Universidade de São Paulo
Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto
Departamento de Psicologia
Programa de Pós-Graduação em Psicobiologia

Fernanda de Barros Vidal

Leitura da fala do Português Brasileiro: elaboração de vídeos para treinamento

Ribeirão Preto / SP
2024

Fernanda de Barros Vidal

Leitura da fala do Português Brasileiro: elaboração de vídeos para treinamento

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Psicobiologia da Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da Universidade de São Paulo para obter o título de Mestre em Ciências (com ênfase em Psicobiologia)
Orientador: Prof. Dr. Sérgio Sheiji Fukusima

Ribeirão Preto / SP
2024

Vidal, Fernanda de Barros.

Título: Leitura da fala do Português Brasileiro: elaboração de vídeos para treinamento / Fernanda de Barros Vidal – Ribeirão Preto/SP, 2024. 71 páginas.

Dissertação de Mestrado – Universidade de São Paulo, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Programa de Pós-Graduação em Psicobiologia, 2024.

Orientador: Sérgio Sheiji Fukusima.

Título em inglês: Speechreading of Brazilian Portuguese: elaboration of videos for training

1. Percepção visual da fala. 2. Leitura labial do Português Brasileiro. 3. Frequência de uso e competição lexical de palavras. I. Fukusima, Sérgio Sheiji. II. Leitura da fala do Português Brasileiro: elaboração de vídeos para treinamento.

AGRADECIMENTOS

Agradeço aos meus pais, Ricardo e Marlene, que me apoiam e me inspiram como pais e como pessoas. Sinto como se já tivesse conquistado o mundo só por ter esse amor incondicional vindo de pessoas tão incríveis, que me ensinam, acreditam em mim, me impulsionam e acolhem ao mesmo tempo, de um modo extraordinário. Jamais serei capaz de expressar toda minha gratidão, tento demonstrar diariamente e com a minha vida o quanto sou grata por vocês.

Agradeço à Deus por sempre iluminar meu caminho, me abençoar e proteger.

Agradeço à minha família por sempre acreditarem em mim e compreenderem as minhas escolhas, mesmo sem entendê-las completamente, e me apoiarem incondicionalmente. Especialmente à minha vó, Glória, que sempre me diz "você tá fazendo aquilo que sua mãe fez né (mestrado)?" e pergunta se está acabando ou se eu estou ganhando bem, e à minha irmã, Maria Luísa, que me recorda sempre da parte da gente que não cresce e é mantida pela essência.

Agradeço à minha psicóloga, Fabiane Neves, que caminhou ao meu lado para que eu não perdesse o equilíbrio e me acompanha no amadurecimento para lidar com a liberdade que vem com a responsabilidade de seguir meus sonhos, me mostrando que eu posso encarar a vida com leveza, seriedade, espontaneidade e paciência comigo mesma.

Agradeço ao meu orientador, prof. Dr. Sérgio S. Fukusima, pela orientação e por ter visto o potencial desse tema e me indicado a pesquisá-lo. Também por ter me permitido trabalhar com autonomia, diálogo e me auxiliado, mapeando caminhos possíveis durante a realização do trabalho, e por inspirar minha trajetória profissional dentro de uma linha de pesquisa interdisciplinar.

Agradeço aos professores que sempre estiveram disponíveis e dispostos a me auxiliar, tiraram dúvidas e indicaram direcionamentos ao longo da pesquisa, sempre visando o caráter interdisciplinar: prof. Dra. Heliana Mello (Linguística - UFMG), Dra. Brasília Chiari (*in memoriam*) (Fonoaudiologia - Unifesp), Dra. Ruth Campbell (Psicolinguística/Neurociência Cognitiva - UCL), Dra. Máiread McSweeney (Psicolinguística/Neurociência Cognitiva - UCL), Dra. Sandra Aluísio (Linguística Computacional - USP/São Carlos), Dr. Gustavo Estivalet (Linguística - UFPB), Dra. Jamila Viegas (Linguística - UFLA), Dra. Marisa Fukuda (Fonoaudiologia - USP/Ribeirão Preto), Dra. Patrícia Pupim (Fonoaudiologia - USP/Ribeirão Preto), Dra. Aline Wolf (Fonoaudiologia - USP/Ribeirão Preto). Agradeço à fonoaudióloga Luísa Stefano, pelo auxílio na compreensão de conceitos da fonoaudiologia e seleção das palavras de estímulo.

