UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CAMPUS TOLEDO
CURSO DE TECNOLOGIA EM SISTEMAS PARA INTERNET

RAFAEL SOUZA GOMES

ANÁLISE DO RECONHECIMENTO AUTOMÁTICO DE FALA APLICADO AO ENSINO DE PRONÚNCIA DE LÍNGUA ESTRANGEIRA

TRABALHO DE CONCLUSÃO DE CURSO

TOLEDO
2021

RAFAEL SOUZA GOMES

ANÁLISE DO RECONHECIMENTO AUTOMÁTICO DE FALA APLICADO AO ENSINO DE PRONÚNCIA DE LÍNGUA ESTRANGEIRA

ANALYSIS OF AUTOMATIC SPEECH RECOGNITION APPLIED TO FOREIGN LANGUAGE PRONUNCIATION TEACHING

Trabalho de Conclusão de Curso apresentado ao Curso de Tecnologia em Sistemas para Internet da Universidade Tecnológica Federal do Paraná, como requisito parcial para a obtenção do título de Tecnólogo.

Orientador: Roberto Milton Scheffel
Universidade Tecnológica Federal do Paraná - UTFPR

TOLEDO
2021

RAFAEL SOUZA GOMES

ANÁLISE DO RECONHECIMENTO AUTOMÁTICO DE FALA APLICADO AO ENSINO DE PRONÚNCIA DE LÍNGUA ESTRANGEIRA

Trabalho de Conclusão de Curso apresentado ao Curso de Tecnologia em Sistemas para Internet da Universidade Tecnológica Federal do Paraná, como requisito parcial para a obtenção do título de Tecnólogo.

Data de aprovação: 01 de dezembro de 2021

Profa. Dra. Rosane Fátima Passarini
Doutora em Engenharia de Automação e Sistemas
Universidade Tecnológica Federal do Paraná - UTFPR

Prof. Me. Eduardo Pezutti Beletato dos Santos
Mestre em Ciências da Computação e Matemática
Universidade Tecnológica Federal do Paraná - UTFPR

Prof. Dr. Roberto Milton Scheffel
Doutor em Ciência da Computação
Universidade Tecnológica Federal do Paraná - UTFPR

TOLEDO
2021

RESUMO

GOMES, Rafael S.. Análise do Reconhecimento automático de Fala aplicado ao ensino de Pronúncia de língua estrangeira. 2021. 28 f. Trabalho de Conclusão de Curso – Curso de Tecnologia em Sistemas para Internet, Universidade Tecnológica Federal do Paraná. Toledo, 2021.

Um dos desafios ao se aprender um novo idioma é a pronúncia. Esse trabalho visa validar o uso da tecnologia de reconhecimento automático de fala aplicada ao ensino de pronúncia. Segundo as pesquisas apresentadas, não há um consenso entre os pesquisadores sobre o quão eficiente a tecnologia pode ser. Diante disso, o trabalho tem como objetivo primeiramente analisar e comparar a taxa de assertividade das API: Google Speech to Text, Rev.Ai e Web Speech Api. A segunda parte do trabalho consiste na realização de um teste prático com alunos de inglês através de um protótipo de um aplicativo de ensino de pronúncia que utiliza a Web Speech Api para transcrições automáticas. Dessa forma espera-se que os resultados do presente trabalho possam contribuir para futuros projetos didáticos que utilizem o reconhecimento automático de fala.

Palavras-chave: Reconhecimento Automático de fala. Pronúncia. Ensino Lingua estrangeira.

ABSTRACT

GOMES, Rafael S.. Analysis of Automatic Speech Recognition Applied to Foreign Language Pronunciation Teaching. 2021. 28 f. Trabalho de Conclusão de Curso – Curso de Tecnologia em Sistemas para Internet, Universidade Tecnológica Federal do Paraná. Toledo, 2021.

One of the challenges when learning a new language is pronunciation. This final paper aims to validate the use of automatic speech recognition technology applied to teaching pronunciation. According to the research presented, there is no consensus among researchers about how efficient the technology can be. Therefore, the work aims primarily to analyze and compare the assertiveness rate of the APIs: Google Speech to Text, Rev.Ai, and Web Speech API. The second part of the work consists of carrying out a practical test with English students through a prototype of a pronunciation teaching application that uses the Web Speech API for automatic transcriptions. Thus, it is expected that the results of this work can contribute to future educational projects that use automatic speech recognition.

Keywords: Automatic speech recognition. Pronunciation. Foreign Language Teaching.

LISTA DE FIGURAS

Figura 1 – O processo de Reconhencimento Automático de Fala 12
Figura 2 – Representacao Genérica do Funcionamento 13
Figura 3 – O protótipo 19
Figura 4 – Suporte a Video Chamada 20

LISTA DE QUADROS

Quadro 1 – Grupo 1 : Frases de Pronunciação Simples 21
Quadro 2 – Grupo 2: Frases de Pronunciação Complexa 21

LISTA DE TABELAS

Tabela 1 – Comparativo das Ferramentas 22

LISTA DE ABREVIATURAS E SIGLAS

API	Application Programming Interface
DECOM	Departamento de Computação
HMM	Hidden Markov Model
web	World Wide Web
MP3	MPEG Layer 3
AAC	Advanced Audio Coding
Kbps	Kilobits per Second
FLAC	Free Lossless Audio Codec
ALAC	Apple Lossless Audio Codec
Hz	Hertz

SUMÁRIO

1 – INTRODUÇÃO 10
1.1 Objetivo Geral 10
1.2 Objetivos Específicos 11

2 – REVISÃO DE LITERATURA 12
2.1 A Pronúncia em Língua Estrangeira 12
2.2 Reconhecimento Automático de Fala 12
2.3 API de Reconhecimento Automático de Fala 14
2.3.1 Google Speech to Text 14
2.3.2 Web Speech API 15
2.3.3 REV.AI 15
2.4 Aplicação didática do reconhecimento automático de fala 16

3 – METODOLOGIA 18
3.1 Comparação entre as ferramentas Google Speech To text, Rev.AI e Web Speech API 18
3.2 Aplicação de testes com voluntários para a validação da Web Speech API . . 19
3.2.1 Sobre o protótipo 19
3.2.2 O Experimento 20

4 – ANÁLISE E DISCUSSÃO DOS RESULTADOS 22
4.1 Análise Comparativa entre as ferramentas Web Speech API, Google Speech to Text e REV.AI 22
4.2 Resultados da aplicação prática da Web Speech API com os voluntários . . . 22
4.2.1 Primeira etapa: Transcrição das frases do grupo 1. 22
4.2.2 Segunda Etapa: Transcrição das frases do grupo 2 23
4.2.3 Terceira Etapa: Conversação livre 23
4.2.4 Problemas Encontrados 23
4.2.5 Trabalhos futuros 24

5 – CONCLUSÃO 25

Referências 27

1 INTRODUÇÃO

Avanços tecnológicos têm proporcionado a criação de ferramentas didáticas e métodos inovadores que facilitam e potencializam o aprendizado em diferentes áreas. No ensino de línguas estrangeiras, plataformas de ensino como o Duolingo (DUOLINGO, 2021) e o Babbel (BABBEL, 2021) vem chamando a atenção dos estudantes e ganhando cada vez mais relevância ao utilizar métodos não tradicionais de ensino e envolvendo recursos tecnológicos no processo de aprendizagem (DUOLINGORESEARCH, 2021).

