UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

CASSIO TRINDADE BATISTA

Uma Proposta de Sistema de Controle Remoto Universal e Multimodal Baseado em Gestos da Cabeça e Comandos de Voz

Belém – Pará
2017

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

CASSIO TRINDADE BATISTA

Uma Proposta de Sistema de Controle Remoto Universal e Multimodal Baseado em Gestos da Cabeça e Comandos de Voz

Dissertação apresentada à Universidade Federal do Pará, como parte dos requisitos do Programa de Pós-Graduação em Ciência da Computação, para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Sistemas Inteligentes

Orientador: Prof. Dr. Nelson Cruz Sampaio Neto

Belém – Pará

2017

Dados Internacionais de Catalogação na Publicação (CIP)
Sistemas de Bibliotecas da Universidade Federal do Pará
Gerada automaticamente pelo módulo Ficat, mediante os dados fornecidos pelo autor.

B333p Batista, Cassio Trindade
Uma Proposta de Sistema de Controle Remoto Universal e Multimodal Baseado em Gestos da Cabeça e Comandos de Voz / Cassio Trindade Batista. – 2017.
74 f. : il. (algumas color.)

Dissertação (Mestrado) – Programa de Pós-Graduação em Ciência da Computação (PPGCC), Instituto de Ciências Exatas e Naturais, Universidade Federal do Pará, Belém, 2017.
Orientação: Prof. Dr. Nelson Cruz Sampaio Neto

1. Estimação da pose da cabeça. 2. Reconhecimento automático de voz. 3. Tecnologia Assistiva. 4. Controle remoto universal. 5. Interfaces multimodais. I. Sampaio Neto, Nelson Cruz, *orient*. II. Título

CDD 005.28

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

CASSIO TRINDADE BATISTA

UMA PROPOSTA DE SISTEMA DE CONTROLE REMOTO UNIVERSAL E MULTIMODAL BASEADO EM MOVIMENTOS DA CABEÇA E COMANDOS DE VOZ

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal do Pará como requisito para obtenção do título de Mestre em Ciência da Computação, defendida e aprovada em 06/12/2017, pela banca examinadora constituída pelos seguintes membros:

Nelson Cruz Sampaio Neto.
Prof. Dr. Nelson Cruz Sampaio Neto
Orientador – PPGCC/UFPA

Adalb Castro
Prof. Dr. Adalbery Rodrigues Castro
Membro Externo – UFPA

Profa. Dra. Andréa da Silva Miranda
Membro Externo – UFRA

Visto: Bianchi Serique Meiguins
Prof. Dr. Bianchi Serique Meiguins
Coordenador do PPGCC/UFPA

Dedico esta dissertação à minha família.

AGRADECIMENTOS

Agradeço primeiramente a Deus por toda sabedoria e força a mim dadas, pois sozinho seria impossível chegar até aqui.

À minha família, por toda ajuda e apoio que foram de fundamental importância tanto para a minha formação pessoal quanto profissional. Em especial, aos meus pais Rosa Batista e Raimundo Batista, pela paciência, compreensão, dedicação, amor incondicional e tantas outras virtudes demonstradas que eu não conseguiria aqui citar. Agradeço também a minha irmã, meus avós e todos os meus tios, que sempre estiveram e estão, direta ou indiretamente, ao meu lado. Espero que degrau mais importante da minha carreira continue sempre sendo o primeiro.

Ao meu colega Erick Campos, pela enorme contribuição para a realização deste trabalho; e ao meu orientador, Nelson Neto, por tantas oportunidades concedidas; pela paciência durante mais de três anos orientando minha vida acadêmica; e por ter confiado nos meus *brainstorms* e me dado todas as condições para fazê-los crescer, não apenas como orientador, mas como colega de trabalho. Sem esse "time", este projeto certamente não teria passado de uma ideia no papel.

A todas as pessoas que se disponibilizaram como voluntários para testar o protótipo do sistema, sem as quais os resultados e conclusões deste trabalho não teriam sido possíveis. Agradeço também aos profissionais do Abrigo Especial Calabriano e da Associação Paraense de Pessoas com Deficiência, em especial à Kelly Pinheiro e à Érilen, as quais foram extremamente solícitas durante o contato com os pacientes.

Deixo um agradecimento também aos meus amigos que me ajudaram na divulgação, pelo Twitter, da página do projeto na plataforma Hackaday.io, o que nos deu certa visibilidade através da publicação de matérias em *blogs* e de prêmios ganhos em competições de incentivo à comunidade de *makers*. Não cito nomes, pois são muitos, mas agradeci diretamente a quem curtiu e deu *retweet* nas minhas persistentes postagens.

Por fim, agradeço a todos os meus amigos que fiz durante todos os períodos da vida até então, os quais não conseguiria listar e, por isso, não ouso nomeá-los, mas que sem os quais o caminho até aqui não teria sido o mesmo.

Cassio Trindade Batista

“Cada louco traz em si o seu mundo e para ele não há mais semelhantes;
o que foi antes da loucura é outro muito outro do que ele vem a ser após.”
(Lima Barreto)

RESUMO

A evolução tecnológica, num sentido amplo, caminha para que as pessoas possam interagir com os equipamentos eletrônicos de forma mais fácil. Por outro lado, para pessoas com deficiência, essa interação torna-se algo mais do que simples: ela se torna possível. Este trabalho apresenta uma proposta de um sistema de controle remoto universal, multimodal, livre e de baixo custo que permite o controle de dispositivos eletrônicos por de comandos de voz e gestos da cabeça. Além disso, um circuito sensor de proximidade foi combinado a módulos de rádio-frequência para agir, em paralelo a uma palavra-chave dada através de comando de voz, como um acionador externo sem fio. Dentre as aplicações encontradas a respeito de controles remotos alternativos para aparelhos eletrônicos convencionais (que recebem informação remota via luz infravermelha), nenhuma dá suporte a gestos da cabeça como entrada, tampouco à combinação deste método de interação com a voz, o que certamente seria uma opção viável para pessoas cujos membros superiores são comprometidos. Uma avaliação objetiva foi realizada especificamente sobre as técnicas implementadas como método de entrada do sistema. Além disso, um teste subjetivo, baseado em um questionário mean opinion score, também foi aplicado com objetivo de avaliar o protótipo do sistema de modo geral. Como resultado, observou-se que a interação por comandos de voz foi mais acertiva e precisa do que a por gestos da cabeça, apesar de este último ter sido a forma de controle preferida pelos usuários. A análise também mostrou grande interesse da parte dos participantes em um possível produto final.

Palavras-chaves: Estimação da pose da cabeça. Reconhecimento automático de voz. Tecnologia assistiva. Controle remoto universal. Interfaces multimodais. Open-source. Mean opinion score.

ABSTRACT

Technological developments converge to make people interact with electronic devices in an easy way. For people with disabilities, however, that interaction becomes something more than simple: it becomes possible. The current work presents a proposal of a low-cost, open-source, multimodal, universal remote control system that allows users to control electronic devices through voice commands and head gestures. In addition, a proximity sensor circuit was combined to radio-frequency modules to act in parallel with a spoken keyword as a wireless AT switch. Among the applications found with respect to alternative remote controls for conventional electronic devices (i.e., devices that receive information commands via infrared light), none of them supports head gestures as input, neither the combination of head gestures with voice commands, which would certainly be a viable option for people whose upper limbs are compromised. An objective evaluation was performed over the techniques implemented as system input. Besides, a mean opinion score questionnaire was applied to volunteers in order to subjectively evaluate the overall system prototype. As result, it was noticed that the interaction via speech was more accurate than the one via head gestures, despite the latter have been chosen as preferred method of control by the users. The analysis also showed great interest of the users in a possible final product in the future.

Key-words: Head pose estimation. Automatic speech recognition. Assistive technology. Universal remote control. Multimodal interfaces. Open-source. Mean opinion score.

LISTA DE FIGURAS

Figura 1 – Eixos tridimensionais de rotação (poses) da cabeça. 25
Figura 2 – Visão geral do sistema de reconhecimento de gestos da cabeça. 26
Figura 3 – Haar-*like features* 26
Figura 4 – Imagem integral. 27
Figura 5 – Arquitetura de um sistema ASR. 29
Figura 6 – Exemplos de PWMs de mesmo período com diferentes *duty cycles*. 33
Figura 7 – Sinal infravermelho do protocolo da Samsung. 34
Figura 8 – Microcomputador C.H.I.P. 35
Figura 9 – Arduino UNO e *shield* Bluetooth SHD 18, baseado no módulo HC-05. . . . 38
Figura 10 – Arquitetura do sistema de controle remoto proposto. 41
Figura 11 – Sensor de proximidade baseado em luz infravermelha. 42
Figura 12 – Diagrama do módulo transmissor (Tx) do acionador externo sem fio. 43
Figura 13 – Diagrama do módulo receptor (Rx) do acionador externo sem fio. 43
Figura 14 – Detecção facial, realizada pelo algoritmo Viola-Jones, e cálculo heurístico da posição dos olhos e do nariz sob duas condições de iluminação diferentes, com e sem o uso de óculos. 45
Figura 15 – Algoritmo implementado para detecção facial. 45
Figura 16 – Algoritmo implementado para estimação da pose da cabeça. 46
Figura 17 – Perfil demográfico dos participantes envolvidos nos testes com o sistema unimodal, somente com o módulo de reconhecimento de gestos da cabeça. . 52
Figura 18 – Perfil demográfico dos participantes envolvidos nos testes do sistema multimodal, já com ambos os módulos de controle por gestos e voz. 53
Figura 19 – Movimentos de rotação da cabeça (ou poses) traduzidos em comandos para um televisor. 54
Figura 20 – Número de tentativas de realização dos movimentos da cabeça por pessoas sem deficiência durante o teste com o sistema unimodal. 56
Figura 21 – Número de tentativas de realização dos movimentos da cabeça por pessoas com deficiência motora durante o teste com o sistema unimodal. 56
Figura 22 – Número de tentativas de realização dos movimentos da cabeça por pessoas com deficiência motora dos membros superiores durante o teste com o sistema ainda unimodal. 57
Figura 23 – Número de tentativas de realização dos movimentos da cabeça por pessoas sem deficiência durante o teste com o sistema multimodal. 58
Figura 24 – Número de tentativas de realização dos comandos de voz por pessoas sem deficiência durante o teste com o sistema multimodal. 58

Figura 25 – Deslocamento da localização do nariz ao longo do eixo x durante o movimento *roll*, contrastando com o modelo de rotação esperado. 60

LISTA DE TABELAS

Tabela 1 – Perfil da população brasileira com deficiência. 19
Tabela 2 – Modos dos pinos 3, 25 e 35 do *header* U14 do C.H.I.P.. 37
Tabela 3 – Escala de avaliação baseada no MOS. 40
Tabela 4 – Ações de comando na televisão e seus respectivos comandos de voz. 47
Tabela 5 – Roteiro de tarefas a serem completadas pelo usuário utilizando gestos da cabeça como entrada. 54
Tabela 6 – Roteiro contendo as ações na TV e seus respectivos comandos de voz. . . . 55
Tabela 7 – Resultado da avaliação subjetiva do sistema unimodal (somente gestos da cabeça) através do questionário MOS. 59
Tabela 8 – Resultado da avaliação subjetiva do sistema multimodal (com entrada por gestos da cabeça e comandos de voz) através do questionário MOS. 59
Tabela 9 – Funcionalidades comuns em dispositivos eletrônicos. 65

LISTA DE ABREVIATURAS E SIGLAS

ADA	*Americans with Disabilities Act*
ARM	*Advanced RISC Machines*
API	*Application Programming Interface*
ASR	*Automatic Speech Recognition*
AGR	*Active Gesture Recognition*
AHRS	*Attitude and Heading Reference System*
BCI	*Brain-Computer Interface*
BSD	*Berkeley Software Distribution*
CAT	Comitê de Ajudas Técnicas
CC BY-SA	*Creative Commons Attribution-ShareAlike*
CMU	*Carnegie Mellon University*
DNN	*Deep Neural Networks*
DTW	*Dynamic Time Warping*
DRAM	*Dynamic Random Access Memory*
EMG	Eletromiografia
FSG	*Finite State Grammar*
G2P	*Grapheme to Phoneme*
GPIO	*General Purpose Input/Output*
GPL	*GNU General Public License*
HMM	*Hidden Markov Model*
IBGE	Instituto Brasileiro de Geografia e Estatística
IEEE	*Institute of Electrical and Electronics Engineers*
IHC	Interação Humano-Computador
IR	*Infrared*

JSGF	*Java Speech Grammar Format*
LED	*Light Emitting Diode*
LiPo	*Lithium Polymer*
LIRC	*Linux Infrared Remote Control*
MFCC	*Mel-Frequency Cepstral Coefficients*
MOS	*Mean Opinion Score*
NAND	*Not AND*
NPN	*Negative-Positive-Negative*
NTSC	*National Television System Comittee*
OpenCV	*Open Source Computer Vision Library*
PCD	Pessoas com Deficiência
PAL	*Phase Alternanting Line*
PT_BR	Português Brasileiro
PIR	*Pyroelectric Infrared*
PWM	*Pulse Width Modulation*
RISC	*Reduced Instruction Set Computer*
RF	Rádio-Frequência
SRAM	*Static Random Access Memory*
SSH	*Secure Shell*
TTS	*Text to Speech*
UART	*Universal Asynchronous Receiver-Transmitter*
UFPA	Universidade Federal do Pará
UN	*United Nations*
USB	*Universal Serial Bus*
VS	*Vishay*
WHO	*World Health Organization*

SUMÁRIO

1 Introdução 16
1.1 Contextualização e Terminologias 16
1.2 Justificativa 18
1.2.1 Trabalhos Relacionados 20
1.2.2 Contribuições 22
1.3 Objetivos 22
1.3.1 Objetivos Específicos 23
1.4 Síntese de Conteúdos 24
2 Referencial Teórico 25
2.1 Reconhecimento de Gestos da Cabeça 25
2.1.1 Viola-Jones e Relações Antropométricas 26
2.1.2 Lucas-Kanade e Relações Trigonométricas 28
2.2 Reconhecimento Automático de Voz 28
2.2.1 Redes Neurais Profundas 31
2.3 Acionador Externo 31
2.4 Protocolos de Comunicação via Luz Infravermelha 32
2.4.1 Modulação por Largura de Pulso 32
2.4.2 O Protocolo da Samsung 33
2.5 Ferramentas de *Hardware* e Plataformas Embarcadas 34
2.5.1 C.H.I.P. 35
2.5.1.1 Expansão da Pinagem dos Headers U13 e U14 35
2.5.2 Arduino 37
2.6 Ferramentas de *Software* 38
2.7 Formas de Avaliação 39
3 Protótipo 41
3.1 Acionador Externo 42
3.2 Reconhecimento de Gestos da Cabeça 44
3.2.1 Detecção Facial 44
3.2.2 Estimação da Pose da Cabeça 45
3.3 Reconhecimento de Voz 47
3.3.1 Acionamento por Palavra-Chave 48
3.4 Controle Remoto do Televisor 49
4 Testes e Resultados 51
4.1 Perfil dos Participantes 51
4.2 Ambiente de Testes 53

4.2.1 Gestos da Cabeça . . . 53
4.2.2 Comandos de Voz . . . 55
4.3 Resultados . . . 55
4.3.1 Avaliação Objetiva: Sistema Unimodal . . . 55
4.3.2 Avaliação Objetiva: Sistema Multimodal . . . 57
4.3.3 Avaliação Subjetiva . . . 59
4.4 Discussão . . . 60
5 Considerações Finais . . . 62
5.1 Dificuldades e Soluções . . . 63
5.2 Trabalhos Futuros . . . 65

Referências . . . 68

1 INTRODUÇÃO

1.1 Contextualização e Terminologias

Com o passar dos anos, a interação das pessoas com as máquinas tem-se tornado cada vez mais simples. O que antes era portado somente por empresas e pessoas com poder financeiro e conhecimento acima da média, em termos de tecnologia, é hoje muito mais acessível para usuários domésticos sem profundo entendimento do assunto. Devido à abrangência dos computadores pessoais e da Internet, novas oportunidades e expectativas em termos de trabalho, estudo e lazer são criadas a fim de melhorar ainda mais essa comunicação, de modo que a máquina se aproxime mais de ações típicas do ser humano, como pensar, agir e falar.

Acredita-se que as tecnologias de reconhecimento de gestos (AGR, do inglês *active gesture recognition*) (DARRELL; PENTLAND, 1996) e de reconhecimento de voz (ASR, do inglês *automatic speech recognition*) (HUANG; ACERO; HON, 2001) tornam a interface humano-computador (IHC) muito mais prática e natural, de forma que a comunicação assemelha-se, de fato, àquela estabelecida entre duas pessoas. O AGR refere-se ao sistema que, utilizando técnicas de computação visual para processar imagens recebidas como entrada, é capaz de definir na saída um significado inteligível a uma ação referente ao movimento motor realizado pelo usuário. Já o sistema ASR refere-se à geração de um texto transcrito quando um sinal de fala digitalizado é posto na entrada. Dentre os inúmeros sistemas que envolvem reconhecimento de gestos e de voz, destacam-se os de automação residencial, os quais promovem comodidade e praticidade às pessoas no controle de equipamentos eletrônicos; e, ainda, os que visam ajudar pessoas que tenham dificuldades na utilização desses equipamentos. Normalmente enquadrados no conceito de Tecnologia Assistiva, tais sistemas tornam-se umas das poucas soluções acessíveis às pessoas com deficiência motora, cujas limitações concentram-se principalmente na locomoção e manipulação de certos objetos.

A Tecnologia Assistiva é um campo do conhecimento de característica interdisciplinar dedicado a aumentar a independência e mobilidade de pessoas com deficiência (PCD), englobando produtos, metodologias, práticas e serviços que objetivam promover sua autonomia, qualidade de vida e inclusão social (BRASIL, 2009b; GELDERBLOM; WITTE, 2002). O termo refere-se à tecnologia que fornece assistência a pessoas com diferentes tipos e graus de limitação, a fim de reduzir os efeitos das deficiências e permitindo-lhes participar ativamente das suas respectivas rotinas. Segundo a Organização das Nações Unidas (UN, do inglês *United Nations*), os olhares do século XXI deixam de enxergar as PCD como “objetos” de caridade, tratamento médico e proteção social, e passam a vê-las sob uma nova perspectiva, na qual as PCD são “sujeitos” ativos na sociedade, capazes de reivindicar seus direitos e tomar decisões baseadas em seu próprio consentimento (UN, 2007).

A Tecnologia Assistiva busca, então, reduzir a dificuldade vivenciada por pessoas que precisam de soluções que não as deixem à margem da utilização de dispositivos eletrônicos. Visando diminuir a exclusão digital imposta às PCD pela dificuldade ou total incapacidade para manipular certos equipamentos, a acessibilidade é vista como elemento fundamental para elevar a autoestima e o grau de independência dessas pessoas. Além disso, as soluções apresentadas também podem ser úteis para as pessoas sem qualquer deficiência, já que o controle de equipamentos torna-se mais prático, confortável e natural (WECHSUNG; NAUMANN, 2009).

Nesse sentido, este trabalho buscou desenvolver um sistema de controle remoto universal de baixo custo baseado nas plataformas embarcadas C.H.I.P. e Arduino, que foram utilizadas em conjunto como um *hub* centralizado para controlar diversos aparelhos eletrônicos. O sistema é capaz de processar i) imagens de vídeo em tempo real, as quais são a base para o reconhecimento de gestos da cabeça; e ii) áudios digitalizados, repassados ao sistema de reconhecimento de voz; os quais servem como interface alternativa de entrada para promover autonomia às pessoas com deficiência nas tarefas de controle. Além disso, um sensor de proximidade foi combinado a módulos de rádio-frequência (RF) para atuar como um acionador externo sem fio, utilizado com o propósito de ativar o sistema de controle.

Os usuários-alvo em potencial são pessoas com limitações motoras dos membros superiores, e que tenham o cognitivo preservado juntamente com os movimentos do pescoço (para o caso do sistema AGR) ou a fala (para o sistema ASR). A ideia é que esse público consiga realizar, através de seis movimentos básicos com a cabeça ou, no mínimo, seis comandos de voz, ações do tipo ligar/desligar, aumentar/diminuir e avançar/recuar em uma grande variedade de equipamentos eletrônicos disponíveis no ambiente residencial. Vale ressaltar que todas as interfaces de programação (APIs, do inglês *application programming interfaces*), plataformas embarcadas e pacotes de *software* utilizados para a criação dos módulos e recursos do sistema são *open-source* e *open-hardware* ou encontram-se disponíveis livremente na Internet.

O Ato de Americanos com Deficiência (ADA, do inglês *Americans with Disabilities Act*) (UNITED STATES, 1990) é um documento publicado em 1990 que regula os direitos dos cidadãos com deficiência nos Estados Unidos, além de prover a base legal dos fundos públicos para compra dos recursos que tais pessoas necessitam. Já o Comitê de Ajudas Técnicas (CAT) (BRASIL, 2009b), criado em 2006, executa papel semelhante, tendo a perspectiva de aperfeiçoar, regulamentar e dar transparência e legitimidade ao desenvolvimento da Tecnologia Assistiva no Brasil com base na legislação da Constituição Federal, difundindo-a por entre a maior parte das esferas da sociedade, como saúde, educação, emprego, cultura, etc.

O projeto, por apresentar um sistema de controle remoto com suporte a reconhecimento de gestos e de voz, enquadra-se em duas categorias da Tecnologia Assistiva (ASSISTIVA, 2017) criadas a partir da ADA e atualmente existentes nas diretrizes do comitê, brevemente descritas a seguir: i) recursos de acessibilidade ao computador; e ii) sistemas de controle de ambiente.

i) Recursos de acessibilidade ao computador: equipamentos de entrada e saída (síntese de voz, Braille), auxílios alternativos de acesso (ponteiras de cabeça, de luz), teclados modificados, *softwares* especiais (reconhecimento de voz, etc.), que permitem as pessoas com deficiência a usarem o computador.

ii) Sistemas de controle de ambiente: sistemas eletrônicos que permitem as pessoas com limitações moto-locomotoras controlar remotamente aparelhos eletro-eletrônicos, sistemas de segurança, etc., localizados em seu quarto, sala, escritório, casa e arredores.

1.2 Justificativa

A definição de deficiência, no âmbito jurídico brasileiro, é a mesma presente na Convenção Internacional sobre os Direitos das Pessoas com Deficiência (UN, 2007) promulgada pelo Decreto N° 6.949 (BRASIL, 2009a), a qual tem *status* de emenda constitucional no Brasil segundo as regras do art. 5° § 3° da Constituição Federal (BRASIL, 1988).

> Pessoas com deficiência são aquelas que têm impedimentos de longo prazo de natureza física, mental, intelectual ou sensorial, os quais, em interação com diversas barreiras, podem obstruir sua participação plena e efetiva na sociedade em igualdades de condições com as demais pessoas (BRASIL, 2009a, art. 1°).

A Convenção das Nações Unidas também aborda o conceito social de deficiência, com foco nas barreiras que as PCD encontram tanto nas relações com o meio quanto nas relações em sociedade, conforme o preâmbulo do mesmo Decreto:

> *Reconhecendo* que a deficiência é um conceito em evolução e que a deficiência resulta da interação entre pessoas com deficiência e as barreiras devidas às atitudes e ao ambiente que impedem a plena e efetiva participação dessas pessoas na sociedade em igualdade de oportunidades com as demais pessoas [...] (BRASIL, 2009a, preâmbulo, e).

O Decreto N° 3.298, que compreende as normas que asseguram os direitos das PCD no Brasil sob a chamada Política Nacional para a Integração da Pessoa Portadora de Deficiência (BRASIL, 1999), também considera os conceitos de incapacidade, deficiência de forma geral e, especificamente, deficiência física:

> Deficiência – toda perda ou anormalidade de uma estrutura ou função psicológica, fisiológica ou anatômica que gere incapacidade para o desempenho de atividade, dentro do padrão considerado normal para o ser humano (BRASIL, 1999, art. 3°, I).
>
> Incapacidade – uma redução efetiva e acentuada da capacidade de integração social, com necessidade de equipamentos, adaptações, meios ou recursos especiais para que a pessoa portadora de deficiência possa receber ou transmitir informações necessárias ao seu bem-estar pessoal e ao desempenho de função ou atividade a ser exercida (BRASIL, 1999, art. 3°, III).