Agradeço aos locutores voluntários por estarem dispostos a contribuir com a pesquisa através da sua imagem, pela disponibilidade em ir até o laboratório e paciência para que pudéssemos realizar um bom trabalho.

Agradeço aos participantes da pesquisa por terem dedicado seu tempo para responder a sessão experimental e, dessa forma, contribuírem para o avanço da pesquisa nessa área.

Agradeço à FAPESP e à CAPES por terem viabilizado a realização dessa pesquisa.

Agradeço ao Programa de Pós-graduação em Psicobiologia pela menção honrosa a esse trabalho na XVII Reunião Anual da Psicobiologia e a secretária Renata pelo auxílio em trâmites burocráticos no decorrer do mestrado.

Agradeço aos meus colegas de laboratório, Melina Urtado, Larissa Kuroishi, Flávia Santiago e Alexandre Gonzaga pelas trocas que tivemos desde que realizei o curso de verão até o final do mestrado. Em especial ao Alexandre, que me auxiliou com direcionamentos e ideias desde a escrita do projeto.

Agradeço aos meus amigos e as pessoas que estiveram ao meu lado nesse período do mestrado, que contribuíram em diferentes proporções, momentos e maneiras para o bom andamento deste trabalho e foram meu suporte em muitos momentos, sem nem perceberem:

Denison, Amanda Hellen, Giovanna, Matheus, Paolla, Eduardo, Paulo, Letícia Costa, Gabrielle, Marcos, Eloísa, Bianca, Alexandre, Aura, Carolina Kato, Hellen, Marcelo, Luísa Stefano, Luiza Vito e Natália. Em especial à Bárbara Munhão, Amanda Ferraz, Andreza Fonseca, Letícia Mesquita, Carolina França e Maria Clara Lopes, que me acolheram nos momentos em que as coisas estavam difíceis e diariamente me ajudaram a me manter firme em um período conturbado, graças a vocês este período foi maravilhoso, mesmo com as dificuldades, e nunca me esquecerei disso.

A vontade de me tornar pesquisadora é presente na minha vida desde o ensino médio, quando tive meu primeiro contato com a pesquisa, por isso, agradeço também ao meu primeiro orientador, prof. Dr. Luiz Fernando Lomba, por ter me ensinado, fomentado esse sonho e agora nesse período ter compartilhado angústias que muitas vezes passam despercebidas.

A vida acadêmica é permeada pela formulação de novas ideias a partir de conceitos previamente desenvolvidos em um caminho que não tem fim e que não tem verdades absolutas. Por isso, o desafio de quem decide seguir por esse caminho é se manter curioso, persistente e resiliente para buscar mais, pois é assim que a ciência se mantém viva.

Meu mestrado começou com um falecimento na minha família. Aquele momento difícil, com meu primo Gabriel, me abriu os olhos para a urgência em viver e amar. Ele e todas as pessoas citadas neste agradecimento, direta ou indiretamente, estarão sempre comigo enquanto eu estiver viva, no meu coração, e foram parte essencial da realização deste trabalho.

RESUMO

Vidal, F. B., Fukusima, S. S. (2024). Leitura da fala do Português Brasileiro: elaboração de vídeos para treinamento. Dissertação de Mestrado. Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto. Universidade de São Paulo, Ribeirão Preto/SP.

