	3
CNN-DNN	3
DNN-GMM-HMM	2
Rede Neural Recorrente-Memória Longa de Curto Prazo (LSTM)	2
Rede Neural Recorrente (RNN)-LSTM	2
Rede Neural Artificial – *Artificial Neural Network* (ANN)	2
Bloco Diagonal Infinito de Modo Oculto de Markov – *Block Diagonal Infinite Hidden Markov Mode* (BDiHMM)	1
Convolucional LSTM redes neurais profundas – *Convolutional* LSTM *Deep Neural Networks* (CLDNN)	1
CNN-HMM	1
Distorcido – *Fuzzy*-Entortamento de Tempo Dinâmico – *Dynamic Time Warping* (DTW)	1
DNN-GMM	1
DTW	1
GMM-RNN-HMM	1
GMM-MLP	1
LSTM-LSTM	1
Mapeamento Estocástico Baseado em Estéreo – *Stereo-Based Stochastic Mapping* (SSM)-HMM	1
Máquina de Vetor de Suporte – *Support Vector Machine* (SVM)-HMM	1

(Continuação)

Modelagem Acústica	**Quantidade**
MLP-GMM-HMM	1
MLP-GMM-SGMM	1
MLP-SGMM-HMM	1
Modelo Oculto de Markov Condicionado – *Conditioned Hidden Markov Model* (CHMM)	1
Multi-distribuição de Redes Neurais Profundas – *Multi-distribution Deep Neural Networks* (MD-DNN)	1
Perceptron Multicamadas – *Multilayer Perceptron* (MLP)-HMM	1
Rede Neural – *Neural Network* (NN)-HMM	1
Rede Neural Profunda Convolução Convolucional – *Convolutional Deep Neural Network* (CDNN)	1
Rede Neural Profunda Convolução Convolucional – *Convolutional Deep Neural Network* (CDNN)-HMM	1
Redes de Crenças Profundas – *Deep Belief Networks* (DBN)-HMM	1
Regressão Temporal de Variação de Peso – *Temporally Varying Weight Regression* (TVWR)	1
GMM-RNN-LSTM	1

Apêndice D

Técnicas de Extração de Características

A Tabela D.1 contém as técnicas utilizadas para extrair as características do sinal da fala. Na coluna da direita pode-se verificar a quantidade de cada tecnologia utilizada.

Tabela D.1: Quantidade de técnicas de extração de características.

Extração de Característica	**Quantidade**
Coeficientes Cepstral de Frequência Mel – *Mel-frequency Cepstral Coefficients* (MFCC)	258
Não informado	258
Percepção Linear Preditiva – *Perceptual Linear Prediction* (PLP)	31
PLP-Espectro Relativo – *RelAtive SpecTrAl* (RASTA)-Espectrograma de Modulação – *Modulation Spectrogram* (MSG)	14
Coeficientes de Previsão Lineares – *Linear Prediction Coefficients* (LPC)	8
Análise preditiva de percepção espectral linear relativa – *Relative Spectral-perceptual Linear Predictive Analysis* (RASTA)-PLP	3
Coeficientes Cepstral de Frequência Gammatone – *Gammatone frequency cepstral coefficients* (GFCC)	2
Amplitude Quadrada Média – *Mean Square Amplitude* (MSE)	1
Banco de Filtro Gammachirp – *Gammachirp Filterbank* (GF)-PLP	1
Banco de Filtro Mel – *Mel-filterbank* (MFB)	1
Características Gammatone – *Gammatone Features* (GT)	1
Coeficiente Cepstral de Modulação Normalizada – *Normalized Modulation Cepstral Coefficient* (NMCC)	1
Coeficientes Amortecidos do Oscilador – *Damped Oscillator Coefficients* (DOC*)*	1
Coeficientes Cepsrais Sincronizado do Oscilador Amortecido – *Synchronized Damped Oscillator Cepstral Coefficients* (SyDOCC)	1
Coeficientes de Filtro Gammatone – *Gammatone Filter Coefficients* (GFCS)	1
Coeficientes de Modulação Normalizados – *Normalized Modulation Coefficients* (NMC)	1
Coeficientes de Transformação do Cosseno Discreto – *Discrete Cosine Transformation Coefficients* (DCTC)	1
Correlação Automática de Fase – *Phase Auto Correlation* (PAC)-MFCC	1
Distribuição Multivariada de Laplace – *Multivariate Laplace Distribution* (MLD)	1
Entortamento de Tempo Dinâmico – *Dynamic Time Warping* (DTW)	1
Escala Log de Espectro Mel – *Mel Scale Log Spectrum* (MSLS)	1