> Deficiência física – alteração completa ou parcial de um ou mais segmentos do corpo humano, acarretando o comprometimento da função física, apresentando-se sob a forma de paraplegia, paraparesia, monoplegia, monoparesia, tetraplegia, tetraparesia, triplegia, triparesia, hemiplegia, hemiparesia, ostomia, amputação ou ausência de membro, paralisia cerebral, nanismo, membros com deformidade congênita ou adquirida, exceto as deformidades estéticas e as que não produzam dificuldades para o desempenho de funções (BRASIL, 1999, art. 4°, I).

Já a Lei N° 13.146, que institui o Estatuto da Pessoa com Deficiência, considera a avaliação biopsicossocial de deficiência (BRASIL, 2015):

> I – os impedimentos nas funções e nas estruturas do corpo;
> II – os fatores socioambientais, psicológicos e pessoais;
> III – a limitação no desempenho de atividades; e
> IV – a restrição de participação (BRASIL, 2015, art. 2°, §1°).

Segundo dados da Organização Mundial de Saúde (WHO, do inglês *World Health Organization*), aproximadamente 15% da população mundial possui algum tipo de deficiência (WHO, 2015). Esse número é realmente expressivo, pois revela que, em uma população de 7,6 bilhões de pessoas, cerca de um sétimo (1 bilhão de pessoas) é portadora de deficiência. A WHO também afirma que, em 2013, 80% das pessoas com deficiência viviam em países ainda em desenvolvimento, o que sugere que o predomínio da condição de deficiência está bastante relacionado com a situação econômica dos países.

No Brasil, segundo o censo realizado em 2010 pelo Instituto Brasileiro de Geografia e Estatística (IBGE), aproximadamente 23,9% da população (cerca de uma entre quatro pessoas, um total de 46 milhões de habitantes) declarou ter alguma deficiência (IBGE, 2010). Os dados também mostram que, desse total, quase 7% (cerca de de 13,2 milhões) apresentam dificuldades motoras. A Tabela 1 mostra o perfil da população brasileira com deficiência.

Tabela 1 – Perfil da população brasileira com deficiência.

Deficiência	**Descrição**	**Número de Pessoas**	**Porcentagem**
Visual	Cegueira ou dificuldades gerais	35.774.392	18,754 %
Motora	**Paralisia ou dificuldades gerais**	**13.265.599**	**6,95 %**
Auditiva	Surdez ou dificuldades gerais	9.717.318	5,094 %
Cognitiva	Problemas mentais ou intelectuais	2.611.536	1,369 %

Fonte: Extraída de IBGE (2010).

Em Costa et al. (2012), uma entrevista foi realizada com algumas PCD a fim de se construir um "controle de TV ideal" baseado nas características que esses cidadãos gostariam de encontrar em controles remotos convencionais, de modo que suas dificuldades em utilizá-los fossem minimizadas. A sugestão de um *design* mais limpo foi unânime entre os PCD, enquanto

os deficientes visuais, em particular, apontaram a necessidade de algum tipo de *feedback* quando os botões fossem pressionados (como um estalo, por exemplo) e a implementação de alguma referência reconhecível pelo tato nos botões. Já os deficientes motores sugeriram um dispositivo menor, mais leve, que não escorregasse facilmente das mãos, contendo um ângulo de controle mais abrangente (ou seja, que não precisasse ser diretamente apontado para a TV) e com botões maiores para aqueles que não conseguem ter controle absoluto sobre as mãos e/ou dedos.

A expressão gestual e a fala são exemplos de métodos alternativos de interação bastante utilizados em sistemas que utilizam interfaces multimodais. Esses sistemas permitem que o acesso à máquina ocorra de várias maneiras através de dispositivos de entrada e saída de dados que tornam a interação mais natural, não sendo, portanto, limitado apenas aos convencionais botões e teclas. Embora haja diminuição na usabilidade e na rapidez da execução da tarefa quando associadas ao controle de equipamentos, tais técnicas de interação proveem uma experiência muito mais prazerosa e prática quando comparadas aos métodos que forçam o uso das mãos por parte do usuário (WECHSUNG; NAUMANN, 2009). Além disso, os métodos alternativos geralmente são a única maneira que as pessoas com deficiência motora dos membros superiores encontram para interagir com determinados aparelhos de maneira independente.

Nesse sentido, esta pesquisa tem como finalidade apresentar uma solução para diminuir a exclusão digital vivenciada especialmente por pessoas com variados graus de necessidade motora dos membros superiores, as quais estão à margem do mundo eletrônico por conta da ausência de recursos que adaptem os dispositivos às suas necessidades. A tecnologia de reconhecimento de gestos da cabeça tornaria acessível a utilização de qualquer dispositivo eletrônico por usuários que apresentam dificuldades na execução de movimentos específicos com membros superiores, os quais são frequentemente exigidos pelos controles remotos convencionais e acabam tornando-se barreiras, como segurar um controle físico e apertar botões ou digitar. Já o reconhecimento de fala conseguiria ir além, abrangendo um público cuja movimentação dos membros superiores é totalmente impossível, o que torna suas dificuldades proibitivas, como é o caso dos amputados e dos tetraplégicos.

1.2.1 Trabalhos Relacionados

Ajudar pessoas com deficiência a utilizar equipamentos eletrônicos de forma independente não é uma ideia recente. Diversos trabalhos já propuseram soluções para tentar minimizar os mais variados tipos de limitação ou, até mesmo, tornar a forma de interação mais natural. O Navigo (POKHARIYA et al., 2006), por exemplo, foi desenvolvido como um sistema multimodal para deduzir informações relevantes extraídas de pessoas com paralisia cerebral através de reconhecimento de fonemas, objetos e gestos (especialmente dos lábios e sobrancelhas). Em Epelde et al. (2011), a acessibilidade na interação com aparelhos televisores do tipo *smart*, dada por um sistema de diálogo, foi abordada com relação às pessoas idosas com cognitivo considerado normal para a idade. Já em Esquef e Caetano (2007), o nariz foi utilizado como

ponto específico para controlar o cursor de um *mouse* através de técnicas de processamento de imagens, enquanto a função de *click* foi atribuída à transição cerrado-aberto dos lábios.

O sistema Handmote (SOLANKI; DESAI, 2011), por outro lado, foi configurado para processar imagens de gestos das mãos, capturadas por uma câmera infravermelha, para controlar remotamente dispositivos domésticos "não-*smart*" também através da tecnologia infravermelha. A principal motivação dos autores concentrou-se na remoção da necessidade de olhar para o controle, procurando teclas específicas. O controle de aparelhos eletrônicos também foi implementado no ActiveIris (LEVY et al., 2013), no qual técnicas de rastreamento ocular atuavam como método de entrada em uma interface gráfica desenvolvida para plataformas *desktop*. Já em Erden e Cetin (2014), seis movimentos das mãos e antebraços (esquerda-direita, direita-esquerda, cima-baixo, baixo-cima, horário e anti-horário), capturados por três sensores de movimento do tipo PIR, eram reconhecidos também com o intuito de controlar aparelhos eletrônicos de forma mais natural, sem a necessidade do contato físico.

Em Barrena et al. (2012), um sistema ASR foi embarcado em um *smartphone* Android, permitindo a pessoas com lesões na medula espinhal controlar aparelhos domésticos através da interface de voz. O resultado do reconhecedor de fala era enviado para a placa IOIO-OTG, a qual, estando fisicamente conectada ao controle remoto de uma TV, acionava um determinado comando. Outros trabalhos focam em *designs* limpos e uniformes para controles remotos universais também em *smartphones*, como é o caso de Ahsan et al. (2014).

A respeito de sistemas que utilizam os gestos da cabeça como método de entrada, foram encontrados muitos que são utilizados para controlar cadeiras de rodas através de acelerômetros (PAJKANOVIĆ; DOKIĆ, 2013; SRIVASTAVA; CHATTERJEE; THAKUR, 2014) e outros sensores de posição, como os AHRS (do inglês *attitude and heading reference system*) (KUMAR; KUMAR, 2015; LEE; KIM; KIM, 2016) colocados sobre a cabeça do usuário. Já em Althoff et al. (2005), movimentos de aprovação e rejeição em ambientes automotivos foram capturados por uma câmera e processados por de técnicas de computação visual.

Também não foram encontrados trabalhos com foco em desenvolvimento de acionadores externos, tampouco baseados em proximidade e *wireless*. Algumas soluções, no entanto, podem ser encontradas disponíveis para venda no mercado em forma de produtos, como é o caso do HoneyBee (ADAPTIVATION, 2017) e do Candy Corn (ABLENET, 2017). Porém, além de a forma de conexão desses *switches* ser através de um conector de áudio *jack* P2, os preços não são nada acessíveis: U$ 149,00 e U$ 195,00 dólares, respectivamente.

Portanto, como se pode perceber, dentre a literatura consultada, nenhum dos trabalhos encontrados apresentou propostas que utilizam gestos da cabeça — capturados por uma câmera e processados por técnicas de computação visual — com foco em acessibilidade para o controle remoto de aparelhos eletrônicos, tampouco buscou combinar a técnica com um módulo de reconhecimento de voz em um sistema multimodal. A escassez de trabalhos envolvendo acionadores

externos de qualquer natureza também foi sentida, bem como a de referências sobre dispositivos equivalentes sem fio e baseados não em contato físico ou pressão, mas sim em proximidade Além disso, a maioria dos trabalhos encontrados não expôs preocupação a respeito das licenças dos recursos de *software* e *hardware* utilizados durante a implementação, fato que tem um impacto direto no custo de produção e, consequentemente, em um futuro custo de aquisição, especialmente por pessoas de países ainda em desenvolvimento.

1.2.2 Contribuições

A implementação do sistema de controle, unicamente via gestos da cabeça, em conjunto com o acionador baseado em proximidade; e os resultados preliminares obtidos através de testes com o público-alvo foram publicados em uma conferência no formato de artigo completo:

- Cassio Trindade Batista, Erick Modesto Campos e Nelson Cruz Sampaio Neto. "A Proposal of a Universal Remote Control System Based on Head Movements". Em: *IHC – XVI Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais*. Joinville, SC, Brasil. 2017 (BATISTA; CAMPOS; SAMPAIO NETO, 2017).

O projeto também participou das finais da competição *The Hackaday Prize*, patrocinada pelas empresas Microchip, Supplyframe, Texas Instruments e Digi-Key Electronics sob o nome "*TV Remote Control Based on Head Gestures*". No total, 120 projetos (incluindo este) foram selecionados para a final para concorrer a prêmios de incentivo à Tecnologia Assistiva.

Por fim, três *blogs* da área de tecnologia publicaram matérias sobre o projeto:

`blog.arduino.cc` "Controlling a TV with Head Movements"[1]

`blog.hackster.io` "Head Movement Remote Control with C.H.I.P. and Arduino"[2]

`hackaday.com/blog` "Hackaday Prize Entry: Remote Control by Head Gestures"[3]

1.3 Objetivos

O objetivo principal consiste em criar um protótipo portável e barato, baseado em plataformas embarcadas *open-hardware*, que seja capaz de ajudar pessoas com diferentes graus de deficiência motora dos membros superiores a controlar, através de gestos da cabeça e comandos de voz, aparelhos eletrônicos através do envio remoto de sinais. Assim, espera-se que a autonomia e a independência das PCD no controle dos dispositivos seja aumentada, já que o sistema ajudaria as pessoas a contornar as barreiras impostas pelos controles convencionais, que obrigatoriamente forçam o uso das mãos pelos usuários.

[1] <https://blog.arduino.cc/2017/09/16/controlling-a-tv-with-head-movements/>
[2] <https://blog.hackster.io/head-movement-remote-control-with-c-h-i-p-and-arduino-93e89cf481c5>
[3] <https://hackaday.com/2017/09/03/hackaday-prize-entry-remote-control-by-head-gestures/>

A princípio, como prova de conceito, uma única TV da marca Samsung será controlada. No entanto, o sistema projetado pode ser considerado de propósito geral e, futuramente, tornar-se universal ao dar suporte a uma maior variedade de dispositivos.

1.3.1 Objetivos Específicos

Os objetivos específicos podem ser divididos em seis subgrupos principais: i) estudo e implementação das técnicas de reconhecimento de gestos da cabeça; ii) estudo e implementação das técnicas de reconhecimento de voz; iii) desenvolvimento de um acionador externo sem fio baseado em proximidade; iv) estudo do microcomputador embarcado; v) estudo e implementação da comunicação via luz infravermelha; vi) testes com voluntários.

Uma breve descrição sobre cada tópico é feita a seguir.

i. Estudo de bibliotecas e algoritmos capazes de identificar os três principais ângulos de rotação da cabeça:
   - Identificação e análise de algoritmos de AGR via *software* em ambientes *desktop* baseados em Linux.

ii. Estudo de algoritmos e pacotes que realizam o reconhecimento de fala em Português Brasileiro:
   - Identificação e análise de algoritmos de ASR via *software* em ambientes *desktop* baseados em Linux.

iii. Desenvolvimento de um acionador externo:
   - Estudo de sensores eletrônicos de proximidade;
   - Estudo de técnicas de baixo custo para comunicação sem fio de média distância.

iv. Estudo do C.H.I.P.:
   - Estudo do sistema operacional embarcado;
   - Investigação dos módulos AGR e ASR em sistemas embarcados;
   - Estudo do mapeamento dos pinos do processador na plataforma embarcada.

v. Estudo da comunicação entre controles remotos e aparelhos eletrônicos domésticos via luz infravermelha:
   - Descrição do padrão dfinido pelo protocolo da Samsung

vi. Testes com voluntários:
   - Avaliação dos resultados a fim de superar as desvantagens e possivelmente melhorar o sistema nas próximas versões.

1.4 Síntese de Conteúdos

Este capítulo fez uma breve introdução sobre as motivações que levaram ao desenvolvimento do projeto, mostrando com breves detalhes o que foi construído na implementação dos sistemas de reconhecimento de gestos da cabeça e de reconhecimento de voz, além de oferecer razões sobre como os sistemas afetam significativamente a vida das pessoas com comprometimento no movimentos dos membros superiores.

A seguir, é apresentada uma síntese dos conteúdos que serão abordados nos próximos capítulos.

Capítulo 2. Referencial Teórico. Uma introdução dos principais módulos desenvolvidos no trabalho será feita. Serão abordados os tópicos de reconhecimento de gestos da cabeça; reconhecimento automático de voz; descrição das plataformas embarcadas C.H.I.P. e Arduino; padrões de comunicação via luz infravermelha, com ênfase no protocolo da Samsung; técnicas de entrada e saída de dados digitais nos pinos do C.H.I.P. e do Arduino; e as formas de avaliação utilizadas para verificar, de forma quantitativa, a qualidade do sistema.

Capítulo 3. Protótipo. Neste capítulo, as atividades realizadas para a construção do projeto serão abordadas. Serão descritos com detalhes os procedimentos utilizados para desenvolver o acionador externo sem fio baseado em proximidade; implementar o sistema de reconhecimento de voz e de gestos da cabeça, ambos configurados no C.H.I.P., bem como o emulador do protocolo de controle remoto da Samsung no Arduino; e estabelecer a comunicação via Bluetooth entre as duas plataformas embarcadas.

Capítulo 4. Testes e Resultados. O contato com voluntários foi feito com o objetivo de avaliar o desempenho do sistema. Alguns resultados de testes objetivos e subjetivos, frutos da observação do pesquisador e do *feedback* dos usuários, serão apresentados e discutidos nesse capítulo, juntamente com a metodologia dos testes aplicados.

Capítulo 5. Considerações Finais. Breves discussões sobre os resultados dos testes feitos com o grupo de voluntários serão abordadas nesse capítulo, juntamente com a conclusão do trabalho e a perspectiva para as próximas versões do projeto, esta última discutida em uma seção de trabalhos futuros. Detalhes sobre todas as falhas cometidas e dificuldades gerais encontradas durante o projeto também serão relatados.

2 REFERENCIAL TEÓRICO

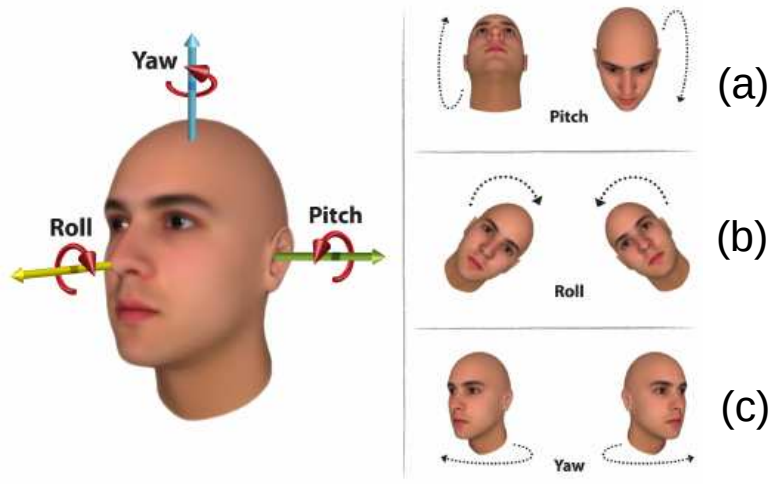
Este capítulo tem como objetivo apresentar o referencial teórico dos módulos mais importantes utilizados e construídos neste trabalho, como i) o reconhecimento de gestos da cabeça; ii) o reconhecimento de voz; iii) o acionador externo; iv) o padrão da comunicação entre o controle remoto e o aparelho eletrônico; v) as plataformas embarcadas (*hardware*); vi) as ferramentas de *software* utilizadas; e, finalmente, vii) as formas de avaliação.

2.1 Reconhecimento de Gestos da Cabeça

Reconhecimento de gestos consiste em reconhecer movimentos relevantes das mãos, braços, cabeça e outras partes do corpo (MITRA; ACHARYA, 2007). A compreensão desses movimentos, para humanos, é uma tarefa trivial. Porém, para um computador, é uma função bastante trabalhosa. Tal dificuldade em realizar o reconhecimento de gestos está em fatores como a grande diversidade de tamanhos do objeto alvo, de cores da pele, posição, iluminação do ambiente, etc. Para "perceber" e "entender" um gesto capturado em um vídeo, é necessário implementar no computador uma série de técnicas de processamento de imagens.

A cabeça, parte do corpo alvo deste trabalho, possui três eixos de rotação, conforme mostrado na Figura 1: (a) *pitch*, que é o dado ângulo formado pelo movimento de subida e descida do nariz; (b) *roll*, rotação semicircular lateral realizada pelo nariz quando as laterais da cabeça se movimentam em direção a um dos ombros; e (c) *yaw*, dado pela rotação do movimento realizado lateralmente pelo nariz.

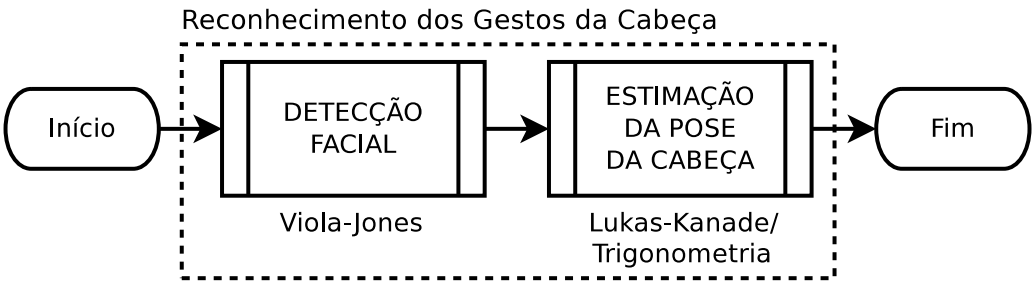
Figura 1 – Eixos tridimensionais de rotação (poses) da cabeça.



Fonte: Extraída de Arcoverde Neto et al. (2012).

É possível realizar o reconhecimento de gestos da cabeça através de uma estimação da pose da cabeça. Em Arcoverde Neto et al. (2014), por exemplo, a estimação da pose da cabeça foi desenvolvida para trabalhar em tempo real sobre plataformas móveis. O algoritmo é dividido em duas etapas, conforme mostrado no diagrama da Figura 2.

Figura 2 – Visão geral do sistema de reconhecimento de gestos da cabeça.



Fonte: Elaborada pelo autor com base em Arcoverde Neto et al. (2014).

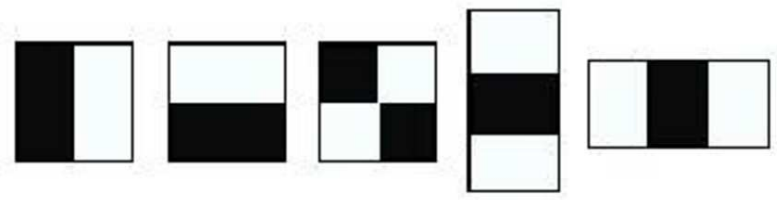
Inicialmente, a face é detectada através de um método proposto por Viola e Jones (VIOLA; JONES, 2001; VIOLA; JONES, 2004). A identificação dos olhos e do nariz é realizada como passo seguinte através de relações que consideram a antropometria da face humana. Os autores propuseram um método próprio que, de posse das coordenadas dos pontos de localização dos olhos e do nariz no plano bidimensional, calcula os eixos de rotação *roll*, *yaw* e *pitch* por relações trigonométricas com base no deslocamento dos três pontos em dois *frames* consecutivos de um vídeo. Os algoritmos são detalhados a seguir.

2.1.1 Viola-Jones e Relações Antropométricas

A proposta do algoritmo Viola-Jones é, de forma geral, baseada em três ideias: uso de características da imagem "parecidas com Haar" (Haar-*like features*, similares às *wavelets* utilizadas no primeiro detector facial em tempo real); uma representação de imagem chamada de "imagem integral"; e uma série de classificadores implementados em cascata.

Haar-*like features* são estruturas retangulares que indicam a presença ou ausência de bordas, barras ou outras características simples da imagem. Durante a fase de treino, alguns desses descritores são selecionados para representar o objeto (no caso, o rosto). A Figura 3

Figura 3 – Haar-*like features*

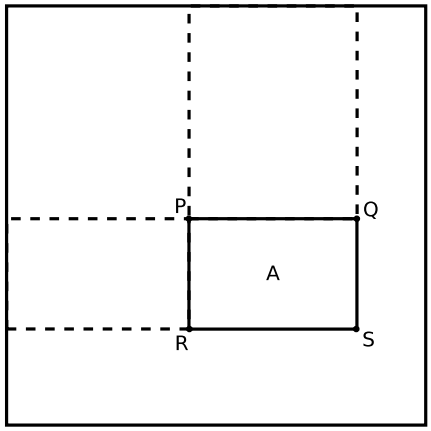


Fonte: Extraída de Martínez-Zarzuela et al. (2011).

mostra cinco exemplos de características Haar, as quais podem ser definidas como a diferença entre a soma dos pixels das áreas pretas e a soma dos pixels das áreas brancas. Além disso, tal representação pode ser otimizada pelo uso das chamadas imagens integrais.

Uma imagem integral é definida de modo que cada ponto (x', y') seja a soma dos pixels da área superior-esquerda ao ponto (x, y) da imagem original. Assim, para calcular a área de um retângulo A, como mostrado no exemplo da Figura 4, a expressão $S - Q - R + P$ poderia ser aplicada utilizando apenas os quatro pontos em questão.

Figura 4 – Imagem integral.



Fonte: Baseada em Viola e Jones (2001).

Através da ideia de foco de atenção, o algoritmo pode ser melhorado quando se tenta procurar apenas nos locais onde a face é mais provável de ser encontrada. Esse processo é feito por uma série de classificadores em cascata que aumentam em complexidade, de modo que uma região com baixa probabilidade de conter uma face é imediatamente descartada por um classificador mais simples, o que poupa recursos computacionais. De forma geral, o funcionamento do algoritmo Viola-Jones em si consiste em avaliar a saída dos classificadores em cascata em uma janela que desliza sobre a imagem.