O tema de percepção da fala é interdisciplinar e estudado por meio de abordagens variadas, podendo ser investigado por associação bimodal (audiovisual) ou unimodal (somente auditiva ou visual). A percepção visual da fala, por meio da leitura labial, é bem explorada em diversos idiomas, como o inglês. No entanto, poucos estudos foram realizados sobre os paradigmas que influenciam essa percepção em Português Brasileiro e nenhum material de estímulo com faces naturais foi desenvolvido e disponibilizado considerando essas características já bem investigadas no inglês. Esta pesquisa exploratória propôs desenvolver uma base de arquivos de vídeo para treinamento de leitura labial em Português Brasileiro, por meio de estímulos naturais, com palavras que contemplem diferentes fonemas. Foram produzidos 110 vídeos com dois locutores fonoaudiólogos pronunciando palavras de estímulo, seguindo diretrizes para produção de vídeos de leitura labial. As palavras de estímulo foram selecionadas a partir do *corpora* C-Oral Brasil, do português falado em ambientes informais, classificadas gramaticalmente como substantivos, dissílabas e que possuíam a mesma estrutura Consoante-Vogal (CV) com os 19 fonemas consonantais do idioma balanceados na primeira e segunda sílaba. Para testar os vídeos produzidos, 281 participantes ouvintes e com idade entre 18 e 55 anos responderam uma sessão experimental computadorizada, hospedada na plataforma Pavlovia, composta por 30 vídeos com as palavras de estímulo. Ao final de cada vídeo os participantes indicavam uma alternativa de resposta, entre a palavra de estímulo e duas alternativas distratoras, que configuravam possíveis competidores lexicais e variavam conforme o posicionamento do fonema consonantal na palavra, e recebiam feedback. Foi realizada a estatística descritiva dos dados e análises exploratórias, a partir da sessão experimental. Dentre as 30 palavras de estímulo, 24 delas tiveram uma proporção de acerto maior que 70%. A correlação entre o tempo médio de resposta e a frequência de resposta por tipo de resposta foi calculada por meio do tau de Kendall ($p<.05$). 23 dos 30 itens apresentaram correlação negativa significativa em relação as respostas corretas. Em alternativas incorretas, a correlação foi positiva e significativa. A proporção de respostas corretas foi maior em palavras em que pelo menos uma das alternativas distratoras divergia quanto ao modo de articulação. A análise exploratória de agrupamento (*cluster*) dos fonemas foi realizada para verificar possíveis agrupamentos quanto as classificações das consoantes. Não foi observado nenhum padrão de agrupamento pela sonoridade, nasalidade ou ponto de articulação, no entanto, o modo de articulação de plosivas, no geral, teve melhor identificação. Novos trabalhos vão poder usar a base de arquivos (110 vídeos) produzidos nessa pesquisa. A investigação do tema em português brasileiro está em caráter inicial. Para desenvolver um teste computadorizado robusto, entender a relação da leitura labial com habilidades cognitivas e de aprendizagem e estabelecer tarefas de treinamento e investigação em outros âmbitos, é necessário considerar particularidades da língua relacionadas ao léxico e realizar mais pesquisas interdisciplinares.

Palavras-chave: percepção visual da fala; leitura labial do Português Brasileiro; frequência de uso e competição lexical de palavras

Apoio: Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP 2022/00801-2)

ABSTRACT

Vidal, F. B., Fukusima, S. S. (2024). Speechreading of Brazilian Portuguese: elaboration of videos for training. Master Thesis. Faculty of Philosophy, Sciences and Letters at Ribeirão Preto. University of São Paulo, Ribeirão Preto/SP.

Speech perception is an interdisciplinary topic studied through different approaches. It can be investigated by bimodal (audiovisual) or unimodal (only auditory or visual) association. The visual speech perception, through lipreading, is well explored in several languages, such as English. Although few studies have been carried out on the paradigms that influence this perception in Brazilian Portuguese and no stimulus material with natural faces has been developed and made available considering characteristics that are already well investigated in English. This exploratory research proposed to develop a database of video files for lipreading training in Brazilian Portuguese, through natural stimuli, with words that include different phonemes. 110 videos were produced with two speech therapists pronouncing stimulus words, following guidelines for producing lipreading videos. The stimulus words were selected from the C-Oral Brazil *corpora*, from Portuguese spoken in informal environments, classified grammatically as nouns, dissyllables and which had the same Consonant-Vowel (CV) structure with the 19 consonant phonemes balanced in the first and second syllable. To test the videos produced, 281 hearing participants aged between 18 and 55 completed a computerized experimental session, hosted on the Pavlovia platform, consisting of 30 videos with the stimulus words. At the end of each video, participants indicated an alternative response, between the stimulus word and two distracting alternatives, which configured possible lexical competitors and varied according to the positioning of the consonant phoneme in the word and received feedback. Descriptive statistics of the data and exploratory analyses were carried out, based on the experimental session. Among the 30 stimulus words, 24 of them had a correct answer rate greater than 70%. The correlation between the average response time and response frequency by response type was calculated using Kendall's tau ($p<.05$). 23 of the 30 items showed a significant negative correlation in relation to correct answers. In incorrect alternatives, the correlation was positive and significant. The proportion of correct responses was higher in words in which at least one of the distracting alternatives differed in terms of manner of articulation. Exploratory cluster analysis was carried out to check possible groupings regarding distinctive features of the phonemes. No grouping pattern by sonority, nasality or manner and place of articulation was observed. However, in general the manner of articulation of plosive had better identification than other phonemes. New research will be able to use the file base (110 videos) produced in this study. The investigation of the visual speech perception in Brazilian Portuguese is at an initial stage. It is vital to consider particularities of the language related to the lexicon and carry out more interdisciplinary research to develop a robust computerized test, understand the relationship between lipreading and cognitive and learning skills, and establish training and research tasks in other areas.