Existem inúmeros desafios no ensino de língua estrangeira. Segundo Gilakjani (2011) se observa que a pronúncia é negligenciada com muita frequência. Consequentemente há um grande número de estudantes que não conseguem ter êxito na comunicação. O uso da tecnologia de reconhecimento automático de fala pode permitir ao estudante a possibilidade de praticar a sua pronúncia em língua estrangeira apenas com o auxílio de alguma aplicação que disponibilize essa funcionalidade. Dessa forma a ferramenta seria capaz de retornar um *feedback* bem preciso do desempenho do estudante.

Alguns pesquisadores são céticos sobre o uso dessa tecnologia no aprendizado de língua estrangeira devido ao alto grau de variação dos idiomas, fator que dificulta o reconhecimento de fala quando o usuário não é fluente (CONIAM, 1999; DERWING, 2000). Entretanto as abordagens mais recentes no reconhecimento automático de fala melhoraram a precisão da transcrição, diminuindo significativamente a taxa de erro com pessoas não fluentes (MCCROCKLIN, 2019).

Este trabalho apresenta uma breve revisão literária abordando outros estudos e conceitos importantes sobre o reconhecimento de fala automático, a dificuldade na pronúncia de língua estrangeira, caminhos para se aplicar essa tecnologia de maneira eficaz no ensino e o quanto eficiente ela pode ser na transcrição de não fluentes na prática.

Com o propósito de potencializar o ensino de língua estrangeira e com foco no aprimoramento da pronúncia, será realizado um experimento para comparar a taxa de assertividade de APIS de reconhecimento automático de fala em uma base de dados pública que disponibiliza áudios em inglês com sua transcrição. Será desenvolvido um protótipo de uma aplicação focada no ensino de pronúncia, implementando a Web Speech Api (LIBBY, 2020). Depois, através do protótipo desenvolvido para esse trabalho, será realizado um teste do serviço de transcrição automática Web Speech API com alunos brasileiros de inglês.

1.1 Objetivo Geral

Esse trabalho tem como objetivo avaliar o funcionamento da tecnologia de reconhecimento automático de fala no ensino da pronúncia do inglês para alunos brasileiros.

1.2 Objetivos Específicos

- Identificar meios efetivos para se aplicar reconhecimento automático de fala no ensino de pronúncia de língua estrangeira, através de pesquisa bibliográfica.
- Comparar os recursos oferecidos e a assertividade da transcrição de algumas API de reconhecimento automático de fala.
- Desenvolver um protótipo de uma aplicação para prática da pronuncia em inglês utilizando uma API para a transcrição automática.
- Validar o uso da API escolhida aplicado ao aprendizado de pronúncia em inglês com voluntários através do protótipo.

2 REVISÃO DE LITERATURA

2.1 A Pronúncia em Língua Estrangeira

Uma comunicação eficiente em uma língua estrangeira depende de uma boa pronunciação. Entretanto, obter o domínio da pronúncia é uma tarefa difícil pois o aluno tende naturalmente a aplicar o padrão de sons de sua língua materna no idioma que está aprendendo, o que causa sotaques estrangeiros (AVERY, 1992).

Os modelos tradicionais de ensino precisam ser aprimorados. Segundo Gilakjani (2011), mesmo em grades curriculares em que a pronunciação não é negligenciada, o tradicional modelo de reprodução de sons não é o suficiente.

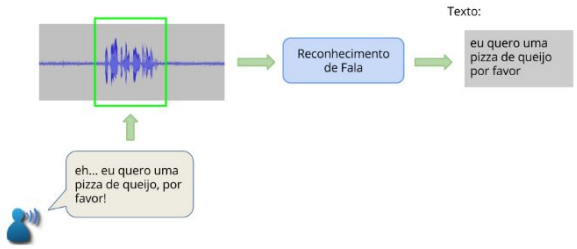
Apesar do intenso avanço tecnológico nas últimas décadas, não há muitos estudos recentes com informações confiáveis sobre as metodologias que funcionam e como técnicas mais inovadoras, como o reconhecimento automático de fala, podem ser adaptadas com eficiência no ensino da pronúncia. É importante aumentar a quantidade de pesquisas acadêmicas sobre esses tópicos (GILAKJANI, 2011).

2.2 Reconhecimento Automático de Fala

O reconhecimento automático de fala é uma área que busca soluções para a decodificação e transcrição de fala. Os sistemas de última geração se baseiam em conceitos multidisciplinares. Linguística, ciência da computação, processamento de sinais, acústica, teoria da comunicação, estatística, fisiologia e psicologia são as principais áreas de estudo (LEVIS, 2012).

Inicialmente é fundamental compreender a distinção entre o reconhecimento de fala e o reconhecimento de voz. Enquanto o reconhecimento de voz é capacidade de um sistema em reconhecer quem disse algo, o reconhecimento de fala diz respeito à habilidade de um sistema em reconhecer as palavras que são ditas em um discurso (LEVIS, 2012).

Figura 1 – O processo de Reconhencimento Automático de Fala



Fonte: Speechweb (2020)

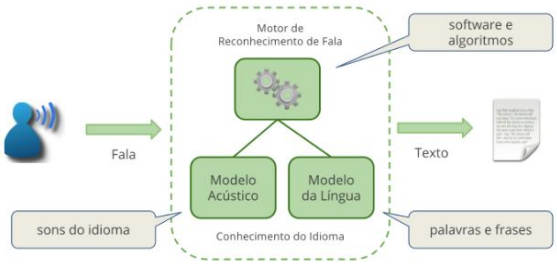
O reconhecimento automático de fala pode ser definido como um campo da ciência da

computação que desenvolve tecnologias e métodos que possibilitem a identificação de padrões no discurso. Geralmente recebe a entrada acústica do som por meio de um microfone para realizar a análise utilizando algum padrão, modelo ou algoritmo. A partir desse processo é produzida uma saída, que normalmente é a transcrição na forma de texto (LEVIS, 2012).