De posse dos limites verticais e horizontais da face, que formam um quadrado, três pontos, definidos por $P_{OE} = (x_{OE}, y_{OE})$, $P_{OD} = (x_{OD}, y_{OD})$ e $P_N = (x_N, y_N)$, são calculados, com base no modelo geo e antropométrico da face humana, de forma heurística e relativamente precisa para representar a posição do olho esquerdo, do olho direito e do nariz, respectivamente, no plano bidimensional. Assim, a coordenada x_{OE} encontra-se a 30% da largura horizontal da face; x_{OD} a 70% da mesma largura; e ambas as coordenadas y_{OE} e y_{OD} encontram-se à 38% da altura vertical (de baixo para cima) da face. Já a coordenada abscissa do nariz é dada por uma relação entre a distância horizontal dos olhos $x_N = (x_{OE} + x_{OD})/2$, enquanto a segunda coordenada do nariz é calculada por $y_N = y_{OE} + (x_{OD} - x_{OE})^*0,45$

2.1.2 Lucas-Kanade e Relações Trigonométricas

O algoritmo de Lucas-Kanade (LUCAS; KANADE, 1981) é responsável por realizar o rastreamento de alguns pontos da face no plano bidimensional da imagem. De forma geral, o algoritmo utiliza apenas dois *frames* consecutivos do vídeo de entrada para estimar o movimento de um objeto (no caso, o rosto) através do cálculo do fluxo óptico entre os *frames*. O fluxo óptico nada mais é do que distribuição aparente de velocidade resultante do movimento de pontos de intensidade: dado um pixel em uma imagem, determinar um pixel próximo, dentro de uma região de interesse, em uma segunda imagem, com a mesma iluminação/brilho.

Assumindo pontos da face disponíveis em um *frame* (considera-se aqui o referido *frame* como sendo o *frame* "anterior"), o algoritmo de Lucas-Kanade calcula o deslocamento desses pontos dentro de uma área de vizinhança definida entre duas imagens: o *frame* anterior e o *frame* "atual". O objetivo desse cálculo é minimizar a soma do erro quadrático entre os dois *frames*, o que resulta na estimação da localização dos pontos em um *frame* com base apenas no seu *frame* predecessor.

Considerando os pontos $P'_{OE} = (x'_{OE}, y'_{OE})$, $P'_{OD} = (x'_{OD}, y'_{OD})$ e $P'_{N} = (x'_{N}, y'_{N})$ como sendo localização do olho esquerdo, olho direito e nariz, respetivamente, no frame anterior; e os pontos $P_{OE} = (x_{OE}, y_{OE})$, $P_{OD} = (x_{OD}, y_{OD})$ e $P_{N} = (x_{N}, y_{N})$ os mesmos pontos no frame atual, é possível calcular ângulos em relação aos três principais eixos de rotação da cabeça através de um conjunto de três relações trigonométricas:

$$roll = \arctan\left(\frac{y_{OE} - y_{OD}}{x_{OE} - x_{OD}}\right) \tag{1}$$

$$yaw = x'_{N} - x_{N} \tag{2}$$

$$pitch = y_{N} - y'_{N}\ , \tag{3}$$

onde a Equação 1 depende somente da localização do olho esquerdo (*OE*) e direito (*OD*) no *frame* atual; e as Equações 2 e 3, que dependem somente da localização do nariz (*N*), precisam das coordenadas dos pontos identificados tanto no *frame* anterior quanto no atual.

2.2 Reconhecimento Automático de Voz

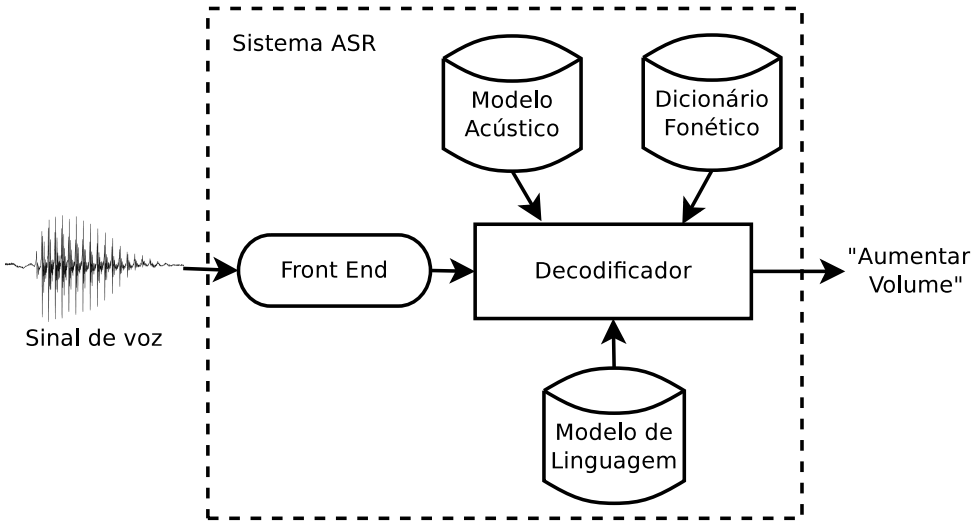
Os sistemas de reconhecimento automático de voz (ASR) têm sido estudados desde os anos 50 (RABINER; JUANG, 1993; HUANG; ACERO; HON, 2001; CUNHA; VELHO, 2003). Por exemplo, nos Laboratórios Bell nessa época, foi desenvolvido o primeiro reconhecedor de dígitos isolados com suporte a apenas um locutor. Na mesma década, foi introduzido o conceito de redes neurais, mas devido a muitos problemas práticos, a ideia não foi seguida no âmbito de ASR. Nos anos 70, a técnica predominante era a *dynamic time-warping* (DTW) (SAKOE; CHIBA, 1978), que consiste em um algoritmo que mede a similaridade en-

tre duas sequências após alinhá-las ao longo do tempo. Com o passar dos anos, a pesquisa foi evoluindo e muitas limitações técnicas foram sendo superadas, além da globalização e popularização dos computadores, o que levou a um aumento no número de pesquisas na área de processamento de voz.

Nos anos 80, os sistemas ASR tentavam aplicar, inicialmente, um conjunto de regras gramaticais e sintáticas à fala (JELINEK, 1997). Caso as palavras ditas caíssem dentro de um certo conjunto de regras, o programa poderia determinar quais eram aquelas palavras. Para isso, era preciso falar cada palavra separadamente (sistemas de palavras isoladas), com uma pequena pausa entre elas. Porém, características como sotaques, dialetos e as inúmeras exceções inerentes à língua dificultaram a disseminação dos sistemas baseados em regras. Foi então que vários pesquisadores iniciaram estudos para reconhecimento de palavras conectadas utilizando métodos estatísticos, com maior destaque para os modelos ocultos de Markov (HMMs, do inglês *hidden Markov models*) (RABINER, 1989; JUANG; RABINER, 1991). Atualmente, o estado da arte em reconhecimento de voz é representado pelo uso de HMMs, incrementado por técnicas como aprendizado discriminativo (WOODLAND; POVEY, 2002), seleção de misturas de gaussianas (LEE; KAWAHARA; SHIKANO, 2001), entre outras.

Um sistema ASR é, tipicamente, composto por cinco blocos: *front-end*, modelo acústico, modelo de linguagem, dicionário fonético e decodificador, como indicado na Figura 5.

Figura 5 – Arquitetura de um sistema ASR.



Fonte: Adaptada de Sampaio Neto et al. (2011).

No processo de *front-end* convencional, tem-se a segmentação do sinal de voz em "janelas" curtas (*frames*) de 20 a 25 milissegundos, com o deslocamento da janela de análise sendo tipicamente de 10 milissegundos. Cada *frame* é, então, convertido em um vetor $\mathbf{x}$ de dimensão L (tipicamente, $L = 39$). É assumido aqui que T *frames* estão organizados em uma matriz $\mathbf{X}$ de $L \times T$, representando uma sentença completa.

Existem várias alternativas no que diz respeito à parametrização do sinal de voz (PICONE, 1993; JUNQUA; HATON, 1996; HUANG; ACERO; HON, 2001). Entretanto, a análise dos coeficientes cepstrais da escala Mel (MFCCs, do inglês *Mel-frequency cepstral coefficients*) (DAVIS; MERLMESTEIN, 1980) tem provado ser eficiente e é geralmente empregada como entrada para os blocos de *back-end* que compõem um sistema ASR (HUANG; ACERO; HON, 2001).

O modelo de linguagem provê a probabilidade $p(\mathscr{T})$ de observar a sentença $\mathscr{T} = [w_1, \ldots, w_P]$ com P palavras. Conceitualmente, o objetivo é encontrar a sentença $\mathscr{T}^*$ que maximiza a probabilidade posterior

$$\mathscr{T}^* = \arg\max_{\mathscr{T}} p(\mathscr{T}|\mathbf{X}) = \arg\max_{\mathscr{T}} \frac{p(\mathbf{X}|\mathscr{T})p(\mathscr{T})}{p(\mathbf{X})}, \tag{4}$$

onde $p(\mathbf{X}|\mathscr{T})$ é dada pelo modelo acústico. Como $p(\mathbf{X})$ não depende de $\mathscr{T}$:

$$\mathscr{T}^* = \arg\max_{\mathscr{T}} p(\mathbf{X}|\mathscr{T})p(\mathscr{T}). \tag{5}$$

Na prática, uma constante empírica é usada para ponderar a probabilidade do modelo de linguagem $p(\mathscr{T})$, antes da mesma ser combinada com a probabilidade do modelo acústico $p(\mathbf{X}|\mathscr{T})$.

Dado o grande volume de sentenças concorrentes, a Equação 5 não pode ser calculada independentemente para cada sentença candidata. Para esse fim, os sistemas ASR utilizam-se de estruturas de dados hierárquicas, como árvores léxicas, e do artifício de separar as sentenças em palavras, e as palavras em unidades básicas, chamadas de fones (HUANG; ACERO; HON, 2001). A busca por $\mathscr{T}^*$ é chamada decodificação e, na maioria dos casos, hipóteses são descartadas ou podadas (*pruning*). Em outras palavras, para tornar viável a busca pela "melhor" sentença, algumas sentenças candidatas são descartadas e a Equação 5 não é calculada para elas (DESHMUKH; GANAPATHIRAJU; PICONE, 1999; JEVTIĆ; KLAUTAU; ORLITSKY, 2001).

Um dicionário fonético (também conhecido por modelo léxico) provê o mapeamento das palavras em unidades básicas (fones) e vice-versa.

Em termos de modelo acústico, no intuito de incrementar o desempenho, HMMs contínuas são geralmente empregadas, onde a distribuição de saída de cada estado é modelada por uma mistura de Gaussianas. A topologia "*left-right*" pode ser adotada, onde as únicas transições permitidas são de um estado para ele mesmo ou para o estado seguinte.

O modelo acústico pode conter uma HMM por fone. Isso seria bastante razoável caso um fone pudesse ser seguido por qualquer outro, o que não é verdade, já que os articuladores do trato vocal não se movem de uma posição para outra imediatamente na maioria das transições de fones. Nesse sentido, durante o processo de criação de sistemas que modelam a fala fluente, busca-se um meio de modelar os efeitos contextuais causados pelas diferentes maneiras que alguns fones podem ser pronunciados em sequência. A solução encontrada é o uso de

HMMs dependentes de contexto, que modelam o fato de um fone sofrer influência dos fones vizinhos, conhecido como coarticulação (LADEFOGED, 2001). Por exemplo, supondo a notação do trifone "*a*-*b*+*c*", "*b*" representa o fone central ocorrendo após o fone "*a*" e antes do fone *c*.

A modelagem da língua estima

$$P(\mathscr{T}) = P(w_1, w_2, \ldots, w_P) \quad (6)$$

$$= P(w_1)P(w_2|w_1)\ldots P(w_P|w_1, w_2, \ldots, w_{P-1}). \quad (7)$$

É impraticável estimar a probabilidade condicional $P(w_i|w_1, \ldots, w_{i-1})$, mesmo para valores moderados de i. Assim, o modelo de linguagem para sistemas ASR consiste em um modelo n-gram, que assume que a probabilidade $P(w_i|w_1, \ldots, w_{i-1})$ depende somente das $n-1$ palavras anteriores. Por exemplo, para $n = 3$, a probabilidade $P(w_i|w_{i-2}, w_{i-1})$ expressa um modelo de linguagem trigrama.

Basicamente, existem dois tipos de aplicação para sistemas de reconhecimento de fala: ditado, e comando e controle (HUANG; ACERO; HON, 2001). O primeiro, por ser capaz de suportar um vocabulário amplo, utiliza o modelo de linguagem n-gram e requer um sistema mais robusto, exigindo mais recursos e processamento da máquina. O segundo é relativamente mais simples, pois utiliza uma gramática livre de contexto como modelo, restringindo, assim, a sequência de palavras aceitas.

Resumindo, após o treinamento de todos os modelos estatísticos, um sistema ASR na etapa de teste usa o *front-end* para converter o sinal de entrada em parâmetros e o decodificador para encontrar a melhor sentença $\mathscr{T}$.

2.2.1 Redes Neurais Profundas

Apesar de o estado da arte em reconhecimento de fala ainda ser dado pela combinação de mistura de gaussianas com HMMs, mesmo após 20 anos de pesquisa, bons resultados vêm sendo recentemente obtidos com as redes neurais profundas (DNNs, do inglês *deep neural networks*) para outros idiomas, como o italiano (COSI, 2015), árabe (ALI et al., 2014), espanhol latino (BECERRA; ROSA; GONZÁLEZ, 2016), inglês e alemão (GAIDA et al., 2014). A principal referência de pacote de *software* na área de reconhecimento de fala com DNNs é o Kaldi (POVEY et al., 2011) e acredita-se que tal técnica possa ser a nova tendência para os sistemas ASR durante os próximos anos.

2.3 Acionador Externo

Um acionador externo sem fio foi construído com o intuito de evitar que o sistema permanecesse sempre ativo, o que acarretaria diversos erros de reconhecimento por conta de movimentos involuntários do usuário. Então, antes de realizar qualquer movimento, o usuário

deve necessariamente inicializar o sistema com o acionador. Por esse motivo, espera-se que a movimentação dos membros superiores não seja totalmente comprometida, como no caso dos tetraplégicos, ou mesmo proibitiva, considerando os amputados. O sistema para automaticamente quando um gesto da cabeça é reconhecido ou quando um movimento é classificado como fora do padrão, resultando em uma situação de erro. Se o usuário desejar ligar a TV e aumentar o volume, por exemplo, o acionador deve ser utilizado duas vezes, uma para cada ação de controle.

2.4 Protocolos de Comunicação via Luz Infravermelha

O funcionamento de controles remotos, com ênfase nos baseados em luz infravermelha (IR, do inglês *infrared*) para TVs, é explicado de forma clara e detalhada em alguns tutoriais para "curiosos" disponíveis na Internet (MUNDO ESTRANHO, 2011; LAYTON, 2005). Atualmente, a maioria dos aparelhos eletrônicos recebe informação através de sensores infravermelhos localizados em painéis frontais. Entretanto, devido à interferência que surge com a vasta quantidade de raios IR no ambiente (luz solar, lâmpadas fluorescentes, etc), a comunicação entre o controle remoto e a televisão ocorre geralmente em 4 passos: um comando *start* inicia a transferência; seguido dos bits do comando específico (como "aumentar o volume", por exemplo) e do endereço do aparelho; por fim, um comando de *stop* encerra o envio de bits. Dessa forma, a chance de a informação ser reconhecida por mais de um aparelho é baixa (exceto se forem equipamentos do mesmo modelo e do mesmo fabricante, como duas TVs da Samsung).

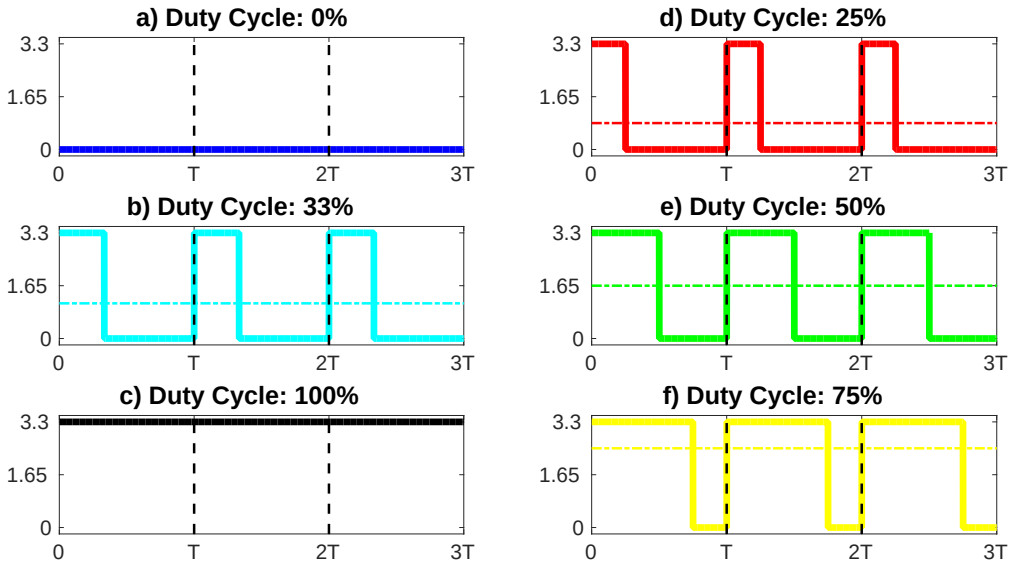
Entretanto, uma das grandes dificuldades relacionadas ao controle de equipamentos eletrônicos é que os 4 passos descritos acima são apenas uma forma genérica de descrever a comunicação. Ou seja, cada empresa praticamente segue o seu próprio padrão para estabelecer a comunicação entre o controle e os dispositivos: o número, a ordem, a duração e o significado dos bits são variados; a modulação e codificação usadas são diferentes; e a frequência dos pulsos pode oscilar entre 32 kHz e 42 kHz, chegando a 50 kHz em determinados aparelhos mais modernos. Além disso, tão rara quanto o seguimento de uma comunicação via infravermelho padronizada é a disponibilização de documentação pelos fabricantes.

2.4.1 Modulação por Largura de Pulso

A modulação por largura de pulso (PWM, do inglês *pulse-width modulation*), é uma técnica para conseguir resultados analógicos por meios digitais (ARDUINO TUTORIALS, 2017). O controle digital é usado para criar um sinal que oscila entre "ligado" e "desligado", produzindo uma onda quadrada cuja tensão média é, no caso do C.H.I.P., um valor entre 0 V e 3,3 V. A maioria dos protocolos de comunicação entre controles remotos e dispositivos eletrônicos utiliza PWM para reduzir a energia ao piscar o LED infravermelho (LED IR) e minimizar interferências e ruídos.

A onda gerada pela PWM possui alguns parâmetros, dos quais destacam-se o período e o *duty cycle*. O primeiro é bem simples, dado pela distância entre um pulso e o próximo (que seriam, em uma onda analógica, duas cristas consecutivas); já o segundo é dado pela percentagem em que esse mesmo período permanece em nível alto (ligado), representando o período apenas do pulso. Existem outros parâmetros como largura do pulso e frequência de PWM, mas estes podem ser obtidos a partir das duas informações mencionadas acima.

Figura 6 – Exemplos de PWMs de mesmo período com diferentes *duty cycles*.



Fonte: Elaborada pelo autor.

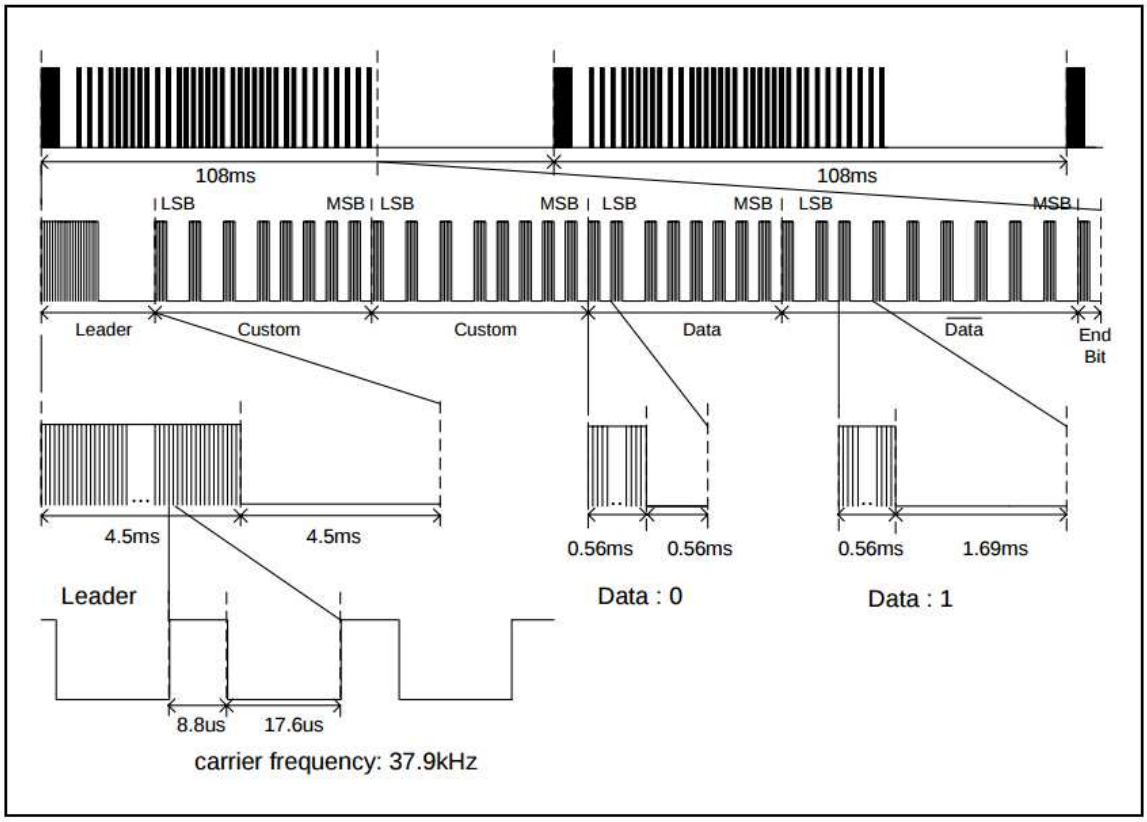
A Figura 6 mostra seis sinais de mesma duração ($3T$) e mesmo período (T), porém com *duty cycle* variado. Há dois casos especiais, mostrados em 6a) e 6c), nos quais o *duty cycle* é, respectivamente, nulo (0%) e igual ao período da PWM (100%). Nesses casos, a tensão média (representada pela linha horizontal nas demais ondas) será mínima (0 V) e máxima (3,3 V).

A maioria dos sistemas de controle remoto baseados em luz infravermelha utiliza a técnica de PWM para codificar o pulso alto de suas respectivas formas de onda. Como o sistema apresentado no trabalho utiliza uma televisão da Samsung, somente o protocolo utilizado por essa fabricante será detalhado a seguir.

2.4.2 O Protocolo da Samsung

A Samsung, assim como a maioria dos demais fabricantes, utiliza seu próprio padrão para comunicação entre os controles e os dispositivos eletrônicos. Felizmente, o documento de *application notes* do microcontrolador S3F80KB (SAMSUNG ELECTRONICS, 2008) presente em controles da Samsung encontra-se disponível na Internet.

Figura 7 – Sinal infravermelho do protocolo da Samsung.



Fonte: Extraída de Samsung Electronics (2008).

O protocolo define uma sequência de 34 bits, onde ambos os valores ‘0’ e ‘1’ são representados por uma mudança no estado e diferenciados pela duração do nível baixo dos pulsos, conforme ilustrado no esquemático da Figura 7 (*Data* 0 e *Data* 1, segundo a legenda). Os bits são definidos sempre como uma transição *high-to-low* durante um período, na qual é estabelecida uma única duração de 560 μs para o nível alto do pulso (modo *burst*) e duas para o nível baixo: 1690 μs para o bit ‘1’ e 560 μs para o bit ‘0’. Ou seja, ambos os bits começam seus respectivos períodos com um nível alto, mas o bit ‘1’ diferencia-se por possuir duração mais longa no nível baixo.

O nível alto dos bits normais é caracterizado por uma PWM com frequência de 37,9 kHz e *duty cycle* que pode variar entre 33% e 50% do período. A PWM caracteriza o modo *burst*, em que o LED IR pisca na frequência definida pelo protocolo. Já o nível baixo é caracterizado pela ausência de sinal, ou seja, é o tempo em que o LED IR permanece desligado ou em estado *idle*. O primeiro bit, chamado de *leader*, tem duração mais longa para garantir a sincronia e o ganho no circuito receptor: 4,5 ms no nível alto e 4,5 ms no nível baixo; os 16 bits seguintes são divididos em dois blocos personalizados (definidos como *Custom*) exatamente iguais de 8 bits cada, provavelmente definindo o endereço do aparelho; outras duas seções de 8 bits definem o bloco de dados, no qual o segundo bloco ($\overline{Data}$) é o complemento dos bits do primeiro (*Data*),

que define os comandos propriamente ditos; o 34° bit (*End Bit*) é sempre baixo e encerra a sequência.