Keywords: visual speech perception; lipreading of Brazilian Portuguese; frequency and lexical competition of words

Support: Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP 2022/00801-2)

SUMÁRIO

LISTA DE TABELAS

LISTA DE IMAGENS

LISTA DE GRÁFICOS

1. INTRODUÇÃO

De acordo com Massaro (2015), a percepção da fala é definida como o processo de experiência perceptual significativa a um estímulo de fala sem sentido. A investigação desse tema amadureceu e se consolidou por um esforço interdisciplinar de áreas como a psicofísica, neurofisiologia, percepção, linguística, psicolinguística, inteligência artificial, sociolinguística e fonoaudiologia, por isso, seu estudo envolve um conjunto variado de abordagens experimentais e teóricas.

Independente da área de estudo, o tema de percepção da fala pode ser pesquisado por meio de associações bimodais (audiovisual) ou unimodais (somente auditiva ou visual). A percepção visual da fala, tema ao qual este estudo está majoritariamente vinculado, pode ser estudada por meio da pesquisa em leitura labial (termo em inglês: *lipreading*). Toffolo et al. (2017) explicam que se entende por leitura labial a compreensão da informação falada por meio da observação dos movimentos da boca e dos lábios e de outras pistas, como expressões faciais, gestos e mudanças de postura. Segundo Oliveira, Soares e Chiari (2014), essa técnica/habilidade também pode ser denominada de leitura orofacial ou leitura da fala (termo em inglês: *speechreading*).

Oliveira, Soares e Chiari (2014) indicam que há uma grande quantidade de estudos internacionais sobre leitura labial e afirmam que poucas pesquisas foram desenvolvidas a respeito dos fatores que influenciam nessa habilidade em diferentes idiomas. Toffolo et al. (2017) também ressaltam a internacionalidade dos estudos a respeito de leitura labial, ou seja, desenvolvidos em outros idiomas, que não o Português Brasileiro (PB). Os autores trazem como exemplo uma peculiaridade da língua inglesa, em que aproximadamente 30% dos sons são visíveis nos lábios e destacam a possibilidade de uma maior regularidade articulatória em Português Brasileiro do que em outros idiomas, como o Francês ou o Inglês. Porém, os autores afirmam que não encontraram pesquisas em Português Brasileiro relacionadas ao tema.

A leitura labial auxilia na oralização (produção da fala) e pode ser um fator relevante para aquisição do português escrito como segunda língua e na fluência em leitura para pessoas com deficiência auditiva (Toffolo et al., 2017). A proficiência em leitura refuta a hipótese de que "fraco desempenho de leitura dos indivíduos surdos seja resultante da falta (completa ou parcial) de um *input* auditivo, considerado necessário para o desenvolvimento de representações fonológicas bem especificadas" (Toffolo et al., 2017, p. 5). *Input* se refere ao estímulo de entrada, ou seja, a informação recebida.

De acordo com Oliveira, Soares e Chiari (2014), todos os indivíduos fazem o uso da leitura labial, no entanto, pessoas ouvintes somente percebem a utilização dessa técnica quando

estão em ambientes em que seja difícil reconhecer o estímulo auditivo, com interferência de ruído, conteúdo ou vocabulário complexo e idioma ou sotaque diferente do qual a pessoa está habituada. Também, conforme as autoras, para pessoas não ouvintes, com grau leve ou moderado de deficiência auditiva, a leitura labial funciona como um complemento ao aparelho auditivo, para compreensão e comunicação efetiva e, na deficiência auditiva profunda, como auxílio na comunicação.