O processo do reconhecimento está exemplificado na Figura 1 . A frase pronunciada pelo usuário é captada pelo microfone como um sinal de áudio. Esse sinal tem as suas ondas de pressão acústicas processadas pelo motor de reconhecimento do sistema. A partir desse ponto, existem muitas abordagens de tratamento dos dados para a geração da transcrição final (SPEECHWEB, 2020).

Na Figura 2 é possível observar uma representação genérica do funcionamento geral do reconhecimento de fala. O Motor de reconhecimento representa os algoritmos necessários para realizar o processamento do áudio antes que o processo de análise seja iniciado. Como por exemplo, soluções para remoção de ruído, detecção de início e fim de fala (SPEECHWEB, 2020).

Figura 2 – Representacao Genérica do Funcionamento



Fonte: Speechweb (2020)

O Modelo acústico (Figura 2) se caracteriza como uma representação lógica de um conjunto de dados de áudios. É um ponto muito relevante para análise pois é a representação lógica dos padrões das palavras, frases e fonemas do idioma. Esses modelos são gerados a partir de um grande volume de fala gravadas com várias pessoas diferentes. Um desafio durante esse processo é o fator do sinal variar conforme cada falante. Existe um grande grau de variabilidade na pronúncia de pessoa para pessoa (SPEECHWEB, 2020).

O modelo da língua (Figura 2) define como um conjunto de palavras podem ser combinadas para formar as frases do idioma de maneira coerente. Podendo ser expresso por uma gramática composta por regras escritas manualmente, ou por modelos de fala gerados a partir de um grande volume de texto sendo aplicado no reconhecimento de fala de maneira espontânea (SPEECHWEB, 2020). Os métodos mais modernos de transcrição costumam utilizar uma abordagem mista, ou seja, é levado em consideração o modelo acústico e o modelo de língua (VOXFORGE, 2021).

Um dos desafios durante o processo de transcrição é a variabilidade dos idiomas. Em um mesmo país podemos encontrar um número considerável de sotaques para o mesmo idioma,

o que pode causar erros de transcrição. Esse problema se torna ainda mais crítico quando é necessário reconhecer a fala de pessoas que não são fluentes no idioma. A limitação ocorre porque a maioria dos sistemas de reconhecimento automático de fala foram projetados para funcionar através de modelos de dados limitados aos de padrões de fala de nativos (LEVIS, 2012).

Os primeiros sistemas de reconhecimento automático, por volta de 1950, eram bem simplificados, pois eram baseados em um pequeno número de padrões acústicos pré-armazenados. Esse método se limitava a obter resultados assertivos apenas em sons foneticamente diferentes. Eficaz apenas em reconhecer vocabulários menores, quando pronunciadas de maneira muito clara por alguém já treinado para isso (LEVIS, 2012).

Atualmente com o poder computacional disponível por meio de processadores multi-core, unidades de processamento gráfico de uso geral e clusters de CPU / GPU vem tornando possível o aprimoramento de modelos treinados mais poderosos. Esses modelos reduziram as taxas de erro dos sistemas de reconhecimento automático de fala (YU, 2015).

Segundo Wang (2019) o melhor desempenho de reconhecimento de fala ainda vem do modelo baseado em HMM (Hidden Markov Model) em combinação com técnicas de *Deep Learning*. Tem sido amplamente utilizado e a maioria dos sistemas implantados industrialmente são baseados em HMM.

2.3 API de Reconhecimento Automático de Fala

2.3.1 Google Speech to Text

A API de reconhecimento automático de fala do Google, chamada Google Speech to Text, é um serviço pago que faz parte do pacote de serviços Google Cloud Plataform. Permite a transcrição da entrada de áudio reconhecendo mais de 110 idiomas e suas variantes (ANGGRAINI, 2018).

O recurso *Streaming* de reconhecimento de fala permite receber resultados do reconhecimento de fala em tempo real. Permitindo configurar a entrada de áudio, sendo a partir do microfone ou através de arquivos de áudios (GOOGLE, 2021).

O serviço também disponibiliza a taxa com índice de confiança de reconhecimento, customização para reconhecimento de termos ou assuntos específicos, e um serviço de pontuação automática que está em aprimoramento (ANGGRAINI, 2018).

A ferramenta de customização fornece um recurso de personalização do vocabulário com o objetivo de melhorar o desempenho na transcrição de termos específicos e palavras raras (GOOGLE, 2021).

A Google Speech to Text se destaca ao permitir a escolha de uma variedade de modelos treinados específicos para cada situação com o intuito de otimizar a transcrição. As opções de recursos personalizados são: *command_and_search, video, phone_call, medical_dictation, medical_conversation*. O modelo *command_and_search* é aconselhado para transcrições de aúdios

curtos que são utilizados para comando de voz (GOOGLE, 2021).

O modelo *video* é aconselhado quando há vários interlocutores para realizar a transcrição simultaneamente. Para se obter êxito nesse modo é preciso que o áudio utilizado seja gravado com um microfone de alta qualidade, esse modelo tem suporte eficiente para ruído de fundo. A documentação aconselha que o áudio seja gravado a uma taxa de amostragem de 16.000 Hz ou mais (GOOGLE, 2021).

Um destaque é o modelo *phone_call* que é utilizado para transcrever áudio de uma chamada telefônica de maneira adequada levando em consideração características técnicas específicas que uma ligação costuma ter (GOOGLE, 2021).

O modelo *medical_dictation* garante um bom nível de assertividade para transcrever discursos de profissionais da saúde, enquanto o modelo *medical_conversation* é o ideal para transcrever conversas entre o paciente e o médico (GOOGLE, 2021).

2.3.2 Web Speech API

A API Web Speech é uma tecnologia que está sendo desenvolvida pelo W3C Speech API Community Group para realizar análise e síntese de fala. O reconhecimento de fala é realizado através de um serviço de transcrição *web*, do qual o desenvolvedor não tem acesso diretamente apesar do serviço possuir uma licença livre, o que torna a conexão com a internet obrigatória para utilizar o serviço (LIBBY, 2020).

API oferece suporte a vários idiomas. Porém, a ferramenta ainda apresenta a necessidade de se especificar explicitamente o idioma a ser reconhecido. A ferramenta ainda não é capaz de detectar o idioma automaticamente. Caso não seja explicitada a língua a ser detectada, por padrão o idioma será definido pelas configurações locais do usuário. Não é possível configurar mais do que um idioma para a aplicação simultaneamente (LIBBY, 2020).