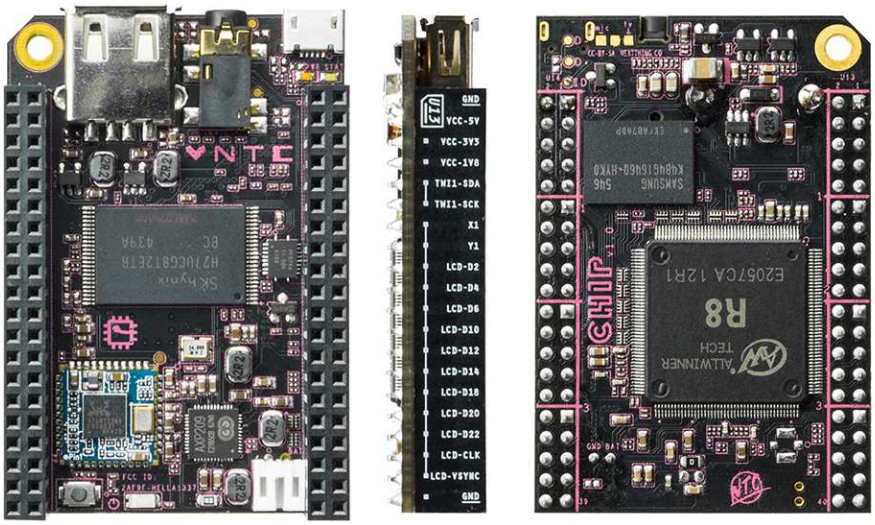
2.5 Ferramentas de *Hardware* e Plataformas Embarcadas

Esta seção apresenta as duas principais ferramentas de *hardware* utilizadas, descrevendo também alguns dispositivos de comunicação e de entrada e saída utilizados para construir o sistema proposto.

2.5.1 C.H.I.P.

C.H.I.P. (C.H.I.P., 2017) é um microcomputador de baixo custo (apenas U$ 9,00 dólares) lançado em 2015 pela fabricante Next Thing Co. A plataforma embarcada, exibida na Figura 8, é baseada na versão Debian do Linux, o que torna fácil a tarefa de compilar e executar *softwares* que já estão disponíveis para a versão *desktop* do sistema operacional. O C.H.I.P. é também *open-hardware*, o que significa que os esquemáticos estão disponíveis livremente na Internet[1] sob a licença *Creative Commons Attribution-ShareAlike* (CC BY-SA)[2] para o caso de entusiastas desejarem construir suas próprias versões da placa.

Figura 8 – Microcomputador C.H.I.P.



Fonte: Extraída de C.H.I.P. (2017).

O microcomputador conta com um processador AllWinner R8 (ARM Cortex-A8) de 1 GHz, memória DRAM de 512 MB, 4 GB de armazenamento em *flash* NAND *onboard*, além de módulos Wi-Fi e Bluetooth já integrados. A placa também possui 80 pinos divididos em dois *headers* (U13 e U14), uma porta USB *host* tipo A para entrada e saída de dados e uma porta USB micro tipo B para alimentação, que é uma alternativa ao conector para uma bateria LiPo.

[1] <https://github.com/NextThingCo/>
[2] <https://creativecommons.org/licenses/by-sa/3.0/us/>

2.5.1.1 Expansão da Pinagem dos Headers U13 e U14

O C.H.I.P. possui 51[3] pinos que podem ser definidos como entrada ou saída de propósito geral (GPIO, do inglês *general purpose input/output*) divididos em 4 bancos. O modo GPIO descreve o comportamento ou direção dos pinos (entrada ou saída) — similar ao que ocorre com o uso da função `pinMode()` do Arduino —, os quais podem ser controlados dinamicamente pelo usuário mediante o ajuste de parâmetros de configuração do *kernel* do Linux.

Cada pino digital do C.H.I.P. possui até 7 modos diferentes de funcionamento (Multi0 até Multi6), os quais incluem: comunicação UART ou SPI, interrupções, PWM, GPIO, etc. O processo de escolha da função do pino, conhecido por multiplexação, exige que o programador saiba exatamente quais as funções que cada pino pode assumir e se o mesmo está disponível, já que alguns pinos do C.H.I.P. não estão conectados ao processador R8. A multiplexação de pinos ocorre pelo ajuste da árvore de dispositivos do Linux (*device tree*), permitindo que o sistema operacional identifique, no momento do *boot*, todos os periféricos de *hardware* disponíveis.

Nas versões mais recentes do *kernel* do Linux embarcado, a recompilação da árvore e reinicialização do sistema operacional em caso de alteração não é mais necessária: o módulo é acrescentado ou "sobreposto" à árvore original e carregado dinamicamente através de um mecanismo chamado *device tree overlay* (PETAZZONI, 2014). De forma alternativa, um dos modos mais simples de controle e acesso aos módulos GPIO ocorre pela simples atribuição de valores a arquivos específicos pelo sistema de arquivos, desde que o usuário Linux tenha as permissões de edição ou seja o *root*. Tal operação oferece uma interface mais amigável para o programador, já que todo o processo de baixo nível de modificação no *kernel* é mascarado.

Os arquivos de configuração do GPIO estão localizados em `/sys/class/gpio`. Ao se escrever o número do GPIO (`XX`) no arquivo `export`, o qual configura automaticamente o pino

[3] Embora o processador R8 permita o uso de 76 pinos como entrada e 70 como saída.

```
1  #!/bin/bash
2  echo 195  > /sys/class/gpio/export
3  echo out > /sys/class/gpio/gpio195/direction
4  for i in `seq 10`; do
5      echo 1 > /sys/class/gpio/gpio195/value
6      sleep 0.5
7      echo 0 > /sys/class/gpio/gpio195/value
8      sleep 0.5
9  done
10 echo 195 > /sys/class/gpio/unexport
```

Listagem 1 – Acesso ao GPIO 195 / U14_3 / PG3 através de arquivos.

para o modo Multi0 (entrada) ou Multi1 (saída), o diretório `gpioXX`, que contém os arquivos necessários para o acesso e controle do pino, é criado. Apenas os arquivos `direction` e `value` são necessários para a tarefa de piscar um LED, os quais indicam, respectivamente, se o pino será de entrada ("*in*") ou saída ("*out*") e o valor digital ("0" ou "1") que será escrito ou lido.

A Listagem 1 mostra um código capaz de piscar um LED conectado ao pino 3 do *header* U14 (UART1-TX, PG3 no processador) com frequência de 1 Hz. O número de GPIO é calculado multiplicando-se o número do banco (de A=0 a G=6) por 32 e somando ao número do módulo (de 0 a 27). No modo Multi1, o pino PG3 do processador será multiplexado para o módulo 3 do banco G=6 dos GPIOs, conforme visto na Tabela 2. Portanto, seu número de GPIO é `XX(PG3)` $= G \times 32 + 3 = 6 \times 32 + 3 = 195$.

Tabela 2 – Modos dos pinos 3, 25 e 35 do *header* U14 do C.H.I.P..

Pino	3	25	35
Pino SoC	PG3	PB18	PE8
Nome	UART1-TX	TWI2-SDA	CSID4
Multi0	Entrada	Entrada	Entrada
Multi1	Saída	Saída	Saída
Multi2	-	TWI2_SDA	TS_D4
Multi3	-	-	CSI_D4
Multi4	UART1_TX	-	SDC2_CMD
Multi5	-	-	-
Multi6	EINT3	-	-
GPIO	**195**	**50**	**136**

Fonte: Elaborada pelo autor.

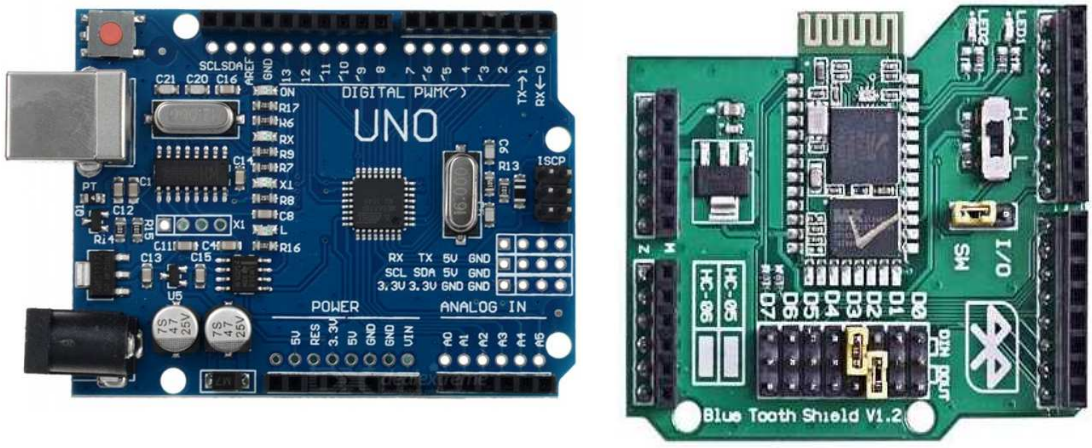
Apesar da facilidade em controlar o nível lógico dos pinos pela manipulação de arquivos, tal forma de atribuição de valores não atende à frequência exigida pela PWM do protocolo da Samsung. A alternância entre os níveis do GPIO dura, na maioria das vezes, 22 μs, conforme constatado através de repetidas escritas no arquivo `value`. Como o período de um ciclo da PWM dura exatos 26,385 μs ($1/37900\ Hz$) e o *duty cycle* dura 33% desse período (8,8 μs), nota-se que é o processo de emular uma PWM por GPIOs é lento demais e, portanto, inviável.

2.5.2 Arduino

O Arduino (ARDUINO, 2017) é uma plataforma embarcada baseada no microcontrolador ATmega328. O modelo UNO R3, utilizado por conta da ausência de controladores dos pinos de PWM do C.H.I.P., possui 16 kHz de frequência de *clock*, 2 kB de memória SRAM e 32 kB de armazenamento em *flash*. A plataforma ainda conta com uma porta USB para comunicação serial e alimentação, além de 20 pinos de GPIO.

Figura 9 – Arduino UNO e *shield* Bluetooth SHD 18, baseado no módulo HC-05.

(a) Arduino (b) *Shield*



Fonte: Extraída da Internet.

A plataforma embarcada também é *open-hardware*, assim como o C.H.I.P., permitindo que qualquer pessoa com conhecimento técnico sinta-se motivada a estudar e fazer modificações nos esquemáticos[4] de *hardware*, disponibilizados novamente sob a licença CC BY-SA. Neste trabalho, o Arduino foi equipado com um *shield* Bluetooth SHD 18, que é baseado no módulo serial HC-05. As duas placas são mostradas lado a lado na Figura 9.

2.6 Ferramentas de *Software*

O Debian Jessie foi instalado e configurado no C.H.I.P. como sistema operacional base. Assim como a maioria das versões baseadas no *kernel* do Linux, o Jessie também é *open-source* e disponibilizado sob a licença GNU *General Public License* (GPL)[5].

As etapas do módulo de reconhecimento de gestos foram implementadas com auxílio das APIs para C++ da biblioteca `OpenCV` (OPENCV, 2017), que também é *open-source* e disponibilizada de forma gratuita sob a licença CC BY-SA. Além disso, a biblioteca também é vantajosa por possuir boa documentação e uma vasta comunidade de usuários.

O sistema de reconhecimento de voz foi implementado graças a API para a linguagem C/C++ do *engine* PocketSphinx (HUGGINS-DAINES et al., 2006), desenvolvido pelo grupo CMU Sphinx (CMUSPHINX, 2017b). A plataforma *open-source*, disponibilizada sob a licença BSD 2-Clause[6], é capaz de processar fala em um vocabulário de até 65 mil palavras. Os recursos para realizar ASR em Português Brasileiro, como dicionário fonético e áudios para treino do modelo acústico, são disponibilizados pelo grupo FalaBrasil (FALABRASIL, 2013).

4 <https://github.com/arduino/>
5 <https://www.gnu.org/licenses/gpl-3.0.en.html>
6 <https://opensource.org/licenses/BSD-2-Clause>

O C.H.I.P. envia os dados referentes ao reconhecimento dos gestos para o Arduino via Bluetooth. Isso é feito através da API para a linguagem C/C++[7] da biblioteca `BlueZ`[8], a qual faz parte da lista de pacotes oficial do Linux. Já biblioteca `SoftwareSerial`[9], que acompanha o ambiente de desenvolvimento do Arduino, foi utilizada para receber dados advindos do C.H.I.P. através do módulo HC-05 presente no *shield* Bluetooth acoplado ao Arduino. A licença das bibliotecas `BlueZ` e `SoftwareSerial` é GPL e CC BY-SA, respectivamente.

Por fim, a biblioteca `IRremote`[10], também *open-source* com licença GPL, foi utilizada para emular o protocolo de comunicação entre o Arduino e o aparelho televisor a ser controlado via luz infravermelha. Assim, como vários dispositivos eletrônicos decodificam informações de diferentes protocolos, a necessidade de implementar vários padrões de comunicação com futuros aparelhos será evitada quando o sistema de controle tornar-se, de fato, universal.

2.7 Formas de Avaliação

O sistema de controle remoto foi avaliado de duas formas. A primeira avaliação foi objetiva e específica em relação à taxa de acerto dos sistemas de reconhecimento de gestos da cabeça e de voz, métodos-chave de entrada do sistema. Já a segunda avaliação, subjetiva, levou em consideração a opinião dos voluntários sobre o desempenho do protótipo em si, de forma geral, como um todo.

Os sistemas AGR e ASR foram avaliados por um método de contagem de tentativas de execução de cada ação de controle na TV. Os usuários foram convidados a realizar cada um dos seis movimentos com a cabeça; e a falar cada um dos comandos de voz por, no máximo, quatro tentativas. Se o sistema não conseguisse identificar de forma correta o movimento realizado ou o comando falado, ou seja, se o respectivo comando de controle não fosse devidamente enviado para o aparelho televisor de acordo com a entrada provida, o usuário era convidado a desistir da interação atual e partir para a próxima. Em outras palavras, a quinta tentativa de movimento ou a quinta repetição do mesmo comando falado representava uma falha definitiva. Nesse teste objetivo, obviamente, quanto menos tentativas de emitir cada um dos comandos de controle para o televisor, melhor.

Já a qualidade do sistema proposto, de forma geral como um conjunto, foi avaliada através de um questionário de duas perguntas ao final do teste. A primeira, de múltipla escolha, foi baseada em uma medida de qualidade de experiência conhecida como MOS (do inglês *mean opinion score*). O MOS surgiu como uma forma de avaliar, de forma quantitativa, a qualidade de sistemas de telecomunicações e dos áudios transmitidos, onde o usuário expressava sua opinião ao escolher um dos valores apresentados em uma escala predefinida (ITU-T, 2006).

[7] <https://people.csail.mit.edu/albert/bluez-intro>
[8] <http://www.bluez.org>
[9] <https://www.arduino.cc/en/Reference/SoftwareSerial>
[10] <https://github.com/z3t0/Arduino-IRremote>

A escala MOS aplicada neste trabalho foi adaptada de Ou et al. (2014) e contém cinco alternativas que classificam o sistema em uma escala que varia de "péssimo" (1) a "excelente" (5). O escore geral é calculado pela média aritmética das respostas de todos os voluntários. A escala de pontuação é mostrada, da mesma forma que foi exibida aos voluntários, na Tabela 3.

Tabela 3 – Escala de avaliação baseada no MOS.

Escore	**Escala**	**Descrição**
5	Excelente	O sistema trará/traria muito conforto e bem-estar às PCD.
4	Bom	O sistema pode/poderia ser útil para as PCD.
3	Razoável	O sistema é/seria pouco útil para PCD.
2	Ruim	O sistema não é/seria útil para PCD.
1	Péssimo	O sistema causa/causaria desconforto às PCD.

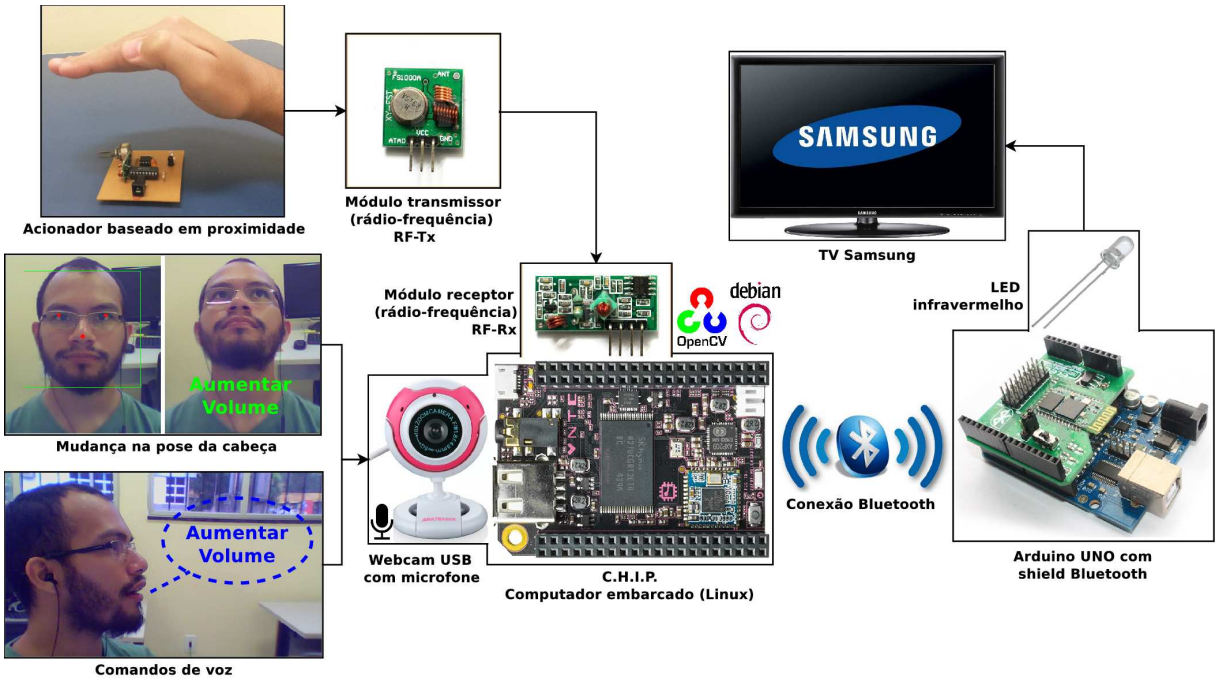
Fonte: Adaptada de Ou et al. (2014).

A segunda pergunta do questionário é subjetiva e questiona sobre a opinião do usuário em forma de palavras, o qual foi encorajado a dar sugestões, criticar e dar sua impressão pessoal sobre o sistema. A preferência entre os métodos de interação, voz ou gestos, também foi questionada. Caso o participante não fosse capaz de responder de forma independente, o responsável foi solicitado a descrever a impressão pessoal a respeito de como o participante reagiu ao utilizar o sistema.

3 PROTÓTIPO

O sistema foi construído sob a ideia do controle remoto universal e acessível às pessoas com deficiência. As plataformas embarcadas C.H.I.P. e Arduino foram configuradas para agir em conjunto como um *gateway* (ou *hub*) centralizado no ambiente residencial, conforme mostrado no esquemático ilustrado na Figura 10.

Figura 10 – Arquitetura do sistema de controle remoto proposto.



Fonte: Elaborada pelo autor.

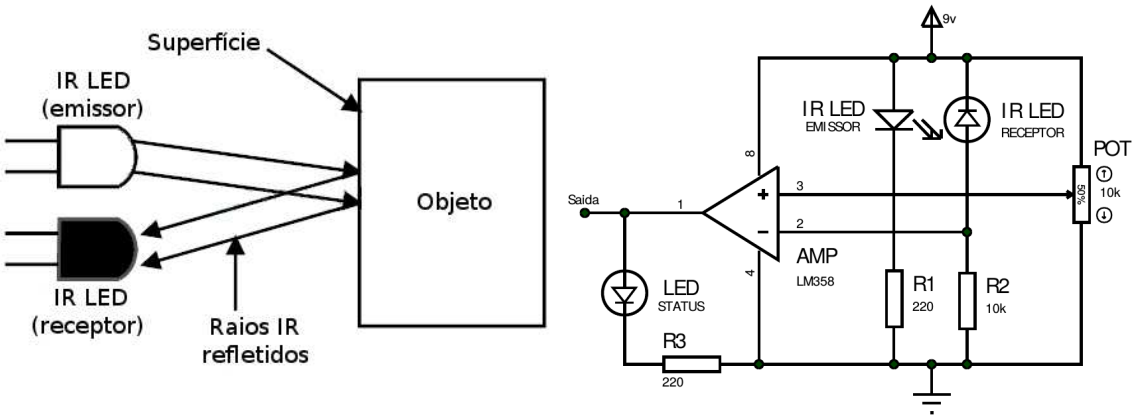
Um acionador externo sem fio foi utilizado para que o usuário pudesse inicializar o sistema AGR. Já o ASR foi inicializado pela própria voz do usuário. Uma *webcam* USB digital, conectada ao C.H.I.P., é responsável por capturar, em tempo real, imagens do rosto e cabeça do usuário, as quais são repassadas aos classificadores que tentam traduzir determinados movimentos em comandos para o aparelho. A mesma *webcam* contém um microfone, utilizado para digitalizar os comandos de voz dados pelo usuário. Obedecendo à exigência imposta pela comunicação infravermelha entre o controle remoto e o dispositivo eletrônico, o *gateway* foi posicionado na "linha de visão" do aparelho a ser controlado. Quando a pose ou o comando de voz são classificados, um sinal é enviado via Bluetooth para um Arduino UNO para que este possa, finalmente, emular o comando de controle correto e enviá-lo para o aparelho eletrônico através de um LED infravermelho (LED IR) acoplado a um circuito amplificador.

3.1 Acionador Externo

Um acionador externo sem fio foi construído com o intuito de evitar que o sistema permanecesse sempre ativo, o que acarretaria diversos erros de reconhecimento por conta de movimentos involuntários do usuário, além de consumo desnecessário de energia. Um sensor de proximidade baseado em luz infravermelha foi configurado para tal, utilizando dois LEDs IR (um emissor e um receptor) posicionados lado a lado, ambos apontando para a mesma direção. Quando um objeto é posto na linha de visão dos LEDs, os raios de luz gerados pelo emissor são refletidos diretamente no receptor, conforme ilustrado na Figura 11a.

Figura 11 – Sensor de proximidade baseado em luz infravermelha.

(a) Princípio de funcionamento (b) Esquemático do circuito



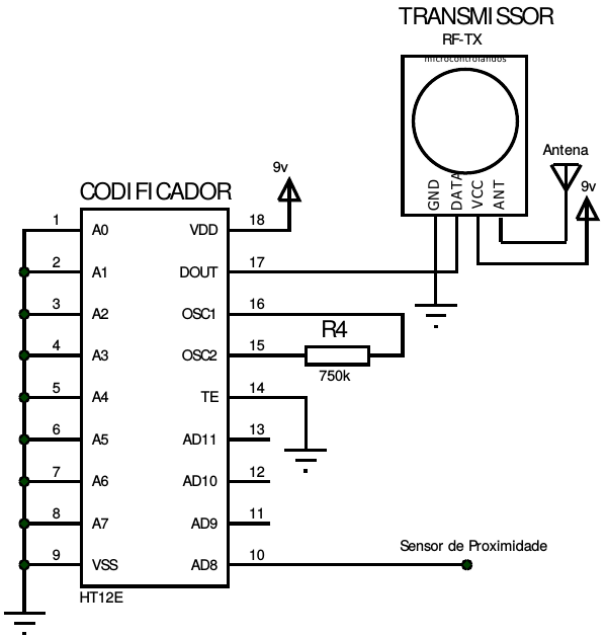
Fonte: (a): Adaptada da Internet; (b): Elaborada pela autor.