Para compreender o desenvolvimento das pesquisas em percepção visual da fala relacionadas a leitura labial, cabe esclarecer os avanços nos estudos de percepção auditiva e audiovisual da fala, uma vez que esses influenciaram os rumos da investigação do reconhecimento visual de diferentes idiomas. Segundo Bernstein (2012), na maior parte do século XX, o estudo da percepção da fala foi principalmente relacionado à percepção auditiva, sendo esse tema muitas vezes visto como resultado do sistema auditivo. No entanto, a percepção visual da fala começou a ganhar destaque quando MacDonald e McGurk (1976) relataram o Efeito McGurk, que demonstrava como a informação visual influenciava na percepção auditiva.

MacDonald e McGurk (1978) investigaram a hipótese de que o modo e ponto de articulação na fala poderiam explicar a confusão entre estímulo de entrada (*input*s) auditivos e visuais associados. Os autores realizaram um experimento para analisar as interações do modo e ponto de articulação das informações audiovisuais percebidas por adultos e crianças ouvintes. O locutor do vídeo de estímulo falava uma sílaba (por exemplo: ba), mas era dublado com outra (por exemplo: ga), assim as informações auditivas e visuais eram associadas. Os autores encontraram diferentes tipos de interação das informações, que classificaram como fusão e combinação. A fusão ocorria quando um estímulo auditivo (“ba”) associado a um estímulo visual (“ga”) resultava na percepção de uma terceira sílaba pela combinação dos estímulos (“da”). Neste experimento, 98% dos adultos e 81% das crianças relataram ter percebido a fusão dos estímulos. Na combinação, a associação do estímulo auditivo (“ga”) ao visual (“ba”) levou a percepção de estímulos combinados (“gabga” e “bagba”), relatados por 54% dos adultos e 57% das crianças. Esses resultados deram origem ao conceito conhecido como Efeito McGurk, que destacou a relevância da informação visual na percepção da fala e como essa informação influenciava principalmente a percepção dos adultos.

Foi durante esse período, na segunda metade do século XX, que a importância da informação visual nas pesquisas de percepção da fala ganhou destaque. Os paradigmas de investigação em percepção visual da fala foram complementados pelos avanços significativos nas pesquisas em percepção auditiva, os quais desempenharam um papel essencial na definição de paradigmas para a investigação da modalidade visual.

Essas influências podem ser observadas a partir do estudo de Luce e Pisoni (1998), que comprovou que o número e a natureza das palavras semelhantes a uma palavra-alvo afetavam a velocidade e a precisão do reconhecimento da palavra pela via auditiva. Os autores apresentaram o termo de densidade de vizinhança, para definir o número de vizinhos de uma palavra, e explicaram que na maioria das pesquisas a vizinhança da palavra era estabelecida como o número de palavras reais que poderiam ser formadas a partir de uma substituição, adição ou exclusão de um fonema naquela palavra. Além disso, os autores trabalharam com a hipótese de que palavras com muitos vizinhos (vizinhança densa) eram mais difíceis de reconhecer do que palavras com poucos vizinhos (vizinhança esparsa). Luce e Pisoni (1998) explicaram, também, a relevância da frequência de uso para o reconhecimento, palavras que ocorrem com alta frequência no ambiente linguístico tem uma vantagem no processamento auditivo em relação a palavras de baixa frequência, que são mais difíceis de reconhecer.

Os autores desenvolveram uma regra de probabilidade de vizinhança que combinava inteligibilidade, confusão e frequência para prever o desempenho da identificação de palavras, e propuseram o Modelo de Ativação por Vizinhança para reconhecimento auditivo (termo em inglês: *Neighborhood Activation Model* (NAM)). O modelo proposto descrevia os efeitos da densidade de vizinhança no processo de reconhecimento das informações acústico-fonéticas das palavras na memória. A publicação desse estudo mostrou a relevância de variáveis específicas para o reconhecimento auditivo de palavras no processo competitivo que ocorre no léxico.