Ao realizar o processo de transcrição, os dados são retornados em uma lista de frases candidatas, em que cada frase está vinculada a um valor que representa o índice de confiança. O item com o maior índice de confiança é listado primeiro (ADORF, 2013).

API é desenvolvida totalmente em JavaScript, considerada uma das linguagens de *script* mais usadas atualmente. É baseada em eventos, permitindo o funcionamento do processamento de fala de maneira assíncrona (ADORF, 2013).

Os eventos também são utilizados para retornar resultados instantâneos de reconhecimento de fala. Dessa forma, a ferramenta é capaz de disponibilizar um *feedback* quase imediato ao usuário. A API Web Speech abrange não só a análise de fala, mas também a síntese de fala. Ou seja, a ferramenta também realiza a conversão de texto em áudio (ADORF, 2013).

2.3.3 REV.AI

A API Rev.ai utiliza o aprendizado de máquina avançado para o reconhecimento automático de fala rápido e com precisão. Oferece recursos de transcrição para mídia gravada e *streaming* (REV, 2021).

Rev.ai oferece suporte para uma boa parte dos formatos de mídia de áudio. A documentação aconselha o uso de formatos que garantem o mínimo de perdas possíveis como o FLAC ou ALAC para se obter um bom desempenho na transcrição. Caso o usuário opte por utilizar algum formato com perdas, como por exemplo, MP3 ou AAC, é necessário que o arquivo possua uma taxa de bits de 192 Kbps ou superior (REV, 2021).

A ferramenta possibilita a configuração de vocabulário personalizado como meio para melhorar a precisão do *software*. É aconselhado o uso desse recurso ao usar palavras ou termos que não são comuns no idioma. Dessa forma, o vocabulário personalizado deve constar nomes próprios, termos técnicos especiais e palavras incomuns no geral (REV, 2021).

Rev.ai oferece o pontuação automática e realiza a normalização do formato padrão de datas, horários e números de telefone na transcrição final. Por exemplo, ao reconhecer como número de telefone o áudio em inglês: "*one two three one two three one two three four*", a ferramenta é capaz de formatar o número na transcrição final: "(123) 123-1234".

Há uma interessante funcionalidade capaz de desconsiderar disfluências durante o discurso. As disfluências são repetições de palavras, prolongamento incomuns de determinados sons, pausas preenchidas e pausas silenciosas (REV, 2021).

2.4 Aplicação didática do reconhecimento automático de fala

Não há um consenso sobre o quão eficaz a tecnologia de reconhecimento automático de fala pode ser no aprendizado de línguas estrangeiras. Cucchiarini (2006) realizou um experimento com um grupo de imigrantes que estavam aprendendo holandês e utilizou um sistema de reconhecimento de fala desenvolvido para a pesquisa como suporte de ensino e obteve um resultado positivo. O grupo que usou a tecnologia com *feedback* melhorou significantemente a pronúncia em sons considerados difíceis para a sua língua materna.

Por outro lado, McCrocklin (2019) em seu experimento dividiu dois grupos de estudantes de língua estrangeira, um estudou apenas com o auxílio de uma plataforma para treino de pronúncia e o outro estudou com base em um modelo tradicional para o ensino de pronúncia. Os resultados não mostraram diferenças significativas de desempenho entre os dois grupos, pois ambos melhoraram a pronúncia.

Estudos conduzidos por Derwing (2000) e Coniam (1999) não consideraram as ferramentas de reconhecimento automático de fala maduras o suficiente para o ensino de língua estrangeira. Os resultados mostraram que a precisão na transcrição com usuários não nativos no idioma reconhecido era muito baixa comparada com a transcrição de usuários nativos, que teve uma porcentagem de 90% de acerto.

As ferramentas de reconhecimento automático de fala mais recentes ainda apresentam problemas ao transcrever frases complexas. Contudo existem boas condições para serem utilizadas em soluções que são capazes de lidar com fala não nativa, e também são capazes de atender a critérios pedagógicos (MCCROCKLIN, 2019).

O estudo conduzido por Neri (2003) define um padrão com 5 aspectos que devem ser

seguidos para se obter um maior êxito no ensino de pronúncia em sistemas computacionais.

O primeiro e mais importante aspecto é o sistema possuir a funcionalidade de reconhecimento automático de fala para transcrever a pronúncia do aluno durante o uso da plataforma. É fundamental que seja garantido uma boa precisão na transcrição com alunos não fluentes (NERI, 2003).

O segundo aspecto é o sistema ser capaz de calcular uma pontuação da qualidade da pronúncia do usuário. O ideal é que o estudante de língua estrangeira consiga medir o seu desempenho apenas com o auxílio do software (NERI, 2003).

A análise da pronúncia pode ser realizada a partir de propriedades temporais, como velocidade da fala, e propriedades acústicas, como o som de cada fonema pronunciado. Quanto mais próximo a pronúncia do aluno for semelhante a de modelos nativos de referência, maior será a pontuação (NERI, 2003).

O terceiro e o quarto aspecto são referentes a detecção específica do erro na pronúncia. O sistema deve ser capaz de indicar em quais palavras houve erro de pronúncia e sugerir como melhorá-la. O autor sugere a utilização de modelo de dados com os erros mais comuns cometidos pelos alunos. E o último aspecto é a apresentação de um *feedback* completo que forneça detalhes do desempenho do aluno levando em consideração as fases anteriores (NERI, 2003).

A plataforma de ensino de idiomas Duolingo tem tentado aplicar esses conceitos no ensino da pronúncia em sua seção de conversação. Em 2014, o sistema de reconhecimento de fala obteve significativas melhorias. A cada módulo existem uma série de atividades de pronúncia, onde o usuário consegue ter um *feedback* de quais palavras pronunciou corretamente e quais ele precisa praticar mais (AHN, 2014).

Alguns críticos apontam que a seção de conversação para fixação da pronúncia do Duolingo ainda não está madura o suficiente e apresenta muitos problemas. Como por exemplo, *feedback* da pronúncia simplório, erros no reconhecimento de fala e o conteúdo para a prática de frases descontextualizadas não úteis para o dia-a-dia (LOTHERINGTON, 2016; BAJOREK, 2017).

3 METODOLOGIA

O método de pesquisa realizado é o exploratório, pois através do estudo acerca da tecnologia de reconhecimento automático de fala existente que foi abordado na revisão de literatura, foram encontradas duas abordagens para a análise da tecnologia em pontos cruciais para validar o seu uso aplicado ao ensino de pronúncia de língua estrangeira.

A primeira abordagem consiste em um experimento comparativo entre as ferramentas Google Speech to Text, Rev.AI e Web Speech API. A segunda abordagem é a aplicação de testes com voluntários em um protótipo de aplicação *web* para ensino de pronúncia com o recurso de reconhecimento automático de fala utilizando a ferramenta Web Speech API.