A incidência de raios infravermelhos, refletidos pela superfície sólida de um objeto, provoca uma reação de mudança de tensão no ânodo do sensor. Essa mudança de tensão, por ser muito baixa, é amplificada pelo amplificador operacional LM358 ao ponto de ser capaz de acender um LED posto em sua saída, o qual age como *status* da detecção de aproximação de algum objeto. A Figura 11b mostra o esquemático do circuito do sensor de proximidade que, portanto, torna o acionador livre de qualquer contato físico, visto que a aproximação de uma mão, um braço ou qualquer objeto é suficiente.

A saída do sensor de proximidade é enviada a um dos quatro canais (no caso, pino AD8) do circuito codificador HT12E (*encoder*), o qual modula uma informação de 12 bits paralelos (8 bits de endereço, do pino A0 ao pino A7; e 4 canais de dados, do pino AD8 ao pino AD11) em uma forma de onda específica para ser enviada de forma serial pelo módulo transmissor de rádio-frequência de 433 MHz (RF-Tx), conforme mostrado no esquemático da Figura 12.

O circuito receptor do acionador externo, por sua vez, possui basicamente um módulo receptor de rádio frequência de 433 MHz (RF-Rx) e um circuito integrado HT12D (*decoder*), como mostrado na Figura 13. O módulo RF-Rx recebe 12 bits de informação serial, oriundos

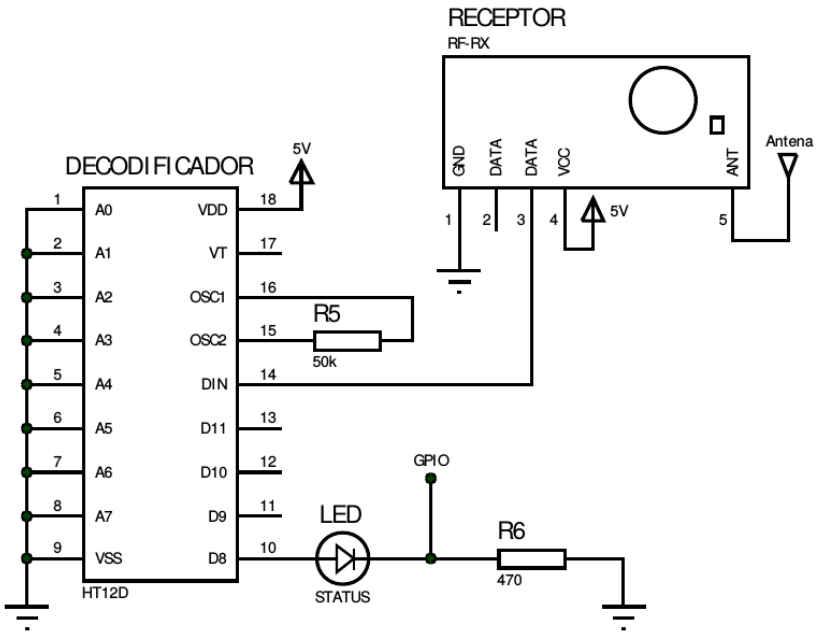
Figura 12 – Diagrama do módulo transmissor (Tx) do acionador externo sem fio.



Fonte: Elaborada pelo autor.

do módulo transmissor, e repassa-os para o decodificador, que paraleliza as informações, comparando os bits de endereço e ativando as respectivas portas de dados ou de saída (no caso, utilizou-se somente a porta D8). Com isso, os dados já estão prontos para serem lidos pelo C.H.I.P., bastando conectar o pino de saída ativado do HT12D a um pino de entrada do micro-

Figura 13 – Diagrama do módulo receptor (Rx) do acionador externo sem fio.



Fonte: Elaborada pelo autor.

computador. No entanto, como os pinos do C.H.I.P. suportam até 3,3 V de tensão de entrada, um pequeno circuito divisor de tensão, composto por um LED (que também funciona como *status* de recebimento de dados) e um resistor de 470 Ω, foi elaborado para diminuir a tensão de 5 V do pino de dados do decodificador, evitando, assim, que o pino de entrada do computador embarcado fosse danificado.

Portanto, antes de realizar qualquer movimento, o usuário deve necessariamente inicializar o sistema com o acionador[1]. Por esse motivo, espera-se que a movimentação dos membros superiores não seja totalmente comprometida, como no caso dos tetraplégicos, ou mesmo proibitiva, considerando os amputados. O sistema para automaticamente quando um movimento da cabeça é reconhecido ou quando um movimento é classificado como fora do padrão, resultando em uma situação de erro. Se o usuário desejar ligar a TV e aumentar o volume, por exemplo, o acionador deve ser utilizado duas vezes, uma para cada ação de controle.

3.2 Reconhecimento de Gestos da Cabeça

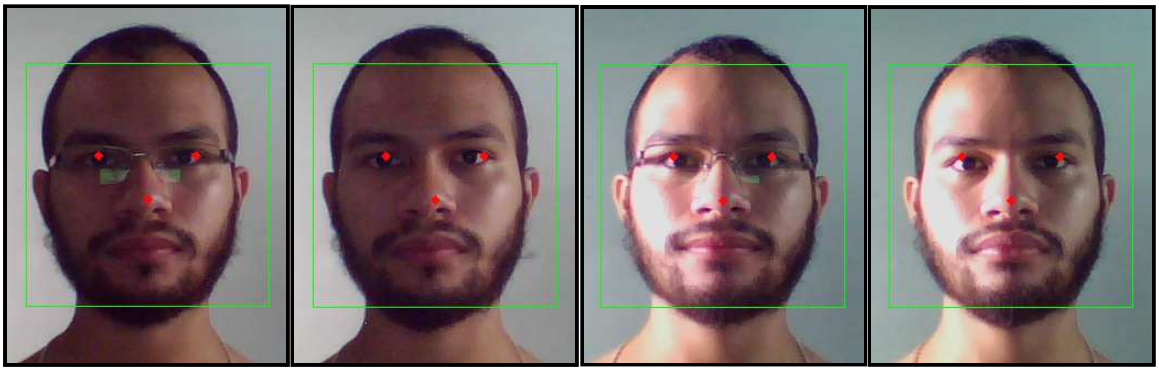
A etapa seguinte à captura das imagens pela *webcam* digital é a de reconhecimento de gestos da cabeça. Como o C.H.I.P. ainda é bastante limitado (apesar de bastante poderoso quando comparado ao seu tamanho), a busca por algoritmos mais simples e com baixa complexidade foi fundamental. Assim, o presente sistema foi desenvolvido com base na solução proposta em Arcoverde Neto et al. (2014), a qual chama a atenção por ter sido desenvolvida para ser executada sobre plataformas embarcadas em tempo real, além de atingir taxas de acerto bastante satisfatórias. Em outras palavras, a proposta é uma junção de algoritmos já consolidados (nenhum algoritmo novo foi proposto), e faz uso das técnicas-chave de detecção facial e estimação da pose da cabeça, implementadas em sequência. Ambas serão abordadas nas seções seguintes.

3.2.1 Detecção Facial

A primeira etapa do reconhecimento dos gestos da cabeça consiste em detectar um rosto através do algoritmo proposto por Viola e Jones (VIOLA; JONES, 2001; VIOLA; JONES, 2004), implementado no método `detectMultiScale()` da `OpenCV`. Classificadores em cascata (*haar cascade*) são aplicados sobre as imagens capturadas pela *webcam* para tentar detectar um rosto. Se encontrado, um LED laranja, conectado ao C.H.I.P., acende, indicando que uma face foi detectada e que o algoritmo está se comportando conforme esperado; e a localização dos olhos e do nariz (baseada no modelo geo e antropométrico da face humana e calculada de forma puramente heurística e relativamente precisa) é estimada no plano bidimensional. Alguns resultados do processo de detecção facial e estimação dos três pontos são ilustrados na Figura 14.

[1] <https://www.youtube.com/watch?v=KMdqEHMWyvA>

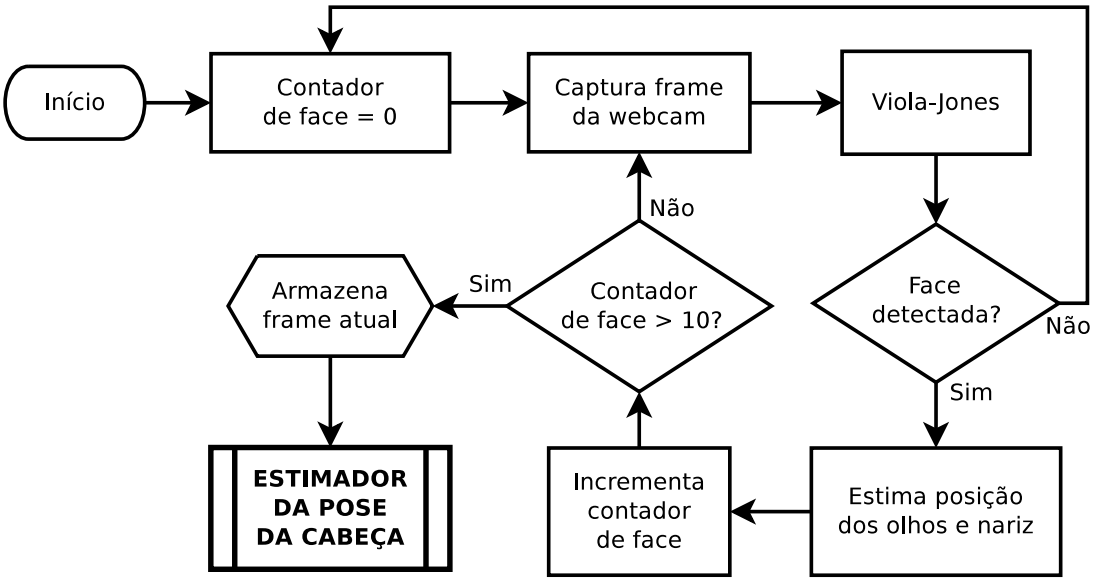
Figura 14 – Detecção facial, realizada pelo algoritmo Viola-Jones, e cálculo heurístico da posição dos olhos e do nariz sob duas condições de iluminação diferentes, com e sem o uso de óculos.



Fonte: Elaborada pelo autor.

O fluxograma do algoritmo de detecção facial descrito é mostrado na Figura 15. Se houver uma face em dez *frames* consecutivos, o 11° *frame* é armazenado, e a etapa seguinte de estimação da pose da cabeça, sinalizada pelo acendimento de um LED verde, é iniciada. Enquanto a condição de haver um rosto em dez quadros consecutivos não for atendida, a captura e classificação dos *frames* continua. Se os classificadores falharem na tarefa de detectar uma face, um LED vermelho sinaliza o erro.

Figura 15 – Algoritmo implementado para detecção facial.



Fonte: Elaborada pelo autor.

3.2.2 Estimação da Pose da Cabeça

A segunda etapa realiza, efetivamente, o reconhecimento dos gestos através da estimação da pose da cabeça. O fluxo óptico é calculado através do algoritmo de Lucas-Kanade, apresentado em (LUCAS; KANADE, 1981) e implementado no método `calcOpticalFlowPyrLK()` da `OpenCV`, o qual utiliza como referência os pontos dos olhos e do nariz que foram passados

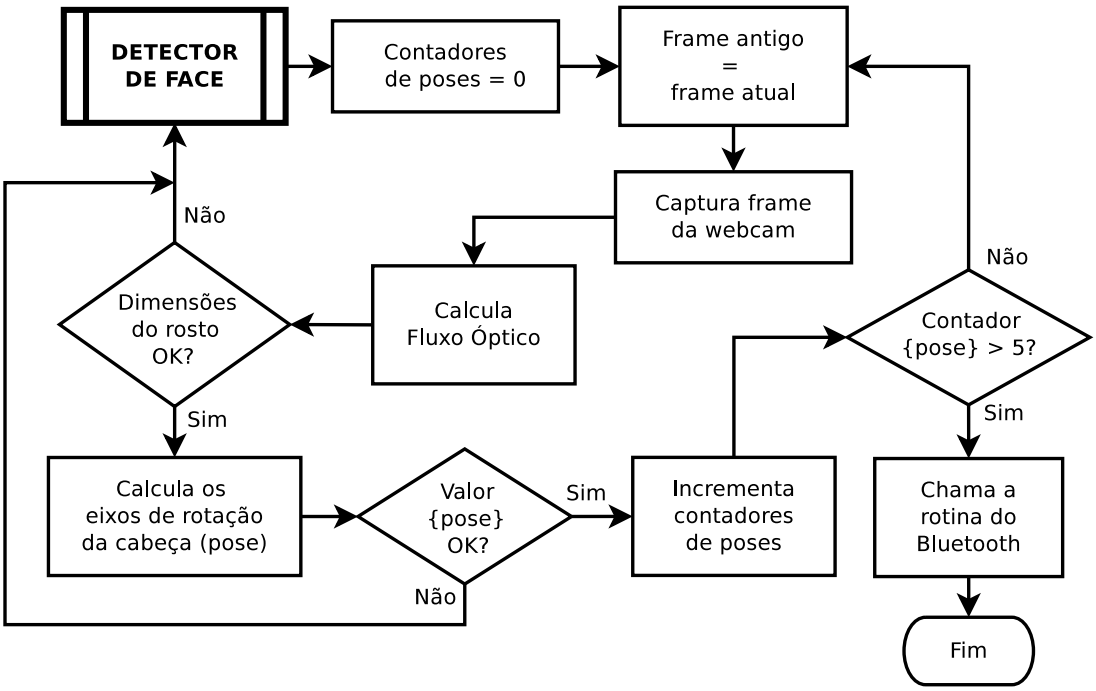
na etapa anterior, a fim de rastreá-los em relação a um próximo *frame* recém capturado pela câmera. Para tornar o algoritmo proposto mais confiável, as dimensões do rosto são verificadas em três condições simples que definem limites heurísticos com base na distância euclidiana. Sendo P_{OE} o ponto de localização do olho esquerdo; P_{OD} a localização do olho direito; e P_N a do nariz, ambos no *frame* atual:

- A razão entre a distância do olho esquerdo para o direito $||P_{OE}P_{OD}||$ e a distância de cada um dos olhos para o nariz $||P_{O\forall}P_N||$ deve estar entre 0,5 e 3,5.
- A distância entre o olho esquerdo e o direito $||P_{OE}P_{OD}||$ deve estar entre 20 e 160.
- A distância entre qualquer dos olhos e o nariz $||P_{O\forall}P_N||$ deve estar entre 10 e 150.

De posse das coordenadas dos pontos P_{OE}, P_{OD} e P_N no plano bidimensional e assumindo que as dimensões da face atendem às condições impostas acima, é possível calcular os eixos tridimensionais de rotação da cabeça (a saber: *roll*, *yaw*, e *pitch*). Os valores das poses foram filtrados com *thresholds* heurísticos baseados em movimentos suaves e pouco amplos, na tentativa de atender às limitações das pessoas com deficiência e, além disso, reduzir ruídos. Os valores são mostrados abaixo.

- *Roll*: entre -60 e -15 para movimentos para a direita e entre 15 e 60 para a esquerda;
- *Yaw*: entre 0 e 20 para movimentos para a direita e entre -20 e 0 para a esquerda;
- *Pitch*: entre 0 e 20 para movimentos para a baixo e entre -20 e 0 para cima;

Figura 16 – Algoritmo implementado para estimação da pose da cabeça.



Fonte: Elaborada pelo autor.

O algoritmo para a estimação da pose da cabeça é ilustrado no fluxograma da Figura 16. Se o movimento de uma das três poses for detectado por 5 *frames* consecutivos, o gesto é reconhecido de forma correta e o respectivo sinal é enviado para o Arduino via Bluetooth. A API em C/C++ desenvolvida sobre a biblioteca `BlueZ` foi utilizada para iniciar a comunicação entre as duas plataformas embarcadas.

3.3 Reconhecimento de Voz

O sistema de reconhecimento de voz implementado tem como base o pacote de *software* PocketSphinx (HUGGINS-DAINES et al., 2006). Os mesmos seis comandos básicos de controle da TV foram implementados em 10 sentenças possíveis, conforme mostrado na Tabela 4, a qual também exibe a frase-chave para acionamento do sistema.

Tabela 4 – Ações de comando na televisão e seus respectivos comandos de voz.

Ação	**Comando de voz**
Ativar o sistema	"acordar sistema"
Ligar o televisor	"ligar televisão"
Mudar para o próximo canal	"canal mais" ou "próximo canal"
Mudar para o canal anterior	"canal menos" ou "canal anterior"
Aumentar o volume	"aumentar volume" ou "volume mais"
Diminuir o volume	"diminuir volume" ou "volume menos"
Desligar o televisor	"desligar televisão"

Fonte: Elaborada pelo autor.

Um modelo acústico para fala contínua foi treinado através do pacote `SphinxTrain` com uma base de dados de aproximadamente 110 mil arquivos do tipo `.wav`, amostrados em 16 kHz, totalizando 164 horas e 41 minutos de áudio transcrito. Já o modelo de linguagem criado é dado por uma gramática livre de contexto, já que o vocabulário é pequeno.

A construção do dicionário fonético para o PT_BR foi feita por meio do conversor grafema-fonema, também chamado de `G2P`, um *software* disponibilizado pelo grupo FalaBrasil que recebe uma lista de palavras (grafemas) como entrada e gera na saída as suas transcrições

```
1  desligar     d e s l i g a xm
2  televisao    t e l e v i z aa ww
3  aumentar     a uu m ee t a xm
4  diminuir     dz ii m ii n u j xm
5  volume       v o l uu m i
6  proximo      p r om sm ii m u
7  anterior     aa t e r i o xm
8  ligar        l i g a xm
9  canal        k a n a w
10 mais         m a j s
11 menos        m ee n u s
12 acordar      a k o rm d a xm
13 sistema      s i s t ee m a
```

Listagem 2 – Dicionário fonético utilizado na aplicação (`.dict`).

fonéticas (fonemas) (SIRAVENHA et al., 2008). O dicionário fonético, mostrado na Listagem 2, é utilizado para converter os símbolos terminais da gramática (palavras) em fonemas, os quais são os *labels*-base para as HMMs utilizados no aprendizado do modelo acústico.

Já a gramática livre de contexto, que é utilizada para restringir o vocabulário a um pequeno conjunto de opções, define as regras de produção das palavras, de modo a gerar somente uma das sentenças listadas previamente definidas. A gramática segue o formato de representação do Java para sistemas de reconhecimento de fala (JSGF, do inglês *Java speech grammar format*). As regras de produção utilizadas para gerar os símbolos (sentenças produzidas pelo decodificador) são definidas no arquivo de extensão `.jsgf`, mostrado na Listagem 3.

```
1 #JSGF V1.0;
2 grammar gram;
3 public <command> = <tv> | <channel> | <volume> ;
4
5 <tv>         = (ligar | desligar) televisao ;
6 <volume>     = (aumentar | diminuir) volume | volume <intensity> ;
7 <channel>    = canal anterior | proximo canal | canal <intensity> ;
8 <intensity> = mais | menos ;
```

Listagem 3 – Gramática estilo JSGF para o Sphinx (`.jsgf`).

Como o tutorial para uso da API do PocketSphinx pela linguagem C (CMUSPHINX, 2017a) ensina apenas a decodificar arquivos de áudio; e a documentação oficial da mesma API (HUGGINS-DAINES, 2017) não faz menção a qualquer função para receber áudio em tempo real do microfone, o código fonte do PocketSphinx foi consultado e, através da análise da função `recognize_from_microphone()` e da ajuda da comunidade de usuários do Raspberry Pi[2], foi possível criar uma função, chamada de `ps_decode_from_mic()`, que consegue receber amostras de áudio pelo microfone e repassá-las ao *engine* de processamento de fala.

3.3.1 Acionamento por Palavra-Chave

O sistema ASR, além de ter sido configurado no modo de comando e controle para agir como método de entrada do protótipo, também foi configurado para agir como um acionador através da identificação de uma frase específica e isolada dentro um vocabulário de apenas uma sentença possível. Tal técnica, conhecida como *keyword spotting* (CHEN; PARADA; HEIGOLD, 2014), permite que o espaço de busca do decodificador seja restringido a apenas uma opção, o que permite que um limiar sobre o nível de confiança da sentença reconhecida seja definido (95%, por exemplo) para que ela seja aceita (reconhecida com sucesso) ou rejeitada.

[2] <http://www.robotrebels.org/index.php?topic=239.0>

A Listagem 4 contém um trecho do código implementado para o módulo de reconhecimento de voz que mostra tanto o modo de acionamento (*keyword spotting*) quanto o modo de comando e controle (gramática). A função `ps_set_keyphrase()` define a sentença "acordar sistema" como palavra-chave (ou, de forma mais literal, frase-chave). Já a função `ps_set_fsg()` aloca os recursos da gramática JSGF, previamente carregada de um arquivo. A troca entre os modos de reconhecimento é efetivamente realizada pela função `ps_set_search()`.

```
1  #include <pocketsphinx.h>
2  ps_decoder_t *ps;      // acesso ao engine decodificador
3  ad_rec_t     *ad;      // acesso ao microfone (conversor A/D)
4  cmd_ln_t     *config; // define configuracoes
5
6  ps_set_keyphrase(ps, "kws", "acordar sistema"); // define frase-chave
7
8  ps_set_jsgf_file(ps, "jsgf", "gram.jsgf");    // carrega arquivo JSGF
9  ps_set_fsg(ps, "fsg", ps_get_fsg(ps, "jsgf")); // define a gramatica
10
11 ps_set_search(ps, "kws"); // troca para o modo 'keyword spotting'
12
13 std::string sent = "";
14 do {
15     sent = ps_decode_from_mic(ps, ad); // procura apenas frase-chave
16 } while(sent == ""); // reijeicao retorna ""
17
18 ps_set_search(ps, "fsg"); // troca para o modo 'gramatica'
19 sent = ps_decode_from_mic(ps, ad); // procura entre 10 sentencas
20 std::cout << "sentence: " << sent << std::endl;
```

Listagem 4 – Trecho do código de reconhecimento de fala que explicita a troca do modo *keyword spotting* para o modo gramática.

3.4 Controle Remoto do Televisor

A plataforma Arduino UNO foi utilizada para receber uma *string* diferente para cada respectivo gesto reconhecido no microcomputador e, por fim, traduzi-la em um comando específico para ser enviado a uma TV da Samsung. A informação do reconhecimento do gesto é recebida através de um *socket* serial criado com auxílio da biblioteca `SoftwareSerial`, conforme exemplificado no trecho de código da Listagem 5.

Os comandos da TV foram obtidos através da biblioteca `IRremote`. O sensor infravermelho VS 18388 foi conectado ao pino digital D11 (PWM) do Arduino UNO para receber os comandos do controle original do aparelho televisor, exibindo no monitor serial a informação

```
#include <SoftwareSerial.h>
SoftwareSerial serialBT(2, 5); // Rx e Tx

int i=0; // index para str
char ch, str[10] = {0};
do {
    ch = (char) serialBT.read(); // recebe char por char
    if (ch != '\n' && ch != '\r')
        str[i++] = ch; // compoe string
    delay(2);
} while (serialBT.available() > 0); // enquanto a lib recebe chars
```

Listagem 5 – Trecho do código de recebimento de uma *string* via Bluetooth no Arduino com auxílio da biblioteca `SoftwareSerial`.

referente aos *bits* do protocolo de comunicação estabelecido entre o controle remoto e a respectiva TV. Esses comandos foram salvos e, então, a mesma biblioteca foi utilizada para enviá-los ao aparelho através de um LED infravermelho, o qual foi conectado ao pino de PWM D2 do Arduino e inserido em um circuito amplificador, o qual foi construído com base no transistor NPN BC548 para garantir um maior alcance do sinal modulado. Trechos dos códigos de recebimento e envio de sinais infravermelhos são mostrados na Listagem 6.

```
#include <IRremote.h>
IRrecv irrecv(11); // IROut: pino PWM D11
IRsend irsend;     // IRIn:  pino PWM D2  (default)

/* Recebimento do sinal IR */
decode_results res;
irrecv.decode(&res);
irrecv.resume();

/* Envio do sinal IR
 * vol+: value=0xE0E0E01F, bits=32 */
irsend.sendSAMSUNG(res.value, res.bits);
```

Listagem 6 – Trecho dos códigos de recebimento e envio de sinais infravermelhos pelo Arduino com auxílio da biblioteca `IRremote`.