(Continuação)

Extração de Característica	Quantidade
Espectro MVDR Robusto-Perceptual de Sequência de Autocorrelação Relativa – *Robust-Perceptual* MVDR *Spectrum of Relative Autocorrelation Sequence* (R-PMSR)	1
Espectro MVDR Robusto-Perceptual de Sequência de Autocorrelação Relativa – *Robust-Perceptual* MVDR *Spectrum of Relative Autocorrelation Sequence* (R-PMSR)	1
Filtrado por frequência – *Frequency-filtered* (FF)	1
Frequência Mel de Previsão Linear Perceptiva – *Mel-frequency Perceptual Linear Prediction* (MF-PLP)	1
GCCC	1
GCMC	1
GTCC	1
GTMC	1
LPC-alisado – *smoothed* – MFCC	1
LPCEPSTRA	1
Média Teager–Kaiser – *Mean Teager-Kaiser* (MTE)	1
MFCC-Dimensão Fractal – *Fractal Dimension* (FD)	1
MFCC-PLP	1
Modulação de Amplitude de Fala de Duração Média – *Modulation of Medium Duration Speech Amplitude* (MMeDuSA)	1
MVLPREF	1
NSGT	1
PLP-HLDA	1
PLP-Variância Média – *Mean Variance* ARMA (MVA)	1
Poder Normalizado de Coeficientes Cepstral – *Power Normalized Cepstral Coefficients* (PNCC)	1
Spectrogramas de Modulação de amplitude – *Amplitude Modulation Spectograms* (AMS)	1
Transformação de Probabilidade Linear Máxima – *Maximum Linear Likelihood Transform* (MLLT)	1
VTL-distorcido – *warped* – PLP	1

Apêndice E

Métricas para Avaliar a Precisão dos Sistemas de Reconhecimento Automático de Fala

Na Tabela E.1 é apresentada as métricas para avaliar a precisão dos SRAFs. Na coluna da direita pode-se verificar a quantidade de cada métrica utilizada.

Tabela E.1: Quantidade de métricas utilizadas para avaliar a precisão dos Sistemas de Reconhecimento Automático de Fala.

Métrica	**Quantidade**
Não Informado	222
Taxa de Erro de Palavra – *Word Error Rate* (WER)	177
Taxa de Reconhecimento de Palavra – *Word Recognition Rate* (WRR)	96
Taxa de Erro de Classificação – *Classification Error Rate* (CER)	6
Taxa de Erro do Fonema – *Phone Error Rate* (PER)	5
Reconhecimento de Fonema Correto – *Correct Phone Recognition*	2
Taxa de Erro Igual – *Equal Error Rate* (EER)	2
Erro de Palavra Real – *Real-Word Error*	1
Precisão de Previsão de Fase – *Phase Prediction Accuracy*	1
Sobreposição de Palavra Simples – *Simple Word Overlap* (SWO)	1
Taxa de Detecção de Erro - *Miscue Detection Rate* (MDR)	1
Taxa de Erro – *Error Rate*	1
Taxa de Erro de Nome – *Name Error Rate* (NER)	1
Taxa de Erro de Quadro – *Frame Error Rate* (FER)	1
Taxa de Palavra Correta – *Correct Word Rate* (CWR)	1

Apêndice F

Corpora para o Treinamento dos Sistemas de Reconhecimento Automático de Fala

Na Tabela F.1 é apresentado os principais *corpora* utilizados para o treinamento dos Modelos Acústicos dos SRAFs. Na coluna da direita pode-se verificar a quantidade de SRAFs que utilizaram o *corpus* para o seu treinamento.

Tabela F.1: Quantidade dos *Corpora* utilizados para o treinamento dos Sistemas de Reconhecimento Automático de Fala para a Língua Portuguesa do Brasil.

Corpus	**Quantidade**
West Point	2
CETUC	1
CSLU: *Spoltech Brazilian Portuguese Version* 1.0	1
LapsBenchmark	1
LapsNews	1
LapsStory	1
Spoltech	1
UFPAdic	1