3.1 Comparação entre as ferramentas Google Speech To text, Rev.AI e Web Speech API

Foi realizado um experimento com o objetivo de comparar quantitativamente a eficiência da transcrição das seguintes API: Google Speech to Text, Rev.AI e Web Speech API.

O experimento consiste no uso de cada API para a transcrição automática de 300 áudios contendo frases em inglês, que juntos têm a duração total de 35 minutos. Devido as ferramentas Google Speech to Text e Rev.AI possuírem um limite de uso no plano gratuito, não foi possível realizar um teste de maior duração.

Os áudios foram extraídos da base de dados pública VoxForge (2021). Esse banco de dados contém um grande número de áudios com frases em inglês com as suas respectivas transcrições. Os áudios foram gravados por voluntários ao redor do mundo e todo o conteúdo está totalmente disponível gratuitamente.

Foram definidos 300 áudios aleatoriamente da base de dados VoxForge para a realização do teste. Durante o experimento, os áudios foram executados sequencialmente e foi realizada a transcrição de cada um. Depois foi feita a comparação da transcrição realizada com a transcrição original do áudio que é fornecida pela base de dados. Esse processo foi repetido para cada uma das API.

Foi necessário implementar um sistema para a aplicar o teste de maneira automatizada. O sistema executa todos os áudios em um navegador de maneira automatizada, efetua a transcrição de cada áudio e depois faz a comparação da transcrição realizada com a transcrição original. Esse processo foi repetido apenas substituindo a implementação da API de reconhecimento automático de fala para cada ferramenta analisada: Google Speech To Text, Rev.io e Web Speech API.

A taxa de acerto da transcrição foi calculada através da divisão do número de palavras reconhecidas corretamente pelo número de palavras existentes na transcrição original multiplicado por 100. Por exemplo, em um áudio onde a transcrição indica que há 10 palavras,

mas só foram transcritas 8 palavras corretamente pela API, então a taxa de acerto nesse caso será de 80%. O resultado final de eficiência da API é a média simples da taxa de acerto de cada um dos 300 áudios transcritos no teste.

Também serão coletados dados qualitativos a respeito das ferramentas, através de testes e da documentação para verificar quais recursos cada API oferece. Com o intuito de identificar recursos que poderiam ser utilizados no ensino de pronúncia de língua estrangeira, levando em consideração os aspectos essenciais que um sistema computacional de ensino de idiomas deveria possuir, como descrito na revisão de literatura.

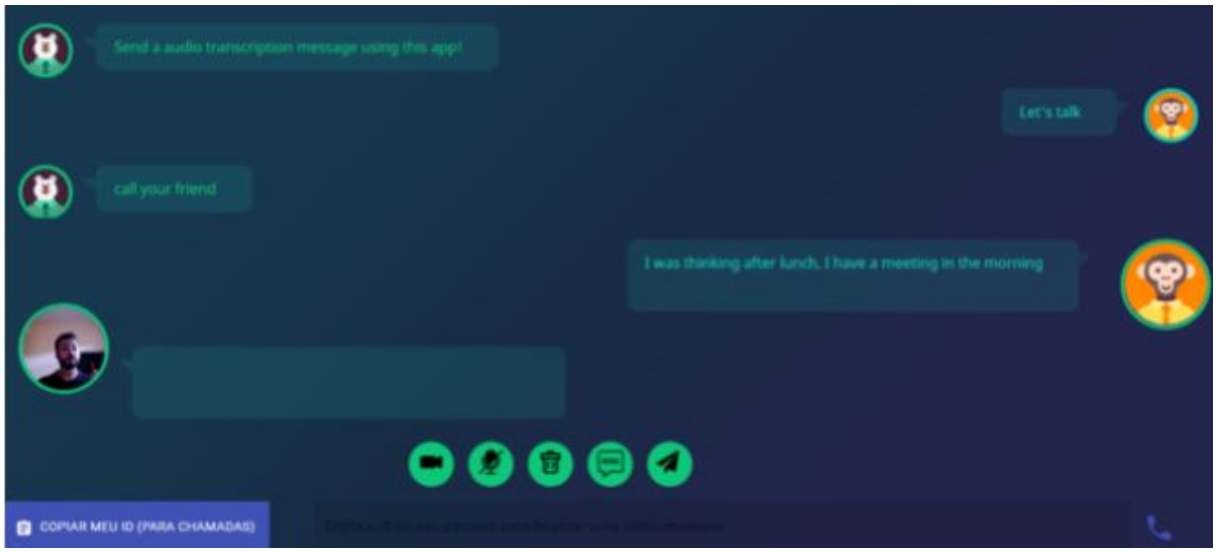
3.2 Aplicação de testes com voluntários para a validação da Web Speech API

Na segunda etapa foi realizado um experimento com 5 voluntários estudantes de inglês brasileiros, utilizando na prática o reconhecimento automático de fala da Web Speech API aplicada ao ensino de pronúncia do inglês. Os voluntários afirmaram estar no nível intermediário de pronúncia e compreenção do idioma. A ferramenta foi selecionada para o teste pelo seu uso ser gratuito e ter tido uma boa pontuação na taxa de acerto por palavra na etapa anterior, ligeiramente menor do que as outras ferramentas avaliadas.

3.2.1 Sobre o protótipo

Foi desenvolvido um protótipo (Figura 3) de uma aplicação *web* de ensino de pronúncia em inglês para a realização do experimento. O protótipo foi projetado apenas para atender as necessidades do experimento com os voluntários.

Figura 3 – O protótipo

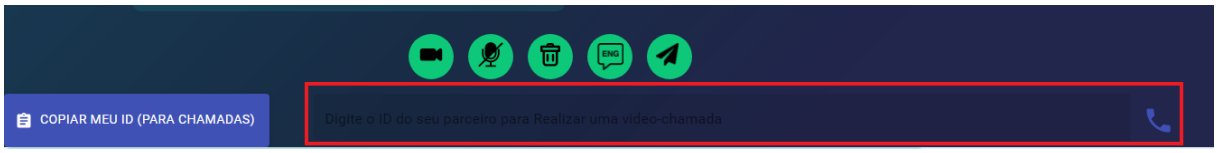


Fonte: Autor

O protótipo consiste em uma página que é uma sala de bate papo com chamada de vídeo disponível e com a transcrição automática, implementada pela Web Speech API, do

conteúdo reconhecido em inglês pelo microfone.