4 TESTES E RESULTADOS

Os testes foram realizados com pessoas de diferentes faixas etárias e graus de instrução, com diagnóstico de alguma doença que ocasionou deficiência ou mesmo sem limitação alguma. Vale ressaltar que deficiência motora dos membros superiores não significa amputação ou paralisia total, assim como deficiência visual ou auditiva, por exemplo, não significa cegueira ou surdez completas, respectivamente.

A realização dos testes acompanhou a conclusão das fases de implementação dos dois métodos de entrada do sistema. Em outras palavras, ao término do desenvolvimento dos módulos de reconhecimento de gestos e de voz, uma seção de testes com voluntários foi executada logo em seguida. Sendo assim, duas baterias de teste ocorreram: uma para o sistema unimodal, realizada quando apenas o sistema AGR encontrava-se funcional; e uma segunda já com o sistema multimodal, na qual os participantes utilizaram tanto os gestos da cabeça quanto a própria voz para controlar o aparelho televisor.

O sistema de reconhecimento de gestos foi testado em conjunto com o circuito acionador externo, enquanto o sistema de comandos de voz foi acionado através do mecanismo de reconhecimento de palavra-chave. A escolha desse par de combinações independentes deve-se ao fato de ambos os sistemas AGR e ASR consumirem bastante recursos da plataforma embarcada, o que tornou impraticável a execução simultânea dos dois métodos de entrada.

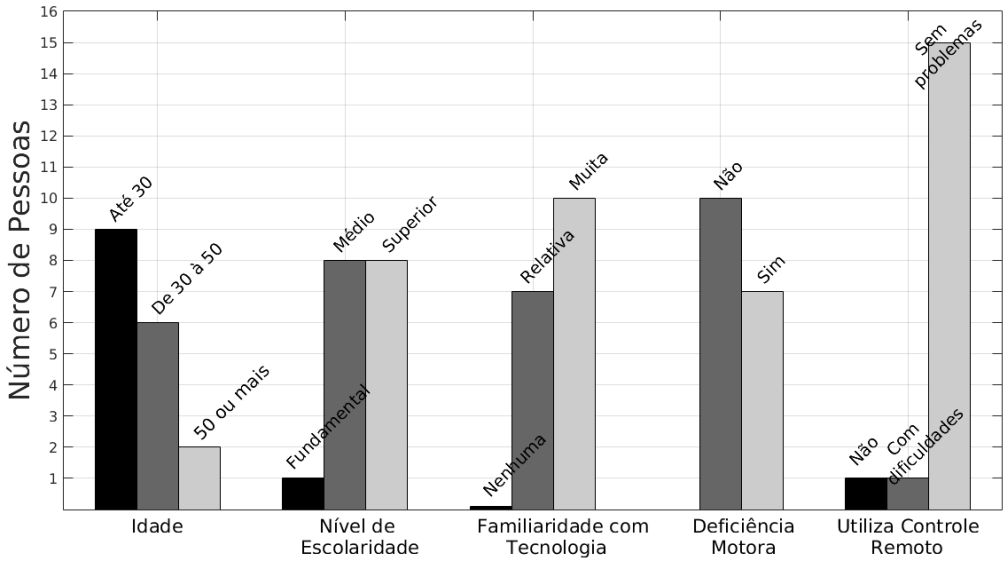
Uma avaliação objetiva foi realizada sobre os sistemas AGR e ASR, de forma individual, com base na contagem do número de tentativas que os usuários precisaram realizar em cada interação para que o comando correto fosse enviado à TV. Em contrapartida, todo o conjunto do protótipo em si foi avaliado de forma geral através de uma análise subjetiva feita pelos usuários através de um questionário pessoal.

Todos os participantes foram convidados a preencher um formulário simples, de modo que um perfil de pessoas fosse formado durante a análise para auxiliar na interpretação dos resultados dos testes. Os voluntários foram questionados sobre faixa etária; grau de escolaridade; tipo de deficiência; familiaridade com o uso de tecnologia e aparelhos eletrônicos; e facilidade no manuseio de controles remotos convencionais.

4.1 Perfil dos Participantes

Durante os testes preliminares, o grupo de participantes que realizou o controle da TV apenas através de gestos da cabeça foi composto por 12 pessoas do sexo masculino e cinco do sexo feminino, totalizando 17 pessoas com idades variando de 12 a 62 anos. O perfil demográfico dos participantes que testaram o sistema AGR é mostrado na Figura 17.

Figura 17 – Perfil demográfico dos participantes envolvidos nos testes com o sistema unimodal, somente com o módulo de reconhecimento de gestos da cabeça.



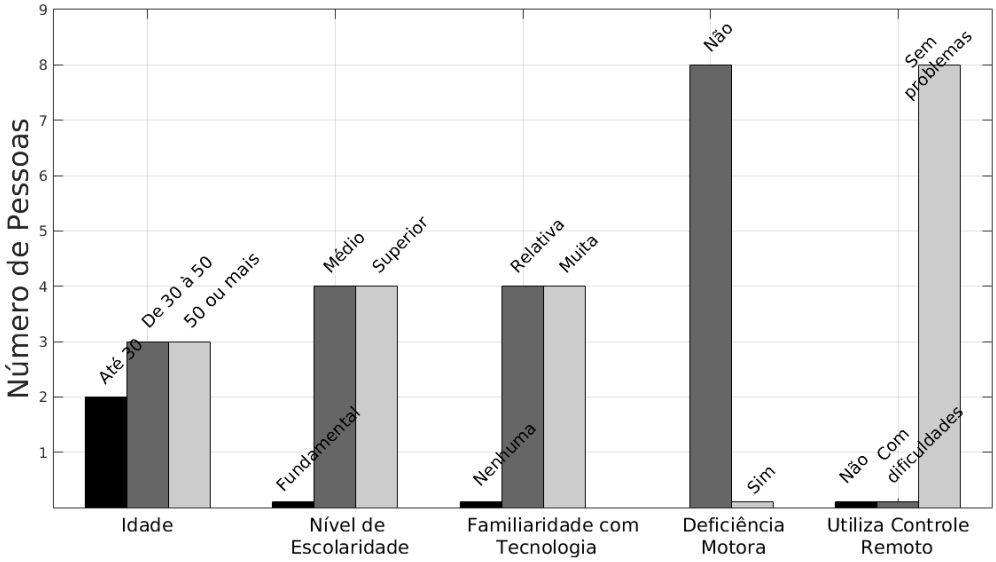
Fonte: Elaborada pelo autor.

Sete participantes apresentaram diagnóstico de deficiência motora. Um deles é portador de vasculite, uma síndrome que ocasiona desequilíbrio e baixa coordenação motora. Este também informou que utiliza controles remotos com certa dificuldade. Outro participante é portador de quadriplegia espástica, o que causa problemas de mobilidade nos membros inferiores e superiores. Este, por sua vez, declarou utilizar controles remotos sem nenhum problema. Um terceiro voluntário, desta vez portador de paralisia cerebral, disse não ser capaz de utilizar controles remotos devido à atrofia severa nos dedos das mãos. Sua coordenação, no entanto, foi suficiente para a realização de alguns movimentos do pescoço com sucesso. Os demais quatro participantes eram cadeirantes com limitações apenas no movimento das pernas (paraplegia).

Três outros voluntários portadores de paralisia cerebral não foram incluídos nos resultados, pois, apesar de terem o cognitivo preservado e entenderem bem os comandos dados, não possuíam controle suficiente sobre os movimentos da cabeça, o que os impediu de executar qualquer um dos movimentos requeridos de forma efetiva.

Já o grupo de participantes que testou o protótipo multimodal, tanto por comandos de voz quanto por gestos da cabeça, foi composto por três pessoas do sexo masculino e cinco do sexo feminino, totalizando oito pessoas com idades variando de 24 a 63 anos. O perfil demográfico dos participantes que testaram ambos os sistemas ASR e AGR é mostrado na Figura 18. É necessário mencionar aqui que nenhum dos voluntários da segunda bateria de testes declarou possuir qualquer tipo de deficiência física, tampouco relatou ter qualquer dificuldade na utilização de controles remotos convencionais.

Figura 18 – Perfil demográfico dos participantes envolvidos nos testes do sistema multimodal, já com ambos os módulos de controle por gestos e voz.



Fonte: Elaborada pelo autor.

4.2 Ambiente de Testes

Sobre o sistema de reconhecimento de gestos, os experimentos com pessoas sem deficiência foram realizados, em sua maioria, no laboratório de teste da universidade, onde a iluminação é favorável e controlada. Já as PCD executaram o teste em uma sala especial das respectivas casas de apoio onde fazem tratamento, também com ambiente de iluminação controlado, enquanto uma única PCD realizou o teste em sua própria residência, onde a iluminação era fraca. Como os classificadores utilizados no Viola-Jones não foram treinados com exemplos de rostos em ambientes pouco iluminados, este é um fator fundamental na acurácia do sistema.

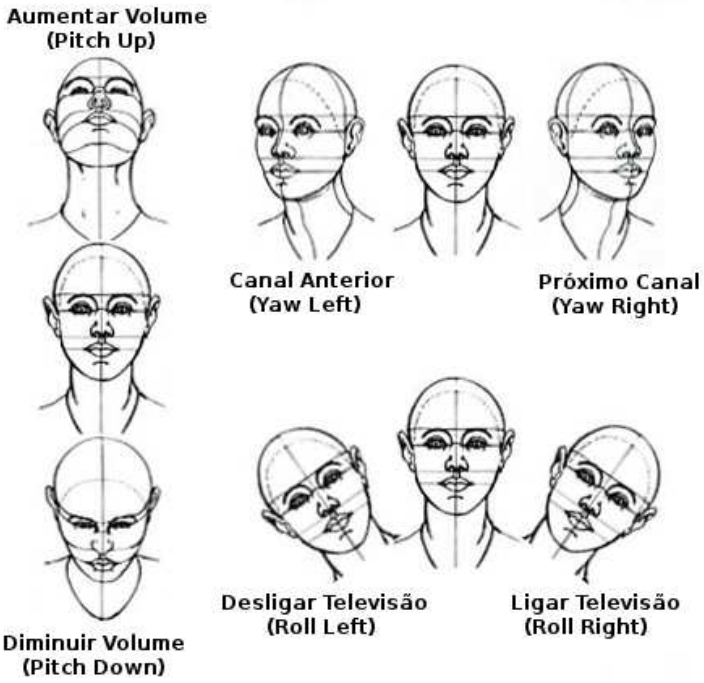
Analogamente, todos os testes com o sistema de reconhecimento de fala foram executados também em um ambiente controlado e pouco ruidoso, o que é fundamental para o seu bom desempenho, já que os exemplos de áudio utilizados no treino do modelo acústico também foram gravados em ambiente com baixo nível de ruído.

Após uma breve introdução sobre o grupo de pesquisa e sobre os objetivos do projeto, o funcionamento do sistema proposto foi explicado de forma verbal a cada um dos participantes e, então, os testes com os dois métodos de entrada (gestos e voz) foram iniciados.

4.2.1 Gestos da Cabeça

Os movimentos da cabeça que deveriam ser executados em frente à câmera foram ilustrados e apresentados aos voluntários como forma de manual, conforme mostrado na Figura 19. O manual esteve à disposição para ser consultado pelos usuários durante todo o período do teste.

Figura 19 – Movimentos de rotação da cabeça (ou poses) traduzidos em comandos para um televisor.



Fonte: Adaptada da Internet.

Os usuários foram, então, instruídos a seguir um roteiro de tarefas com os seis comandos para controle de um televisor através de gestos da cabeça. As tarefas foram listadas em ordem de execução, juntamente com as respectivas descrições, conforme mostrado na Tabela 5. Os participantes também foram instruídos a desistir após a quarta tentativa de cada comando, caso o sistema não se comportasse conforme desejado, e a utilizar o acionador externo sem fio para inicializar o sistema AGR.

Tabela 5 – Roteiro de tarefas a serem completadas pelo usuário utilizando gestos da cabeça como entrada.

	Ação	**Movimento**	**Descrição**
1	Ligar televisão	*Roll right*	Encostar a orelha no ombro direito
2	Próximo canal	*Yaw right*	Encostar o queixo no ombro direito
3	Aumentar volume	*Pitch up*	Olhar para o teto
4	Canal anterior	*Yaw left*	Encostar o queixo no ombro esquerdo
5	Diminuir volume	*Pitch down*	Olhar para baixo
6	Desligar televisão	*Roll left*	Encostar a orelha no ombro esquerdo

Fonte: Elaborada pelo autor.

4.2.2 Comandos de Voz

Os comandos de voz a serem dados para o sistema foram ilustrados e apresentados aos voluntários juntamente com a respectiva ação referente, conforme mostrado na Tabela 6. Novamente, a partir desse ponto, os usuários foram instruídos a seguir o roteiro de tarefas e convidados a desistir após uma quarta tentativa frustrada de cada comando. Os voluntários também foram continuamente lembrados de inicializar o sistema através da "frase-chave".

Tabela 6 – Roteiro contendo as ações na TV e seus respectivos comandos de voz.

	Ação	**Comando de voz**
#	Ativar o sistema	"acordar sistema"
1	Ligar televisão	"ligar televisão"
2	Próximo canal	"canal mais" ou "próximo canal"
3	Aumentar volume	"aumentar volume" ou "volume mais"
4	Canal anterior	"canal menos" ou "canal anterior"
5	Diminuir volume	"diminuir volume" ou "volume menos"
6	Desligar televisão	"desligar televisão"

Fonte: Elaborada pelo autor.

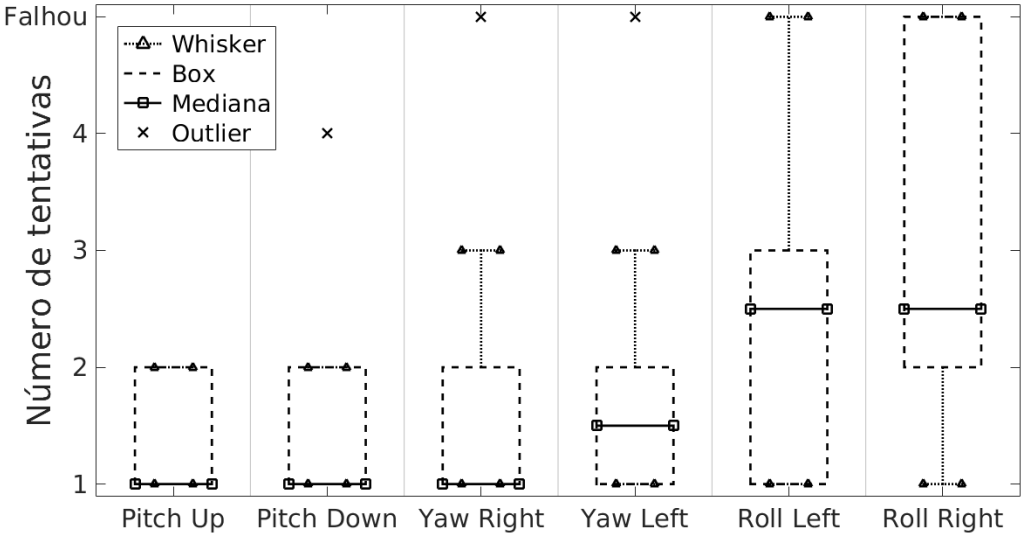
4.3 Resultados

Os resultados da avaliação objetiva serão apresentados através de *boxplots*, os quais mostram a distribuição do número de tentativas executadas por cada usuário a fim de realizar cada um dos seis comandos de controle da TV. Os *whiskers*, representados por linhas pontilhadas delimitadas por dois triângulos, são os números mínimo e máximo de tentativas que não são considerados *outliers*. Já as "caixas" (*boxes*), representadas por linhas tracejadas, abrigam a metade central da distribuição, ou seja, 25% dos valores acima da mediana (indicada pela linha sólida entre marcadores quadrados) e 25% dos valores abaixo. Por fim, o resultado da avaliação subjetiva, com base na opinião dos usuários sobre o protótipo, também será apresentado.

4.3.1 Avaliação Objetiva: Sistema Unimodal

Analisando a Figura 20, pode-se perceber que, para as pessoas sem deficiência, os movimentos de *pitch* e *yaw*, além de terem os limites das caixas entre 1 e 2, ainda obtiveram valor de mediana entre 1 e 1,5, o que significa que a maioria das pessoas precisou de apenas uma ou duas tentativas para realizar os respectivos comandos com sucesso. No entanto, houve um único caso em que ambos os movimentos de *yaw* não foram executados com sucesso, o que explica o valor *outlier* exibido sobre as respectivas caixas. Já a rotação lateral (*roll*) foi o movimento que mais foi executado de forma incorreta, o que tornou os valores mais próximos de 4, que é o limite máximo de tentativas permitidas no roteiro de tarefas.

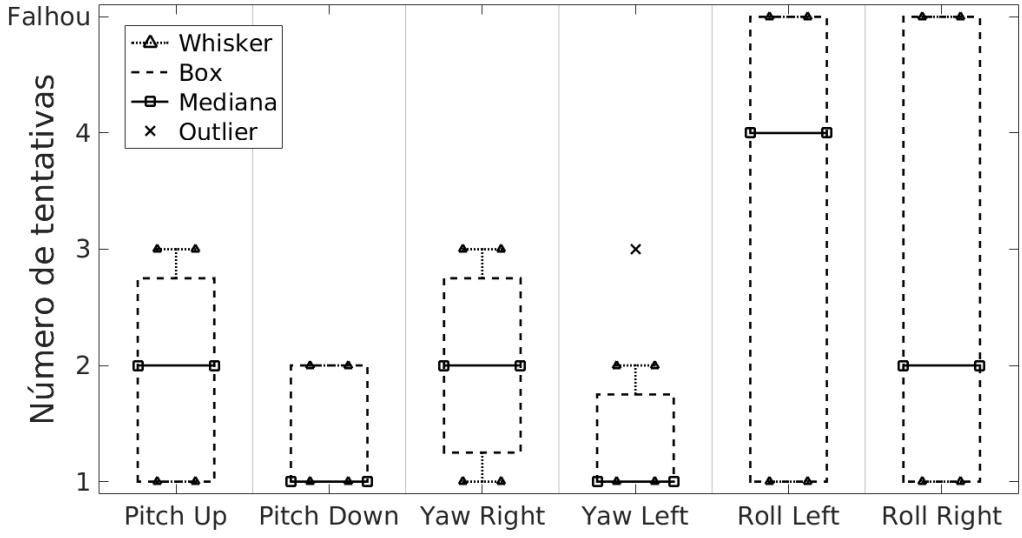
Figura 20 – Número de tentativas de realização dos movimentos da cabeça por pessoas sem deficiência durante o teste com o sistema unimodal.



Fonte: Elaborada pelo autor.

A análise da Figura 21 confirma que o mesmo problema do *roll* ocorreu com as pessoas com deficiência, porém com uma frequência um pouco maior devido a algumas limitações na rotação da cabeça. No entanto, mais uma vez, *pitch* e *yaw* tiveram bons resultados, o que pode ser observado pelas caixas limitadas pelos valores 1 e 2,75, com medianas entre 1 e 2. Ou seja, as PCD precisaram de no máximo 3 tentativas para realizar os movimentos de forma correta.

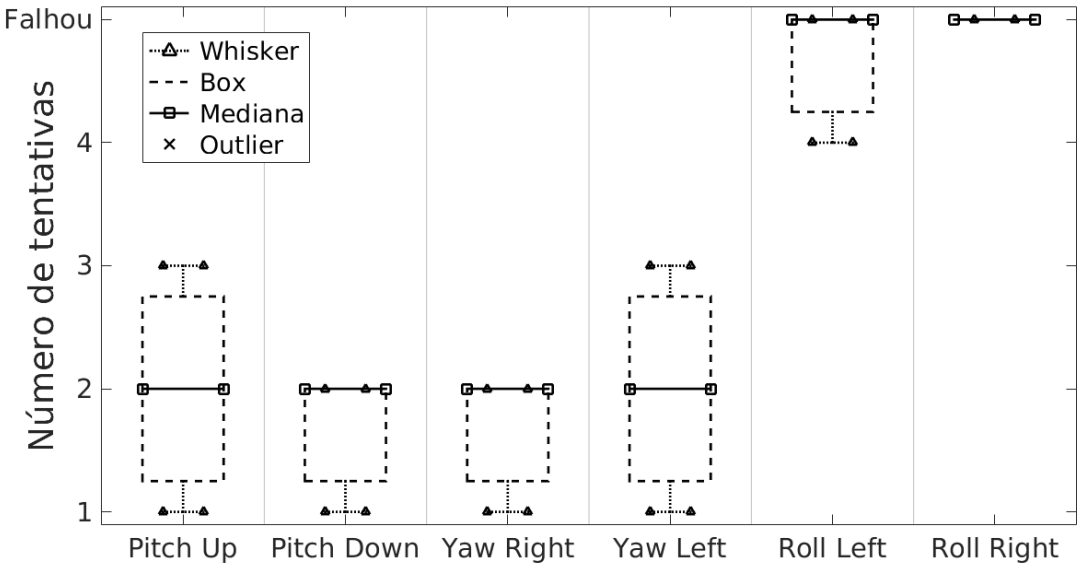
Figura 21 – Número de tentativas de realização dos movimentos da cabeça por pessoas com deficiência motora durante o teste com o sistema unimodal.



Fonte: Elaborada pelo autor.

Já a Figura 22 mostra a distribuição das tentativas de movimento da cabeça somente para as pessoas com deficiência dos membros superiores. Como suas respectivas deficiências ocasionam comprometimento da maior parte do tronco, o que inclui também o pescoço, a dificuldade em executar o *roll* foi sentida com maior intensidade. Apenas um dos três voluntários conseguiu desligar o aparelho televisor já na última tentativa, enquanto houve falha em todas as outras tentativas de ligar e desligar o aparelho, o que explica o valor de mediana tão alto. Já o desempenho nos movimentos de *yaw* e *pitch* foram novamente bons, mantendo as medianas em 2 e os limites entre 1,25 e 2,75.

Figura 22 – Número de tentativas de realização dos movimentos da cabeça por pessoas com deficiência motora dos membros superiores durante o teste com o sistema ainda unimodal.

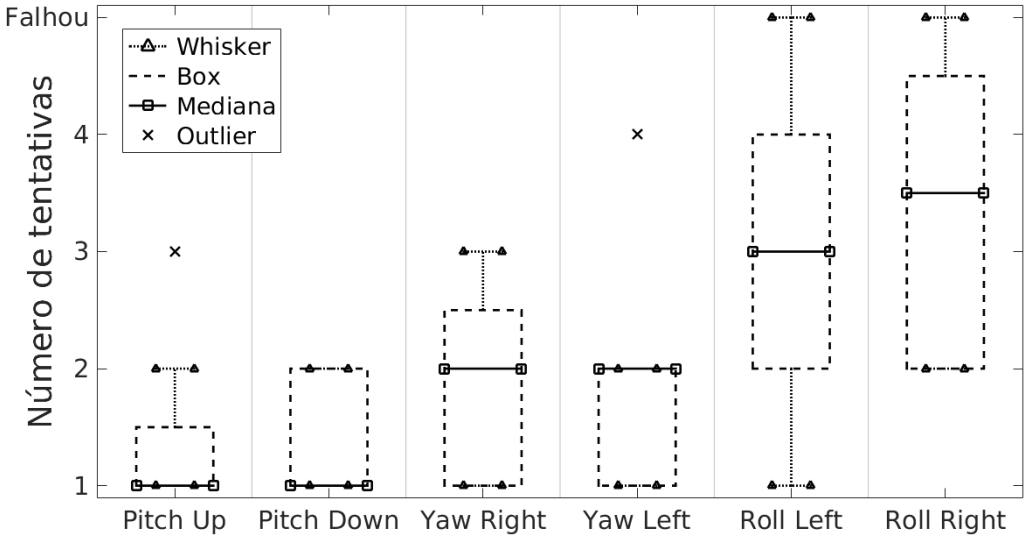


Fonte: Elaborada pelo autor.

4.3.2 Avaliação Objetiva: Sistema Multimodal

Com relação a segunda bateria de testes, agora com o sistema multimodal, a Figura 23 reforça a análise dos resultados exibidos nos *boxplots* anteriores, mais especificamente no da Figura 20, o qual aborda o números de tentativas para pessoas sem deficiência durante o uso do sistema de reconhecimento de gestos. A distribuição de no máximo três tentativas foi mantida para os movimentos *pitch* e *yaw*, com respectivas medianas de valor 1 e 2 e dois *outliers* que representam relativas e isoladas dificuldades ao aumentar o volume e voltar para o canal anterior (três e quatro tentativas em *pitch up* e *yaw left*, respectivamente). Já o movimento *roll* apresentou novamente o pior desempenho, resultando em uma ampla distribuição com mediana de 3 e 3,5 tentativas. De forma exclusiva em relação aos outros dois pares de movimentos, o *roll* também gerou algumas falhas nas tentativas de ligar e desligar a TV.