Figura 4 – Suporte a Video Chamada



Fonte: Autor

Ao entrar no *link*, o sistema gera automaticamente um número de id, que é a chave utilizada para iniciar uma chamada. Ao inserir id do usuário no campo em destaque na Figura 4 é possível ligar para o usuário e estabelecer uma vídeo chamada.

Ao clicar no botão com o ícone de microfone, a transcrição é habilitada. Enquanto o usuário estiver falando, será reconhecido a sua fala em tempo real e o texto será apresentado no balão.

O botão com o ícone de avião envia a transcrição para o *chat* que contém apenas as transcrições realizadas. A partir do chat, o avaliador poderá analisar se a transcrição realizada foi correta levando em consideração a pronúncia do voluntário.

O protótipo foi desenvolvido na linguagem de programação javascript, com o uso do framework Node.Js. Foi dividido em duas camadas: frontend, a camada visual em que é possível interagir com a aplicação, e o backend, a camada responsável em gerenciar as trocas de mensagem do chat da aplicação.

O arquivo server.js, localizado na camada do backend, é responsável pela comunicação do *chat* através da implementação da biblioteca Socket.io. O Socket.io possibilita comunicações bidirecionais entre cliente e servidor, o que permite a troca de mensagens entre os usuários.

O React, uma biblioteca usada para a criação de interfaces, foi utilizada para o desenvolvimento da camada do frontend. As estilizações da página foram criadas utilizando css. A ferramenta de transcrição Web Speech API foi implementada utilizando a biblioteca React Speech Recognition.

3.2.2 O Experimento

Com o auxílio de alguns professores de inglês, foram definidas 14 frases (Quadro 1 e Quadro 2) em inglês divididas em dois grupos. As frases foram classificadas levando em consideração o grau de dificuldade pronúncia para brasileiros, pois como visto na revisão de literatura, alguns fonemas são mais desafiadores em algumas regiões devido a estrutura da língua materna. Um grupo com frases de fácil pronúncia e o outro com frases com sons mais complexos.

O experimento foi individual e teve três etapas. Na primeira etapa foi solicitado ao aluno que tentasse pronunciar as frases, sem nenhuma instrução acerca da pronúncia na primeira

tentativa. O sistema exibe a transcrição pelo *chat*, e o avaliador analisava se a pronúncia do aluno estava correta e marcava o que o sistema havia transcrito.

Quadro 1 – Grupo 1 : Frases de Pronunciação Simples

Nº	**Frase**
1.1	Do you want some water?
1.2	How is it going?
1.3	the part when I break free
1.4	Where are you from?
1.5	I'm reading a book
1.6	Waiter, can you bring me the menu?
1.7	Excuse me, how much is this?

Fonte: Autor

Quadro 2 – Grupo 2: Frases de Pronunciação Complexa

Nº	**Frase**
2.1	Therefore, although all countries are vigorously studying.
2.2	Everything starts somewhere, although many physicists disagree
2.3	The bartender made me a Pink Squirrel.
2.4	Hierarchy and classification are two notions always very close
2.5	Eighth, both parties and tribunal can appoint experts.
2.6	Clothes have nothing to do with chemistry
2.7	Explore a detailed open world environment.

Fonte: Autor

Na segunda etapa do experimento, o voluntário deveria tentar pronunciar mais uma vez as frases que havia errado a pronúncia. Mas dessa vez foi mostrado para o aluno uma gravação com a pronúncia correta, antes da tentativa.

Na última etapa, foram disponibilizados alguns minutos de conversação livre entre o aluno e o avaliador utilizando a transcrição automática. Cada um deveria falar alternadamente frases curtas e aguardar a transcrição ser enviada no *chat*. A partir dos dados no histórico, era possível acompanhar todo o andamento da conversa como uma legenda e o avaliador poderia verificar o quão assertiva era a ferramenta.

4 ANÁLISE E DISCUSSÃO DOS RESULTADOS

4.1 Análise Comparativa entre as ferramentas Web Speech API, Google Speech to Text e REV.AI

O resultado mostrou que as ferramentas têm um desempenho muito próximo. Google Speech to Text obteve 94% de taxa de acerto por palavra, enquanto a Rev.io teve 92% e a Web Speech API 89%.

Dentre os 300 áudios transcritos, foi verificado que ocorreram mais erros nas gravações de frases que foram pronunciadas mais rapidamente. Enquanto frases pronunciadas mais pausadamente tinham uma taxa de acerto próximo a 100

A Web Speech API se destaca por ser a única API totalmente gratuita. Por outro lado, o seu uso é totalmente limitado aos navegadores *web*. Atualmente estando disponível apenas nas versões mais atualizadas dos navegadores Safari, Edge e Google Chrome.

Segue abaixo, um quadro comparativo com as principais características de cada ferramenta analisada.

Tabela 1 – Comparativo das ferramentas.

	Google Speech to Text	Rev.AI	Web Speech Api
Taxa de Acerto Por Palavra.	94%	92%	89%
Suporte	multiplataforma	multiplataforma	apenas navegadores web
Pontuação	sim	sim	não
Licença	serviço pago	serviço pago	grátis
Adaptação transcrição por contexto	sim	não	não

Fonte: Autor

4.2 Resultados da aplicação prática da Web Speech API com os voluntários

4.2.1 Primeira etapa: Transcrição das frases do grupo 1.

O experimento mostrou que a Web Speech API teve um excelente desempenho nas transcrições de frases do grupo 1. Foi levado em consideração que quando os usuários acertavam a pronúncia, a ferramenta deveria transcrever corretamente.

Com todos os voluntários, as frases do grupo 1 foram pronunciadas de maneira correta, na primeira ou na segunda rodada. Por se tratar de frases comuns, os voluntários já tinham conhecimento prévio de como pronunciá-las, e as transcrições também não falharam.

A ferramenta foi muito eficiente ao transcrever as frases mais simples, mesmo quando o sotaque do voluntário não estava próximo do inglês norte americano, a transcrição acontecia sem erros.

4.2.2 Segunda Etapa: Transcrição das frases do grupo 2

Na segunda etapa, ao tentar reconhecer frases do grupo 2, as mais complexas de se pronunciar, foram identificados alguns problemas. Na primeira tentativa, a maior parte dos voluntários não conseguiram pronunciar da maneira correta.

A partir da segunda tentativa, foi disponibilizado um áudio com a frase pronunciada corretamente pelo google tradutor. Os voluntários podiam tentar várias vezes. Em alguns casos, como na frase 2.5, mesmo pronunciando corretamente a palavra "eighth", a ferramenta não foi capaz de reconhecê-la com nenhum voluntário.

No geral, a ferramenta foi capaz de reconhecer aproximadamente 60% das palavras pronunciadas corretamente desse grupo. Os problemas na transcrição serão discutidos na seção Problemas Encontrados.

4.2.3 Terceira Etapa: Conversação livre

Durante a aplicação da última etapa do experimento, onde o avaliador e o voluntário poderiam conversar livremente, foram encontrados pontos positivos que podem contribuir não só para o aprimoramento da pronúncia, mas também para a prática da conversação.

A aplicação permitia que os voluntários pudessem consultar o histórico da conversa transcrita. Com isso, mesmo quando não eram capazes de compreender o que foi dito, poderiam olhar no histórico a transcrição da última frase e manter o andamento da conversação sem a necessidade da outra parte precisar repetir o que foi dito. Entretanto é questionável efetividade didática desse comportamento, se faz necessário que sejam realizados experimentos para valida-lo como uma boa prática.

4.2.4 Problemas Encontrados

A ferramenta tem uma resposta bem ágil, enquanto o voluntário estava falando, a ferramenta já exibia um resultado prévio da transcrição, conforme configurado pelo atributo interimResults.

Entretanto, nos casos onde o voluntário errava a pronúncia das frases do grupo 2, a ferramenta alterava o resultado da transcrição bruscamente. Foi observado, e também confirmado ao ler na documentação, que ao identificar alguma frase incongruente, a API tenta adaptar a transcrição para alguma frase coerente mais próxima do som captado.

Esse comportamento não é desejado em um sistema didático, porque desconsidera boa parte do que havia reconhecido anteriormente ao erro, e altera totalmente a transcrição final. A opção encontrada mais viável na documentação para contornar parcialmente esse problema foi através do uso do recurso 'confidence' que será explorada na seção Futuros Projetos.

Uma limitação da ferramenta é a obrigação de ter que configurar apenas um idioma e um sotaque para ser reconhecido. Estava configurado para reconhecer o Inglês norte americano,

e com isso, os voluntários precisaram ser avisados que suas pronúncias não poderiam ser parecidas com outros sotaques, como o sotaque britânico.

Outro problema foi encontrado na última etapa do experimento, em que os voluntários poderiam improvisar uma conversa livremente. A Web Speech API não foi capaz de reconhecer nomes próprios quando esses nomes não são comuns no idioma inglês. Por exemplo, sempre que algum voluntário tentava dizer o próprio nome, a transcrição cometia erros a partir desse momento.

4.2.5 Trabalhos futuros

A Web Speech API oferece a propriedade confidence vinculado a cada transcrição realizada, que é um índice de confiabilidade da transcrição que retorna um número entre 0 a 1. Foi verificado durante o experimento, que quando o grau de confiabilidade era abaixo de 0.7, normalmente significava que o voluntário havia errado a pronúncia.

Como abordado na revisão literatura, segundo Neri (2003) um dos aspectos importantes que um sistema de ensino de língua estrangeira deve possuir para validar a pronúncia: é a pontuação. O índice de confiabilidade poderia ser utilizado como um parâmetro para determinar se a pronúncia da frase foi correta ou não.

O cenário ideal para se ter boa aplicação didática, é permitir que o usuário seja capaz de visualizar com detalhes se a sua pronúncia foi boa ou não foi. O *feedback* deve ser ainda mais específico mostrando quais palavras dentro da frase tiveram uma pronúncia correta e quais não tiveram.

A Web Speech API também oferece o recurso de síntese de fala, uma funcionalidade que pode agregar positivamente no sistema. Durante o aprendizado é comum que o estudante não saiba a maneira correta de pronunciar algumas palavras. O sistema poderia contar com um módulo de treinamento de pronúncia que emitiria áudios com a pronúncia correta de palavras complexas configuradas para o estudante ter uma referência sem precisar consultar fora da aplicação. Esses são alguns pontos técnicos da API que podem ser levados em consideração em projetos futuros.

5 CONCLUSÃO

O reconhecimento automático de fala aplicado ao ensino de idiomas mostrou resultados promissores, embora tenha apresentado alguns obstáculos para uma aplicação didática mais eficiente. Conforme mencionado durante a revisão literária, não há um consenso entre os pesquisadores a respeito da sua efetividade como mecanismo de aprendizagem de pronúncia de língua estrangeira.

Alguns estudos, como Derwing (2000) e Coniam (1999), apontaram problemas como ineficiência ao transcrever pessoas não fluentes. Entretanto o resultado do experimento com voluntários realizado neste trabalho mostrou que a ferramenta Web Speech API foi capaz de transcrever a fala de diferentes voluntários, com nível de inglês intermediário, com poucos erros na transcrição de frases mais comuns e de pronúncia mais fácil.

O problema encontrado foi o constante erro na transcrição de frases mais raras e de pronúncia complexa. Durante a frase, quando não era reconhecida a pronúncia de alguma palavra com fonema complexo, o comportamento da API era tentar adaptar imediatamente a transcrição para alguma frase coerente.

Conclui-se que o aspecto mais crítico identificado não foi em relação ao desempenho na transcrição, mas foi a respeito do real propósito da tecnologia. Com o objetivo de ser mais eficiente, a API de reconhecimento automático de fala tenta adaptar o áudio captado através de modelos linguísticos para formar frases que façam sentido gramaticalmente, aspecto que foi abordado por Speechweb (2020).

Esse comportamento é inapropriado nesse contexto por tentar deduzir e não apenas levar em consideração os sinais de áudio referente a transcrição. Dessa forma foi observado que era desconsiderado boa parte da pronúncia do usuário para tentar formar alguma frase coerente. O que prejudica o propósito didático pois não permite que o usuário identifique com detalhes a qualidade da sua pronúncia sem esse tratamento realizado pelo *software*.

As ferramentas de reconhecimento automático de fala testadas não foram projetadas para o uso didático, apesar de possuírem uma boa taxa de acerto por palavra no uso comum. O experimento de comparação entre as ferramentas analisadas mostrou que todas tiveram uma taxa alta de acerto, aproximadamente 90%, com os registros da base de dados pública. Entretanto é necessário que as ferramentas disponibilizem mais recursos de customização do mecanismo da transcrição para evitar comportamentos que são inapropriados no contexto didático, como o de dedução e adaptação da fala captada.

O cenário ideal para se obter um eficiente uso didático de ferramentas computacionais para o aprendizado de pronúncia descrito por Neri (2003) envolve alguns aspectos cruciais como a capacidade do sistema em dar *feedbacks* específicos sobre quais palavras foram pronunciadas corretamente e quais não foram. Durante os experimentos não foram encontrados nas API recursos capazes de customizar propriedades na transcrição que evite que os modelos linguísticos

mudem o resultado e desconsidere a pronúncia captada.