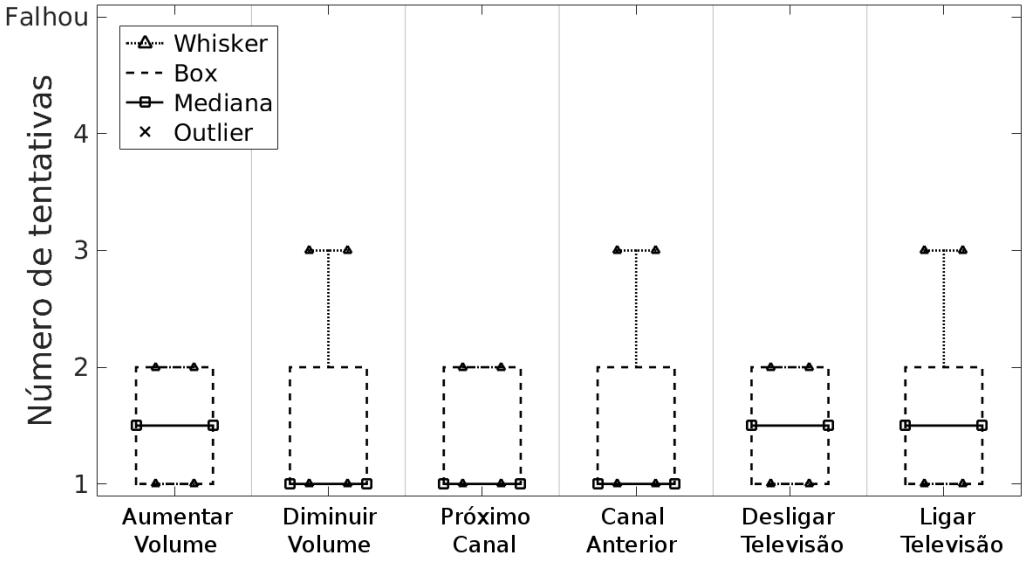
Figura 23 – Número de tentativas de realização dos movimentos da cabeça por pessoas sem deficiência durante o teste com o sistema multimodal.



Fonte: Elaborada pelo autor.

Por fim, a Figura 24 mostra os resultados obtidos para o teste com o sistema de reconhecimento de voz. De início, pode-se imediatamente perceber que o método de entrada por comando de voz obteve o melhor desempenho, se comparado aos resultados anteriores do sistema de reconhecimento de gestos. O valor da mediana de todas as *boxes* não passou de 1,5 e o número de tentativas (repetições do mesmo comando de voz) chegou no máximo a 3. A maioria dos participantes conseguiu executar o comando no aparelho televisor já na primeira tentativa.

Figura 24 – Número de tentativas de realização dos comandos de voz por pessoas sem deficiência durante o teste com o sistema multimodal.



Fonte: Elaborada pelo autor.

4.3.3 Avaliação Subjetiva

Em relação ao questionário preliminar aplicado aos voluntários após o teste com o sistema unimodal, a distribuição das respostas entre pessoas sem deficiência (não PCD) e com deficiência (PCD) é mostrada na Tabela 7. Dos 17 entrevistados, seis deram nota 4 (bom) para o sistema e 11 deram nota 5 (excelente), resultando em um escore geral de 4,647 pontos.

Tabela 7 – Resultado da avaliação subjetiva do sistema unimodal (somente gestos da cabeça) através do questionário MOS.

Pontuação	**Votos (não PCD)**	**Votos (PCD)**	**Total**
5	6	5	11
4	4	2	6
3	∅	∅	∅
2	∅	∅	∅
1	∅	∅	∅
MOS	4,6	4,7	**4,647**

Fonte: Elaborada pelo autor.

Já a Tabela 8 mostra o escore dado pelos participantes que testaram tanto a entrada por gestos quanto por comandos de voz do sistema multimodal. Dos oito voluntários, quatro deram nota 4 (bom) e os quatro restantes deram nota 5 (excelente), totalizando 4,5 pontos no geral.

Tabela 8 – Resultado da avaliação subjetiva do sistema multimodal (com entrada por gestos da cabeça e comandos de voz) através do questionário MOS.

Pontuação	**Votos**
5	4
4	4
3	∅
2	∅
1	∅
MOS	**4,50**

Fonte: Elaborada pelo autor.

Os escores gerais alcançados tanto na avaliação do sistema unimodal quanto na do multimodal provam que a proposta tem, de fato, um potencial para ajudar pessoas com limitações motoras dos membros superiores e movimentos do pescoço ou fala preservados a ter mais independência e autonomia na tarefa de controle de equipamentos eletrônicos.

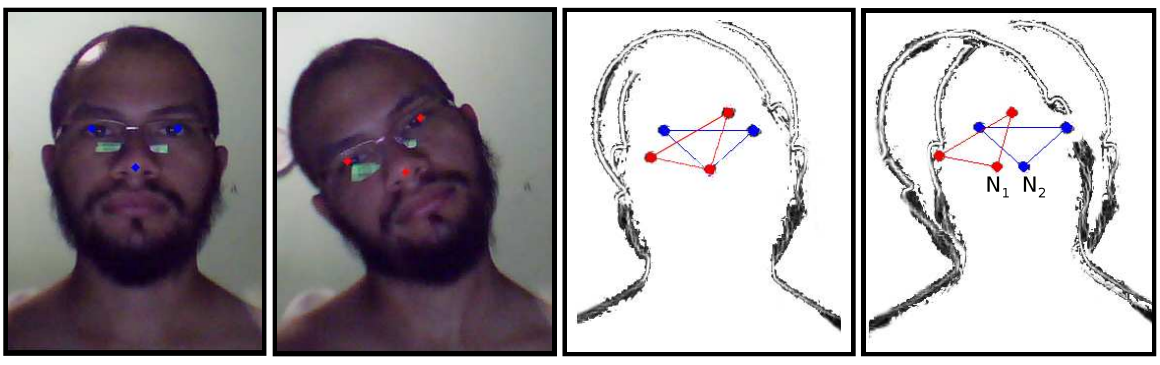
4.4 Discussão

Os participantes que possuem deficiência mostraram-se ansiosos pela eventual disponibilização do protótipo como um produto para ter um módulo "particular em casa", conforme relatado, e interessaram-se por custos e pela possibilidade de controlar diversos outros aparelhos, incluindo aparelhos de som, ares-condicionados, dentre outros. Os profissionais das casas de apoio, por outro lado, relataram o potencial inovador do protótipo. Os demais participantes também apontaram o benefício de não precisar procurar pelo controle remoto de cada aparelho, já que o sistema ficaria centralizado no ambiente. Além disso, a funcionalidade de não precisar pressionar qualquer botão no acionador externo surpreendeu positivamente os voluntários sem comprometimento total dos membros superiores, os quais aprovaram a "nova tecnologia" de apenas aproximar umas das mãos sobre o aparelho para ativar o sistema.

Por outro lado, o sistema também mostrou desvantagens. A limitação na execução de certos movimentos, especialmente o *roll*, frustrou tanto os pesquisadores quanto os usuários, os quais mostraram-se bastante chateados quando não conseguiram concluir a tarefa, o que acabava por fazê-los sentir culpa por tal. Um participante, inclusive, sugeriu que alguma característica facial, como uma piscada longa dos olhos, poderia ser utilizada no lugar do *roll*, movimento que o deixou mais desconfortável durante o teste com o sistema AGR. Além disso, movimentos muito rápidos ou bruscos acabavam por tornar verdadeiras as condições definidas para aumentar a confiabilidade do sistema, pois dificultavam o cálculo do fluxo óptico, o que forçava o sistema a reportar uma situação de erro. Do mesmo modo, o fluxo calculado para movimentos muito lentos é tão baixo que não pode ser interpretado como angulação, pois teria a mesma magnitude de um ruído, o que influenciaria diretamente no desempenho do sistema.

A razão observada para o insucesso do *roll*, relatada também por Arcoverde Neto et al. (2014), recai sobre a dificuldade em realizar o movimento de forma isolada, já que a própria natureza do movimento compreende a execução de um *yaw* parcial. Em outras palavras, o modelo

Figura 25 – Deslocamento da localização do nariz ao longo do eixo x durante o movimento *roll*, contrastando com o modelo de rotação esperado.



Fonte: Elaborada pelo autor.

esperado de rotação lateral do *roll* considera um eixo fixo sobre o nariz, o que não ocorre na prática, conforme ilustrado na Figura 25: percebe-se que o ponto do nariz desloca-se levemente ao longo do eixo horizontal x, fato que se enquadra no critério estabelecido pela Equação 2.

A iluminação não controlada no ambiente residencial de um dos voluntários também foi um problema, já que o algoritmo Viola-Jones teve dificuldades em identificar um rosto, o que gerou certo desconforto e impaciência. Alguns participantes também possuíam uma postura incorreta, de modo que o algoritmo de detecção facial não funcionava por conta de a posição inicial do rosto não ser exatamente frontal e ereta. Além disso, a ausência de um *feedback* sonoro foi sentida durante os movimentos *pitch* e *yaw*, quando o participante perdia o contato visual com a tela e não sabia, por consequência, se o comando havia sido corretamente executado.

Sobre o sistema de reconhecimento de fala, o desempenho do PocketSphinx foi bastante satisfatório, o que pode ser refletido nos poucos erros apresentados no gráfico de tentativas. A parte mais incômoda, segundo os participantes, foi a necessidade de repetir a palavra-chave "acordar sistema" para inicializar o sistema antes de qualquer comando de controle. No mais, a maioria dos voluntários não demonstrou tanta surpresa ao controlar a TV com a voz como demonstrou ao controlá-la com os gestos da cabeã. Ao fim da segunda bateria de testes, inclusive, quando os participantes foram questionados a respeito de uma eventual preferência em relação aos dois métodos de interação adotados, a resposta quase unânime reporta que ter controlado a TV com os gestos da cabeça foi "mais legal e diferente" do que controlá-la através da voz, apesar da dificuldade encontrada na execução de determinados movimentos.

Segundo os resultados do estudo reportado em Kardaris et al. (2016) a respeito de aplicações multimodais utilizando o dispositivo Kinect, a taxa de acerto dos comandos dados via voz foi bem maior do que a dos comandos dados via gestos: 84% contra 57%. Ainda, em Teixeira et al. (2014) e Cavallo et al. (2013), por exemplo, dois estudos que tratam o acesso a serviços *online* e a interação com robôs assistentes pessoais, respectivamente, afirma-se que a voz é o método de interação mais adequado para pessoas idosas, onde as limitações são bastante frequentes e variadas. Os resultados deste trabalho mostram que o sistema ASR de fato obteve o melhor desempenho na execução da tarefa proposta, mas ainda assim foi o sistema de reconhecimento de gestos que provocou uma maior satisfação nos voluntários.

Sobre os participantes portadores de paralisia cerebral, todas as suas formas de interação são, geralmente, bastante restritas, o que torna a interface com aparelhos eletrônicos um desafio para a comunidade científica (POKHARIYA et al., 2006). Além disso, as limitações de duas pessoas com o mesmo diagnóstico de paralisia cerebral normalmente não são parecidas, tampouco iguais, já que tal condição pode resultar em déficit cognitivo, perda significativa da coordenação motora e problemas de grau variado na fala devido ao comprometimento do aparelho oromotor (SANKAR; MUNDKUR, 2005).

5 CONSIDERAÇÕES FINAIS

Este trabalho contribuiu com detalhes do desenvolvimento e implementação de um protótipo de controle remoto universal, multimodal, livre e de baixo custo, voltado especialmente para pessoas com deficiência motora dos membros superiores, com o objetivo de tornar o controle de equipamentos eletrônicos do ambiente doméstico mais prático, confortável e, principalmente, acessível nos métodos de interação e no custo total de desenvolvimento.

Os sistemas revisados que apresentam funcionalidade semelhante a do sistema proposto normalmente utilizam, como método de entrada não convencional, técnicas de rastreamento ocular; reconhecimento de gestos dos braços e/ou mãos ou de características faciais, como piscadas e movimentos da boca; ou somente o reconhecimento de fala. Dentre os sistemas que utilizam gestos da cabeça como interface de interação, apenas aplicações na área de controle de cadeira de rodas e de cursor do *mouse* foram encontradas.

Em suma, nenhuma das aplicações encontradas utiliza um sistema de reconhecimento de gestos da cabeça, baseado em técnicas de computação visual, para controlar aparelhos eletrônicos domésticos, tampouco o propõe em conjunto com um módulo de reconhecimento de fala para formar um sistema multimodal. Além disso, nenhum trabalho acadêmico encontrado foca no desenvolvimento de acionadores externos sem fio baseados em proximidade. Já as soluções encontradas no mercado já são fechadas como produtos e geralmente possuem um preço absurdamente elevado.

A expectativa é que o projeto continue a sua fase de desenvolvimento e otimização e que, nas próximas versões, o sistema atenda a um maior número de pessoas de diversos grupos demográficos, envolvendo, por fim, uma maior gama de deficiências assistidas. O C.H.I.P. é uma plataforma relativamente nova no mercado que ainda tem muito a ser explorada, o que abre a possibilidade de o computador embarcado vir a tornar-se melhor ao longo do tempo, do ponto de vista do desenvolvedor, à medida que sugestões vão sendo propostas pela comunidade e atualizações disponibilizadas pelos fabricantes.

Por outro lado, no âmbito de saída do sistema, é possível que as *smart* TVs também se tornem mais populares e facilmente adquiridas. A comunicação por rede local via Wi-Fi, por conta da praticidade e conforto, poria fim na tecnologia infravermelha que atualmente domina o controle dos equipamentos eletrônicos domésticos. Porém, de acordo com as estatísticas consultadas, é impossível prever por quanto tempo a população de países em desenvolvimento, como o Brasil, continuará a usar majoritariamente aparelhos eletrônicos que utilizam luz infravermelha como protocolo de controle remoto.

5.1 Dificuldades e Soluções

Durante a construção do projeto, diversos problemas foram enfrentados no que diz respeito à instalação, configuração e implementação dos diversos módulos individuais que compõem o sistema, em especial no microcomputador embarcado. Os mais relevantes são listados e discutidos a seguir.

Documentação do C.H.I.P.. Como o C.H.I.P. é uma plataforma de desenvolvimento relativamente nova no mercado, a documentação ainda é um pouco escassa. Recorrer aos fóruns[1] seria uma solução, mas a comunidade que o sustenta ainda não é tão ativa quanto a de plataformas similares, como Raspberry Pi ou BeagleBone. Felizmente, boa parte da documentação dos pacotes utilizados em plataformas *desktop* foi suficiente para reproduzir o funcionamento no computador embarcado.

Saída de vídeo do C.H.I.P.. O sistema de transmissão de vídeo do C.H.I.P., cuja saída é dada por uma porta de áudio *jack* P2 fêmea, segue o padrão norte americano do Comitê Nacional de Sistemas de Televisão (NTSC, do inglês *National Television System Comittee*). Já o padrão dos televisores brasileiros segue a codificação de linha de fase alternada do tipo M (PAL-M, do inglês *phase alternating line type M*). A documentação sugere o uso de um cabo UART específico para alterar o sistema de transmissão do microcomputador para PAL-M, porém a solução mais viável economicamente foi a de instalar o sistema operacional Debian em modo servidor, ou seja, sem interface gráfica, e utilizar o *software* `screen` para acessar a plataforma por meio de uma conexão USB com um computador *desktop*, o qual agia como *host*. Posteriormente, após a configuração da conexão Wi-Fi, o `SSH` foi utilizado para acesso via rede local.

PWM no C.H.I.P.. O uso do Arduino justifica-se unicamente pelo fato de, nas versões do *kernel* do Linux embarcado para o C.H.I.P., a modulação por largura de pulso (PWM) não estar devidamente acessível[2] ao programador: o *driver* que permite acesso ao processador não está implementado. Como o Arduino UNO possui 6 pinos de PWM e a biblioteca `IRremote` gerencia alguns desses pinos de forma otimizada, a plataforma conseguiu realizar com sucesso a tarefa de enviar comandos para a TV através do LED infravermelho.

Aquecimento do C.H.I.P.. Durante a fase de desenvolvimento, percebeu-se que o C.H.I.P. superaquecia quando os algoritmos de computação visual eram executados. Como consequência, o microcomputador normalmente desligava automaticamente. Este problema foi fundamental para a utilização do acionador, o qual também evita o consumo excessivo de processamento e energia do C.H.I.P..

1 <https://bbs.nextthing.co/>,<http://www.chip-community.org/>
2 <http://www.chip-community.org/index.php/PWM_support>

Acionador externo baseado em contato. Inicialmente, a ideia do acionador concentrou-se apenas em criar um dispositivo sem fio, com um botão comum, físico e pressionável como dispositivo de interação. Durante a fase de testes, entretanto, percebeu-se por observação que algumas pessoas com limitação motora dos membros superiores ainda sentiriam dificuldade ao pressionar o botão, visto que o contato físico e a aplicação de certa força muscular continuariam sendo requisitos para o funcionamento do controle remoto, da mesma forma que ocorre com os métodos convencionais. Diante desse problema, surgiu a ideia de criar um acionador baseado em proximidade.

Sensor de proximidade baseado em luz infravermelha. Em condições normais, o sensor de proximidade acende um LED de *status* quando algum objeto de aproxima do circuito. Já quando há luz solar no ambiente, a polaridade do sensor parece se inverter: o LED de *status* permanece aceso e, quando um objeto se aproxima, ele apaga. Isso acontece porque a luz do sol incide uma quantidade enorme de radiação infravermelha sobre o sensor, o que torna a superfície do objeto aproximado um bloqueador, e não mais um refletidor da luz do LED IR. Como este problema ainda não foi devidamente contornado, a solução adotada de imediato foi apenas utilizar o acionador em ambientes pouco afetados pela luz do sol.

Reconhecimento de voz com entrada pelo microfone. A API do PocketSphinx não possui função nativa para lidar com a entrada de áudio em tempo real pelo microfone. Esse fato, apesar de inusitado, é justificado na seção de perguntas frequentes da documentação oficial[3], a qual explica que, como o *software* é multiplataforma, ou seja, desenvolvido para ser executado em diferentes dispositivos e sistemas operacionais, seria impraticável disponibilizar uma função para acessar o microfone em vários ambientes distintos. A solução foi estudar o código do *engine* para *desktop* e adaptar a função que captura áudio em tempo real.

Contato com público alvo. Ao entrar em contato com as casas de apoio às PCD, houve certa resistência quanto à permissão para testar o protótipo com os pacientes, sob a justificativa de que vários outros grupos de pesquisa já haviam passado pelo local, realizado os testes e deixado poucos ou nenhum benefício para a instituição. Segundo a coordenação, o não retorno dos pesquisadores, favorecidos diretamente com as publicações geradas com ajuda da instituição, provocou a falta de interesse em auxiliar outros grupos da universidade. A solução encontrada foi desenvolver o protótipo com antecedência e levá-lo já finalizado, com o intuito de dar a impressão que o trabalho não seria apenas "mais uma promessa" das muitas que passaram pelo instituto sem deixar frutos; e testar com pessoas sem qualquer deficiência para aumentar a amostra populacional dos voluntários.

[3] <https://cmusphinx.github.io/wiki/faq/#q-how-to-record-sound-for-pocketsphinx-from-microphone-on-my-platform>

5.2 Trabalhos Futuros

Apesar de o sistema ter sido configurado para controlar apenas uma marca de TV, o projeto foi concebido com base no modelo de controle remoto universal. Portanto, é fundamental que um dos próximos passos seja expandir o controle para outros equipamentos dentro do ambiente doméstico. Dispositivos como televisores, ares-condicionados, DVD *players*, *home theaters*, aparelhos de som, dentre outros, possuem três pares de funções semelhantes (AHSAN et al., 2014), como mostrado na Tabela 9.

Tabela 9 – Funcionalidades comuns em dispositivos eletrônicos.

Dispositivo	**Ligar / Desligar Começar/Terminar**	**Aumentar/Diminuir Subir / Descer**	**Próximo/Anterior Mudar de estado**
Televisor	✓	✓	✓
Ar-condicionado	✓	✓	✓
DVD *player*	✓	✓	✓
Home theater	✓	✓	✓
Aparelho de som	✓	✓	✓

Fonte: Adaptado de Ahsan et al. (2014).

Além da acessibilidade, um segundo ponto chave no *design* do projeto é o baixo custo. Atualmente, um Arduino UNO é utilizado para mandar informação de controle para a TV. Isso acontece porque o acesso direto aos pinos de PWM do C.H.I.P. ainda não é bem documentado e não atende à velocidade exigida pela frequência dos protocolos de comunicação entre controles e aparelhos. Nesse sentido, a meta é estudar a fundo a documentação sobre o mapeamento dos pinos do microcomputador para tentar utilizá-lo como única plataforma embarcada do sistema, eliminando o atual uso do Arduino. A biblioteca `LIRC`[4], que faz parte do repositório oficial do Linux, poderia ser, então, utilizada no próprio C.H.I.P. para substituir a `IRremote` usada no Arduino. Alguns esforços da comunidade já começam a render frutos funcionais, embora bastante "estranhos", como a utilização dos pinos de saída de áudio do C.H.I.P. para gerar a PWM[5].

Atualmente, o controle remoto funciona somente com os módulos de reconhecimento de comandos de voz e de gestos da cabeça como métodos de entrada. Já o *feedback* ao usuário é dado por três LEDs de cores vermelha, amarela e verde. A fim de antender uma maior diversidade de problemas físicos, novos métodos de interação, tanto de entrada quanto de saída, devem ser incorporados ao sistema. A detecção do movimento dos braços (direita-esquerda, cima-baixo e vice-versa, por exemplo), poderia ser implementada também utilizando técnicas de computação visual através de uma *webcam*, como a subtração de fundo (do inglês *background*

[4] <http://www.lirc.org/>
[5] <https://bbs.nextthing.co/t/installing-lirc-on-c-h-i-p/2449>

subtraction) (DHANANJAYA; RAMAMURTHY; THIMMAIAH, 2015). Outra alternativa seria o "reconhecimento do pensamento", alcançado através da interface cérebro-computador (BCI, do inglês *brain-computer interface*) (LEE et al., 2013) por meio de aparelhos como o Emotiv EPOC (EMOTIV, 2017). Como interface de saída, um *speaker* poderia ser utilizado para prover informações ao longo do funcionamento do sistema via voz artificial, gerada através de um módulo de síntese de voz (TTS, do inglês *text-to-speech*) (TAYLOR, 2009). Dessa forma, espera-se que haja flexibilização quanto as deficiências atendidas pelo sistema, já que o número de limitações físicas abordado seria mais abrangente.

A respeito do acionamento externo, novos métodos de inicialização também devem ser explorados. Um sensor piezoelétrico, muito utilizado para medir níveis de pressão, poderia ser utilizado como base para a construção de um acionador baseado em sopro (que causa uma mudança brusca de pressão no ar), de forma similar à implementada no IntegraMouse, um *mouse* controlado pela boca (INTEGRAMOUSE, 2017). A segunda ideia seria acionar o sistema através da oclusão (movimento realizado ao cerrar os dentes no momento da mordida) (WOŹNIAK et al., 2015): dois eletrodos seriam postos nos músculos das têmporas ou no masseter, ambos acionados durante a mastigação, e um circuito de eletromiografia (EMG) seria utilizado para detectar a contração muscular provocada pela ação de morder. As duas soluções seriam úteis passa pessoas que possuem somente o movimento dos músculos da cabeça e, obviamente, o do aparelho respiratório preservado. Ambas já estão sendo desenvolvidas e devem gerar, em breve, trabalhos de conclusão de curso da graduação.

O consumo de energia do sensor de proximidade também deve ser avaliado, já que os componentes infravermelhos permanecem ligados durante todo o tempo. Assim, a utilização de microcontroladores simples e de baixo custo como o ATtiny10 da Microchip, que custa apenas U$ 0,33 dólares[6], por exemplo, deve ser investigada, já que um de seus pinos de PWM poderia ser útil na tarefa de economizar a energia drenada da bateria para acender o LED IR. Além disso, a PWM poderia ser utilizada também para codificar um protocolo simples de comunicação entre o LED IR e o sensor IR, de modo que a interferência causada pela radiação infravermelha espalhada pelo meio, especialmente aquela produzida pela luz do sol, seja eliminada ou pelo menos minimizada.

Através dos testes, foi percebido que o sistema apresenta dificuldades na etapa de detecção da face que, dependendo de fatores como iluminação, posição inicial da cabeça e distância entre o usuário e a câmera, não ocorre de maneira correta. Para tentar solucionar este problema, soluções alternativas ao algoritmo Viola-Jones precisam ser investigadas, sempre mantendo o foco no baixo custo computacional devido à utilização de um microcomputador embarcado. Além disso, a documentação da `OpenCV`, em especial a do método `detectMultiScale()`, deve ser estudada de forma mais rigorosa.

[6] <http://www.microchip.com/wwwproducts/en/ATtiny10>

De forma análoga, muitos voluntários sentiram dificuldades ao executar o movimento *roll*. Tal desempenho pode ser atribuído tanto à própria natureza do movimento, que normalmente se confunde com o *yaw*, quanto à técnica utilizada, a qual considera o modelo da cabeça como uma esfera livre no espaço, sem considerar o pescoço como base. Como a estimação da pose da cabeça — processo usado para realizar efetivamente o reconhecimento de gestos — compreende uma vasta série de técnicas distintas (MURPHY-CHUTORIAN; TRIVEDI, 2009), é importante investigar outros algoritmos alternativos ao Lucas-Kanade, implementado no método `calcOpticalFlowPyrLK()`, e observar até que ponto o microcomputador embarcado suporta o nível de processamento exigido por eles.

Conforme constatado durante a fase de desenvolvimento, a execução conjunta dos sistemas de reconhecimento de gestos e de fala requer mais processamento e recursos do que o C.H.I.P. suporta. Uma solução seria explorar a fundo os métodos da OpenCV para eliminar quaisquer etapas do processo de detecção facial e estimação da pose da cabeça cuja ausência não comprometa o estado funcional do sistema. Já em relação ao sistema de reconhecimento de voz, espera-se que, ao reamostrar os áudios para 8 kHz e utilizá-los no treino de um modelo acústico semi-contínuo, o tempo de processamento da fala diminua juntamente com a demanda de recursos da plataforma embarcada.

Segundo Costa et al. (2017), as soluções que realizam o reconhecimento dos movimentos da cabeça podem ser classificadas como i) dependentes do ambiente, as quais utilizam abordagens *video-based* e têm o processamento baseado em técnicas de computação visual; e ii) dependentes do usuário, quando uma tecnologia vestível é, normalmente, acoplada fisicamente à cabeça do usuário. A solução deste trabalho pertence ao primeiro grupo, porém as tarefas para a criação de um sistema que seja dependente do usuário já foram iniciadas: dois circuitos integrados MPU-6050, os quais contém sensores de aceleração e giro integrados, serão conectados ao microcontrolador presente em um *chip* baseado no módulo ESP8266 para indicar a mudança de pose da cabeça, substituindo, portanto, o uso da *webcam* e das técnicas de computação visual. É esperado que tal projeto tenha um custo ainda mais baixo do que o atual, já que o ESP8266 custa apenas U$ 6,95 dólares[7].

Sobre os testes, a dificuldade maior foi encontrar PCD que se adequassem ao perfil do público-alvo. A meta é formar parcerias com mais casas de apoio às PCD e encontrar mais voluntários com deficiências motoras dos membros superiores e limitações no uso de controles remotos convencionais, atendidos pelo módulo AGR; e visuais, sendo estes beneficiados pelo reconhecimento e síntese de voz. O *feedback* de usuários impactados de forma mais direta pela tecnologia assistiva, tanto em forma de críticas quanto em forma de sugestões e elogios, é essencial para a identificação de pontos fracos e posterior melhoria do sistema. Apesar da resistência enfrentada pelas instituições, não se pode deixar de tentar.

[7] <https://www.sparkfun.com/products/13678>,<https://www.adafruit.com/product/2491>

REFERÊNCIAS

ABLENET. *Candy Corn Proximity Sensor Switch*. 2017. Disponível em: <https://www.ablenetinc.com/candy-corn-proximity-sensor-switch>. Acesso em: 16 novembro 2017. Citado na página 21.

ADAPTIVATION. *Honeybee Proximity Switch*. 2017. Disponível em: <https://www.adaptivation.com/product-page/honeybee>. Acesso em: 16 novembro 2017. Citado na página 21.

AHSAN, K. et al. Unicon remote control model – a mobile system for assistive technology. *Research Journal of Recent Sciences*, v. 3, n. 4, p. 95–102, April 2014. Citado 2 vezes nas páginas 21 e 65.

ALI, A. et al. A complete kaldi recipe for building arabic speech recognition systems. In: *2014 IEEE Spoken Language Technology Workshop (SLT)*. [S.l.: s.n.], 2014. p. 525–529. Citado na página 31.

ALTHOFF, F. et al. Robust multimodal hand and head gesture recognition for controlling automotive infotainment systems. In: . [S.l.]: VDI, 2005. v. 1919, p. 187. Citado na página 21.

ARCOVERDE NETO, E. N. et al. Real-time head pose estimation for mobile devices. In: *Intelligent Data Engineering and Automated Learning - IDEAL*. [s.n.], 2012. p. 467–474. Disponível em: <https://doi.org/10.1007/978-3-642-32639-4_57>. Citado na página 25.

ARCOVERDE NETO, E. N. et al. Enhanced real-time head pose estimation system for mobile device. *Integrated Computer-Aided Engineering*, v. 21, n. 3, p. 281–293, 2014. Disponível em: <http://dx.doi.org/10.3233/ICA-140462>. Citado 3 vezes nas páginas 26, 44 e 60.

ARDUINO. *Arduino*. 2017. Disponível em: <https://www.arduino.cc/>. Citado na página 37.

ARDUINO TUTORIALS. *Arduino: PWM*. 2017. Disponível em: <https://www.arduino.cc/en/Tutorial/PWM>. Citado na página 32.

ASSISTIVA. *Assistiva: Categorias de Tecnologia Assistiva*. 2017. Disponível em: <http://www.assistiva.com.br/tassistiva.html#categorias>. Citado na página 17.

BARRENA, S. et al. Designing android applications with both online and offline voice control of household devices. In: *2012 38th Annual Northeast Bioengineering Conference (NEBEC)*. [S.l.: s.n.], 2012. p. 319–320. ISSN 2160-6986. Citado na página 21.

BATISTA, C. T.; CAMPOS, E. M.; SAMPAIO NETO, N. C. A proposal of a universal remote control system based on head movements. In: *Proceedings of the 16th Brazilian Symposium on Human Factors in Computing Systems*. Porto Alegre, Brazil: Brazilian Computer Society, 2017. (IHC'17). ISBN 978-85-7669-278-2. Citado na página 22.

BECERRA, A.; ROSA, J. I. de la; GONZÁLEZ, E. A case study of speech recognition in spanish: From conventional to deep approach. In: *2016 IEEE ANDESCON*. [S.l.: s.n.], 2016. p. 1–4. Citado na página 31.

BRASIL. Constituição da República Federativa do Brasil. Supremo Tribunal Federal, Brasília, DF, Brasil, 1988. Citado na página 18.

BRASIL. Decreto n° 3.298, de 20 de dezembro de 1999. *Regulamenta a Lei N° 7.853, de 24 de outubro de 1989, dispõe sobre a Política Nacional para a Integração da Pessoa Portadora de Deficiência, consolida as normas de proteção, e dá outras proviidências*, Diário Oficial, Brasília, DF, Brasil, dec. 1999. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2007-2010/2009/decreto/d6949.htm>. Citado 2 vezes nas páginas 18 e 19.

BRASIL. Decreto n° 6.949, de 25 de agosto de 2009. *Promulga a Convenção Internacional sobre os Direitos das Pessoas com Deficiência e seu Protocolo Facultativo, assinados em Nova York, em 30 de março de 2007*, Diário Oficial, Brasília, DF, Brasil, ago. 2009. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2007-2010/2009/decreto/d6949.htm>. Citado na página 18.

BRASIL. Tecnologia assistiva: Comitê de ajudas técnicas. Subsecretaria Nacional de Promoção dos Direitos da Pessoa com Deficiência, CORDE., Brasília, Brasil, p. 138p, 2009. Citado 2 vezes nas páginas 16 e 17.

BRASIL. Lei n° 13.146, de 06 de julho de 2015. *Institui a Lei Brasileira de Inclusão da Pessoa com Deficiência (Estatuto da Pessoa com Deficiência)*, Diário Oficial, Brasília, DF, Brasil, jul. 2015. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2015/lei/l13146.htm>. Citado na página 19.

CAVALLO, F. et al. On the design, development and experimentation of the astro assistive robot integrated in smart environments. In: *2013 IEEE International Conference on Robotics and Automation*. [S.l.: s.n.], 2013. p. 4310–4315. ISSN 1050-4729. Citado na página 61.

CHEN, G.; PARADA, C.; HEIGOLD, G. Small-footprint keyword spotting using deep neural networks. In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2014. p. 4087–4091. ISSN 1520-6149. Citado na página 48.

C.H.I.P. *C.H.I.P.: The Smarter Way to Build Smart Things*. 2017. Disponível em: <https://getchip.com/pages/chip>. Citado na página 35.

CMUSPHINX. *Building an application with PocketSphinx*. 2017. Disponível em: <https://cmusphinx.github.io/wiki/tutorialpocketsphinx/>. Citado na página 48.

CMUSPHINX. *CMUSphinx: Open Source Speech Recognition Toolkit*. 2017. Disponível em: <https://cmusphinx.github.io/>. Acesso em: 16 novembro 2017. Citado na página 38.

COSI, P. A kaldi-dnn-based asr system for italian. In: *2015 International Joint Conference on Neural Networks (IJCNN)*. [S.l.: s.n.], 2015. p. 1–5. ISSN 2161-4393. Citado na página 31.

COSTA, L. C. P. et al. Accessibility in digital television: Designing remote controls. In: *2012 IEEE International Conference on Consumer Electronics (ICCE)*. [S.l.: s.n.], 2012. p. 676–677. ISSN 2158-3994. Citado na página 19.

COSTA, V. K. da et al. Best practices for graphical user interfaces design with interaction based on head movements. In: *Proceedings of the 16th Brazilian Symposium on Human Factors in Computing Systems*. Porto Alegre, Brazil: Brazilian Computer Society, 2017. (IHC'17). ISBN 978-85-7669-278-2. Citado na página 67.

CUNHA, A. M. da; VELHO, L. *Métodos probabilísticos para Reconhecimento de Voz*. [S.l.], 2003. Citado na página 28.

DARRELL, T.; PENTLAND, A. Active gesture recognition using learned visual attention. *Advances in Neural Information Processing Systems*, MORGAN KAUFMANN PUBLISHERS, p. 858–864, 1996. Citado na página 16.

DAVIS, S.; MERLMESTEIN, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. on ASSP*, v. 28, p. 357–366, Aug. 1980. Citado na página 29.

DESHMUKH, N.; GANAPATHIRAJU, A.; PICONE, J. Hierarchical search for large-vocabulary conversational speech recognition. *IEEE Signal Processing Magazine*, p. 84–107, 1999. Citado na página 30.

DHANANJAYA, B.; RAMAMURTHY, B.; THIMMAIAH, P. Moving object tracking with opencv on arm cortex-a8 in surveillance applications. *International Journal of Current Engineering and Technology*, v. 5, n. 2, p. 843–848, April 2015. ISSN 2347-5161. Citado na página 66.

EMOTIV. *Emotiv EPOC+: 14 Channel Wireless EEG Headset*. 2017. Disponível em: <https://www.emotiv.com/epoc/>. Citado na página 66.

EPELDE, G. et al. Tv as a human interface for ambient intelligence environments. In: *2011 IEEE International Conference on Multimedia and Expo*. [S.l.: s.n.], 2011. p. 1–6. ISSN 1945-7871. Citado na página 20.

ERDEN, F.; CETIN, A. E. Hand gesture based remote control system using infrared sensors and a camera. *IEEE Transactions on Consumer Electronics*, v. 60, n. 4, p. 675–680, Nov 2014. ISSN 0098-3063. Citado na página 21.

ESQUEF, P. A. A.; CAETANO, R. Rocc – a webcam-mouse for people with disabilities. *Proc. XXV Simpósio Brasileiro de Telecomunicações*, 2007. Citado na página 20.

FALABRASIL. *FalaBrasil: Reconhecimento de Voz para o Português Brasileiro*. 2013. Disponível em: <http://www.laps.ufpa.br/falabrasil/>. Citado na página 38.

GAIDA, C. et al. *Comparing open-source speech recognition toolkits*. [S.l.], 2014. Project OASIS. Citado na página 31.

GELDERBLOM, G. J.; WITTE, L. P. de. The assessment of assistive technology: Outcomes, effects and costs. In: . [S.l.]: IOS Press, 2002. v. 14, n. 3, p. 91–94. Citado na página 16.

HUANG, X.; ACERO, A.; HON, H. *Spoken Language Processing*. [S.l.]: Prentice-Hall, 2001. Citado 5 vezes nas páginas 16, 28, 29, 30 e 31.

HUGGINS-DAINES, D. *PPocketSphinx API Documentation*. 2017. Disponível em: <https://cmusphinx.github.io/doc/pocketsphinx/>. Citado na página 48.

HUGGINS-DAINES, D. et al. Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices. In: *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. [S.l.: s.n.], 2006. v. 1, p. I–I. ISSN 1520-6149. Citado 2 vezes nas páginas 38 e 47.

IBGE. *Censo Demográfico*. Instituto Brasileiro de Geografia e Estatística, 2010. Disponível em: <www.ibge.gov.br/home/estatistica/populacao/censo2010/>. Citado na página 19.

INTEGRAMOUSE. *IntegraMouse: The Mouth Montrolled Mouse*. 2017. Disponível em: <https://www.integramouse.com/en/home/>. Citado na página 66.

ITU-T. *ITU-T Recommendation P.10 – Vocabulary for performance and quality of service*. [S.l.], 2006. Citado na página 39.

JELINEK, F. *Statistical methods for speech recognition*. [S.l.]: MIT Press, 1997. Citado na página 29.

JEVTIĆ, N.; KLAUTAU, A.; ORLITSKY, A. Estimated rank pruning and Java-based speech recognition. In: *Automatic Speech Recognition and Understanding Workshop*. [S.l.: s.n.], 2001. Citado na página 30.

JUANG, H.; RABINER, R. Hidden Markov models for speech recognition. *Technometrics*, v. 33, n. 3, p. 251–272, 1991. Citado na página 29.

JUNQUA, J.-C.; HATON, J.-P. *Robustness in Automatic Speech Recognition*. [S.l.]: Kluwer, 1996. Citado na página 29.

KARDARIS, N. et al. A platform for building new human-computer interface systems that support online automatic recognition of audio-gestural commands. In: *Proceedings of the 2016 ACM on Multimedia Conference*. New York, NY, USA: ACM, 2016. (MM'16), p. 1169–1173. ISBN 978-1-4503-3603-1. Disponível em: <http://doi.acm.org/10.1145/2964284.2973794>. Citado na página 61.

KUMAR, S.; KUMAR, S. Design and development of head motion controlled wheelchair. *International Journal of Advances in Engineering & Technology*, IAET Publishing Company, v. 8, n. 5, p. 816, 2015. Citado na página 21.

LADEFOGED, P. *A Course in Phonetics*. 4. ed. [S.l.]: Harcourt Brace, 2001. Citado na página 30.

LAYTON, J. *How Remote Controls Work*. 2005. Disponível em: <http://electronics.howstuffworks.com/remote-control.htm>. Citado na página 32.

LEE, A.; KAWAHARA, T.; SHIKANO, K. Gaussian mixture selection using context-independent HMM. *In Proceedings IEEE-ICASSP*, 2001. Citado na página 29.

LEE, G.; KIM, K.; KIM, J. Development of hands-free wheelchair device based on head movement and bio-signal for quadriplegic patients. *International Journal of Precision Engineering and Manufacturing*, v. 17, n. 3, p. 363–369, 2016. ISSN 2005-4602. Disponível em: <http://dx.doi.org/10.1007/s12541-016-0045-5>. Citado na página 21.

LEE, W. T. et al. A brain computer interface for smart home control. In: *2013 IEEE International Symposium on Consumer Electronics (ISCE)*. [S.l.: s.n.], 2013. p. 35–36. ISSN 0747-668X. Citado na página 66.

LEVY, P. C. et al. Activeiris: Uma solução para comunicação alternativa e autonomia de pessoas com deficiência motora severa. In: *Proceedings of the 12th Brazilian Symposium on Human Factors in Computing Systems*. Porto Alegre, Brazil: Brazilian Computer Society, 2013.

(IHC '13), p. 42–51. ISBN 978-85-7669-278-2. Disponível em: <http://dl.acm.org/citation.cfm?id=2577101.2577113>. Citado na página 21.

LIMA BARRETO, A. H. de. *Triste Fim de Policarpo Quaresma*. [S.l.]: Coleção L&PM POCKET, 2017. Citado na página 6.

LUCAS, B. D.; KANADE, T. An iterative image registration technique with an application to stereo vision. In: *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1981. (IJCAI'81), p. 674–679. Disponível em: <http://dl.acm.org/citation.cfm?id=1623264.1623280>. Citado 2 vezes nas páginas 28 e 45.

MARTÍNEZ-ZARZUELA, M. et al. Adaboost face detection on the gpu using haar-like features. In: ____. *New Challenges on Bioinspired Applications: 4th International Work-conference on the Interplay Between Natural and Artificial Computation, IWINAC*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. p. 333–342. ISBN 978-3-642-21326-7. Disponível em: <https://doi.org/10.1007/978-3-642-21326-7_36>. Citado na página 26.

MITRA, S.; ACHARYA, T. Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, v. 37, n. 3, p. 311–324, May 2007. ISSN 1094-6977. Citado na página 25.

MUNDO ESTRANHO. *Como funciona o controle remoto?* 2011. Disponível em: <http://mundoestranho.abril.com.br/materia/como-funciona-o-controle-remoto>. Citado na página 32.

MURPHY-CHUTORIAN, E.; TRIVEDI, M. M. Head pose estimation in computer vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, IEEE Computer Society, Washington, DC, USA, v. 31, n. 4, p. 607–626, abr. 2009. ISSN 0162-8828. Disponível em: <http://dx.doi.org/10.1109/TPAMI.2008.106>. Citado na página 67.

OPENCV. *OpenCV: Open Source Computer Vision Library*. 2017. Disponível em: <http://opencv.org/>. Acesso em: 9 abril 2017. Citado na página 38.

OU, Y. Y. et al. A happiness-oriented home care system for elderly daily living. In: *2014 International Conference on Orange Technologies*. [S.l.: s.n.], 2014. p. 193–196. Citado 2 vezes nas páginas 39 e 40.

PAJKANOVIĆ, A.; DOKIĆ, B. Wheelchair control by head motion. *Serbian Journal of electrical engineering*, v. 10, n. 1, p. 135–151, 2013. Citado na página 21.

PETAZZONI, T. *Device Tree for Dummies*. [S.l.: s.n.], 2014. Citado na página 36.

PICONE, J. Signal modeling techniques in speech recognition. *Proceedings of the IEEE*, v. 81, n. 9, p. 1215–47, Sep. 1993. Citado na página 29.

POKHARIYA, H. et al. Navigo–accessibility solutions for cerebral palsy affected. In: *2006 International Conference on Computational Inteligence for Modelling Control and Automation and International Conference on Intelligent Agents Web Technologies and International Commerce (CIMCA'06)*. [S.l.: s.n.], 2006. p. 143–143. Citado 2 vezes nas páginas 20 e 61.

POVEY, D. et al. The kaldi speech recognition toolkit. In: *In IEEE 2011 workshop*. [S.l.: s.n.], 2011. Citado na página 31.

RABINER, L. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, v. 77, n. 2, p. 257–86, Feb. 1989. Citado na página 29.

RABINER, L.; JUANG, B. *Fundamentals of Speech Recognition*. Englewood Cliffs, N.J.: PTR Prentice Hall, 1993. Citado na página 28.

SAKOE, H.; CHIBA, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. on ASSP*, v. 26, n. 1, p. 43–49, 1978. Citado na página 28.

SAMPAIO NETO, N. et al. Free tools and resources for Brazilian Portuguese speech recognition. *Journal of the Brazilian Computer Society*, v. 17, p. 53–68, 2011. Citado na página 29.

SAMSUNG ELECTRONICS. *S3F80KB Microcontroller for IR Remote Controller*. [S.l.], 2008. Disponível em: <http://elektrolab.wz.cz/katalog/samsung_protocol.pdf>. Citado 2 vezes nas páginas 33 e 34.

SANKAR, C.; MUNDKUR, N. Cerebral palsy–definition, classification, etiology and early diagnosis. *The Indian Journal of Pediatrics*, v. 72, n. 10, p. 865–868, Oct 2005. ISSN 0973-7693. Disponível em: <https://doi.org/10.1007/BF02731117>. Citado na página 61.

SIRAVENHA, A. et al. Uso de regras fonológicas com determinação de vogal tônica para conversão grafema-fone em Português Brasileiro. *7th International Information and Telecommunication Technologies Symposium*, 2008. Citado na página 47.

SOLANKI, U. V.; DESAI, N. H. Hand gesture based remote control for home appliances: Handmote. In: *2011 World Congress on Information and Communication Technologies*. [S.l.: s.n.], 2011. p. 419–423. Citado na página 21.

SRIVASTAVA, P.; CHATTERJEE, S.; THAKUR, R. A novel head gesture recognition based control for intelligent wheelchairs. *International Journal of Research in Electrical & Electronics Engineering*, v. 2, n. 2, p. 10–17, 2014. Citado na página 21.

TAYLOR, P. *Text-To-Speech Synthesis*. [S.l.]: Cambridge University Press, 2009. Citado na página 66.

TEIXEIRA, A. et al. Speech-centric multimodal interaction for easy-to-access online services – a personal life assistant for the elderly. *Procedia Computer Science*, v. 27, n. Supplement C, p. 389–397, 2014. ISSN 1877-0509. 5th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion, DSAI 2013. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1877050914000453>. Citado na página 61.

UN. Convention on the rights of persons with disabilities. *Treaty Series*, United Nations, v. 2515, December 2007. Disponível em: <http://www.un.org/disabilities/convention/conventionfull.shtml>. Citado 2 vezes nas páginas 16 e 18.

UNITED STATES. *Americans with Disabilities Act of 1990*. 1990. Public Law 101-336, 104 Stat. 327. Disponível em: <http://www.gpo.gov/>. Citado na página 17.

VIOLA, P.; JONES, M. Rapid object detection using a boosted cascade of simple features. In: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*. [S.l.: s.n.], 2001. v. 1, p. I–511–I–518 vol.1. ISSN 1063-6919. Citado 3 vezes nas páginas 26, 27 e 44.

VIOLA, P.; JONES, M. Robust real-time face detection. *International Journal of Computer Vision*, v. 57, n. 2, p. 137–154, 2004. ISSN 1573-1405. Disponível em: <http://dx.doi.org/10.1023/B:VISI.0000013087.49260.fb>. Citado 2 vezes nas páginas 26 e 44.

WECHSUNG, I.; NAUMANN, A. B. Evaluating a multimodal remote control: The interplay between user experience and usability. In: *Quality of Multimedia Experience, 2009. QoMEx 2009. International Workshop on*. [S.l.: s.n.], 2009. p. 19–22. Citado 2 vezes nas páginas 17 e 20.

WHO. *WHO Global Disability Action Plan 2014–2011: Better Health For All People With Disability*. Geneva, Switzerland: World Health Organization, 2015. Disponível em: <http://www.who.int/disabilities/actionplan/en/>. Citado na página 19.

WOODLAND, P.; POVEY, D. Large scale discriminative training of hidden Markov models for speech recognition. *Computer Speech and Language*, v. 16, p. 25–47, 2002. Citado na página 29.

WOŹNIAK, K. et al. Muscle fatigue in the temporal and masseter muscles in patients with temporomandibular dysfunction. *Biomed Research International*, Hindawi Publishing Corporation, v. 2015, 2015. ISSN 2314-6133. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4391653/>. Citado na página 66.