Referências

ADORF, J. **Web Speech API**. [S.l.], 2013. Disponível em: <https://www.juliusadorf.com/pub/web-speech-api.pdf>. Acesso em: 16 de outubro de 2021. Citado na página 15.

AHN, L. V. **Vastly Improved Speaking Exercises in Chrome**. [S.l.], 2014. Disponível em: <https://www.duolingo.com/comment/1880538>. Acesso em: 21 de outubro de 2021. Citado na página 17.

ANGGRAINI, N. **Speech Recognition Application for the Speech Impaired using the Android-based Google Cloud Speech API**. [S.l.], 2018. Disponível em: <https://www.researchgate.net/publication/329865231_Speech_Recognition_Application_for_the_Speech_Impaired_using_the_Android-based_Google_Cloud_Speech_API>. Acesso em: 20 de outubro de 2021. Citado na página 14.

AVERY, P. **Teaching American English Pronunciation**. [S.l.], 1992. Disponível em: <https://global.oup.com/academic/product/teaching-american-english-pronunciation-9780194328159?lang=en&cc=be#>. Acesso em: 20 de outubro de 2021. Citado na página 12.

BABBEL. **Babbel**. [S.l.], 2021. Disponível em: <https://about.babbel.com/pt/about-us/>. Acesso em: 1 de dezembro de 2021. Citado na página 10.

BAJOREK, J. P. **L2 Pronunciation in CALL: The Unrealized Potential of Rosetta Stone, Duolingo, Babbel, and Mango Languages**. [S.l.], 2017. Disponível em: <https://www.researchgate.net/publication/323268458_L2_Pronunciation_in_CALL_The_Unrealized_Potential_of_Rosetta_Stone_Duolingo_Babbel_and_Mango_Languages>. Acesso em: 29 de outubro de 2021. Citado na página 17.

CONIAM, D. **Voice recognition software accuracy with second language speakers of English**. [S.l.], 1999. Disponível em: <https://www.researchgate.net/publication/223890863_Voice_recognition_software_accuracy_with_second_language_speakers_of_English>. Acesso em: 20 de outubro de 2021. Citado 3 vezes nas páginas 10, 16 e 25.

CUCCHIARINI, C. **ASR-based corrective feedback on pronunciation: does it really work?** [S.l.], 2006. Disponível em: <https://www.researchgate.net/publication/221482694_ASR-based_corrective_feedback_on_pronunciation_does_it_really_work>. Acesso em: 31 de outubro de 2021. Citado na página 16.

DERWING, T. M. **Does Popular Speech Recognition Software Work with ESL Speech?** [S.l.], 2000. Disponível em: <https://www.jstor.org/stable/3587748>. Acesso em: 20 de outubro de 2021. Citado 3 vezes nas páginas 10, 16 e 25.

DUOLINGO. **Duolingo**. [S.l.], 2021. Disponível em: <https://www.duolingo.com/approach>. Acesso em: 1 de dezembro de 2021. Citado na página 10.

DUOLINGORESEARCH. **Research Duolingo: A ciência fortalece nossa missão de tornar o ensino de idiomas gratuito e acessível a todos.** [S.l.], 2021. Disponível em: <https://research.duolingo.com/>. Acesso em: 1 de dezembro de 2021. Citado na página 10.

GILAKJANI, A. P. **Why is Pronunciation So Difficult to Learn?** [S.l.], 2011. Disponível em: <https://files.eric.ed.gov/fulltext/EJ1080742.pdf>. Acesso em: 20 de outubro de 2021. Citado 2 vezes nas páginas 10 e 12.

GOOGLE. **An Overview of End-to-End Automatic Speech Recognition**. [S.l.], 2021. Disponível em: <https://cloud.google.com/speech-to-text>. Acesso em: 22 de novembro de 2021. Citado 2 vezes nas páginas 14 e 15.

LEVIS, J. M. **Automatic Speech Recognition**. [S.l.], 2012. Disponível em: <https://www.researchgate.net/publication/261287458_Automatic_Speech_Recognition>. Acesso em: 21 de outubro de 2021. Citado 3 vezes nas páginas 12, 13 e 14.

LIBBY, A. **Introducing the HTML5 Web Speech API**. [S.l.], 2020. Disponível em: <https://link.springer.com/book/10.1007/978-1-4842-5735-7>. Acesso em: 15 de outubro de 2021. Citado 2 vezes nas páginas 10 e 15.

LOTHERINGTON, H. **What's app? Negotiating the good, bad, and ugly of apps for (English and other) language learning.** [S.l.], 2016. Disponível em: <https://www.researchgate.net/publication/303896153_What's_app_Negotiating_the_good_bad_and_ugly_of_apps_for_English_and_other_language_learning>. Acesso em: 22 de outubro de 2021. Citado na página 17.

MCCROCKLIN, S. **ASR-based dictation practice for second language pronunciation improvement**. [S.l.], 2019. Disponível em: <https://www.jbe-platform.com/content/journals/10.1075/jslp.16034.mcc>. Acesso em: 31 de outubro de 2021. Citado 2 vezes nas páginas 10 e 16.

NERI, A. **Automatic speech recognition for second language learning: How and why it actually works**. [S.l.], 2003. Disponível em: <https://www.researchgate.net/publication/228604457_Automatic_speech_recognition_for_second_language_learning_How_and_why_it_actually_works>. Acesso em: 22 de outubro de 2021. Citado 4 vezes nas páginas 16, 17, 24 e 25.

REV. **Rev.ai API Overview, Documentation**. [S.l.], 2021. Disponível em: <https://www.rev.ai/docs/overview>. Acesso em: 16 de novembro de 2021. Citado 2 vezes nas páginas 15 e 16.

SPEECHWEB. **Reconhecimento de fala**. [S.l.], 2020. Disponível em: <https://speechweb.cpqd.com.br/asr/docs/latest/get_started/asr_intro.html>. Acesso em: 22 de novembro de 2021. Citado 3 vezes nas páginas 12, 13 e 25.

VOXFORGE. **About VoxForge**. [S.l.], 2021. Disponível em: <http://www.voxforge.org/home/about>. Acesso em: 10 de outubro de 2021. Citado 2 vezes nas páginas 13 e 18.

WANG, D. **An Overview of End-to-End Automatic Speech Recognition**. [S.l.], 2019. Disponível em: <https://www.mdpi.com/2073-8994/11/8/1018/html>. Acesso em: 22 de novembro de 2021. Citado na página 14.

YU, D. **Automatic Speech Recognition**. [S.l.], 2015. Disponível em: <https://link.springer.com/book/10.1007/978-1-4471-5779-3>. Acesso em: 31 de outubro de 2021. Citado na página 14.