Os paradigmas de distinção lexical e a frequência de uso no reconhecimento de palavras pela via auditiva motivou pesquisas em percepção visual da fala usando paradigmas similares. Bernstein, Jordan, Auer e Eberhardt (2022) esclareceram que a competição lexical ainda é uma explicação bem aceita para o reconhecimento de palavras pela via auditiva. Uma palavra de estímulo ativa palavras semelhantes no léxico mental e o reconhecimento ocorre a depender de uma palavra superar outras palavras candidatas, ou seja, quando suas pistas fonéticas ou categorias fonêmicas divergem de outras palavras disponíveis.

A competição lexical, a partir da informação exclusivamente visual, começou a interessar pesquisadores de percepção visual da fala. Conforme Files, Tjan, Jiang e Bernstein (2015), até o final do século XX e início do século XXI era consenso entre os pesquisadores de leitura labial do Inglês que fonemas que se referiam a um mesmo visema não poderiam ser distinguidos somente a partir da informação visual. De acordo com Bernstein (2012), Fisher (1968) conceituou o termo "*viseme*" (termo em Português Brasileiro: visema), para se referir ao "*visual phoneme*" (fonema visual - tradução própria). Ainda conforme a autora, em visemas

há uma capacidade reduzida para diferenciar palavras, devido as características fonéticas visualmente empobrecidas. Kyle et al. (2013) explica que visemas são fonemas confundidos por serem visualmente semelhantes, como o /f/ e /v/ em “faca” e “vaca”. No entanto, pesquisadores da área se dedicaram a demonstrar que outros paradigmas influenciavam a competição lexical e poderiam diminuir o empobrecimento fonético na informação visual, contribuindo para que fonemas visualmente similares fossem distinguíveis.

Bernstein (2012) explica que a identificação isolada de fonemas em sílabas sem sentido não responde se a informação fornecida visualmente é ou não adequada e que, para compreender isso, é necessário conhecer outros fatores como a confusão entre os fonemas e os efeitos da distinção deles em palavras. Segundo a autora, as semelhanças entre a representação do estímulo de entrada e as formas das palavras armazenadas ao longo do tempo é o que impulsiona os níveis de ativação dessas formas. Ainda, a estrutura do léxico, em boas condições de percepção fonética, é favorável a níveis altos de reconhecimento visual de palavras faladas. A autora também ressalta a dificuldade em medir e manipular estímulos de fala visuais e enfatiza que compreender a percepção visual da fala exige entendimento da percepção fonética e das implicações da informação fonética relacionada ao processamento de palavras armazenadas no léxico.

Dessa forma, a autora relata que alguns estudos seguiram uma abordagem baseada nas observações de Shepard e Chipman (1970) para traçar as relações de dissimilaridade dos estímulos. Shepard e Chipman (1970) observaram que é improvável que uma representação interna de um objeto seja estruturalmente isomórfica com o estímulo. Essa dissimilaridade consistia em não estabelecer a busca de isomorfismo em uma relação de primeira ordem entre o estímulo e sua representação interna (como por exemplo, um estímulo de uma forma geométrica como um quadrado ser “representado internamente” como um quadrado), mas sim em segunda ordem entre o estímulo e objetos externos alternativos e as relações entre as representações internas de objetos próximos (por exemplo, de um quadrado com um retângulo e não com estímulos completamente divergentes como a cor verde ou um número).

Para estabelecer métricas da competição lexical de palavras do Inglês Americano, em 1997, Auer e Bernstein elaboraram uma metodologia de modelagem computacional do léxico. Os autores conceituaram o termo em Inglês *Phoneme Equivalence Class* (PEC) (Classe de Equivalência Fonêmica - tradução própria) para classificar a taxa de confusão perceptual no acesso lexical de fonemas que pertenciam a um mesmo sinal óptico/uma mesma classe. A PEC foi refinada para a *Lexical Equivalence Class* (LEC) (Classe de Equivalência Lexical - tradução própria). A diferença da LEC para a PEC é que a PEC abrange todos os fonemas que podem

competir e é definida como os conjuntos de fonemas de acordo com a semelhança perceptual entre eles, independente de formarem uma palavra ou logatoma (pseudopalavra) no idioma. Enquanto a LEC se refere ao conjunto de palavras que são similares a partir de um conjunto de PEC. Bernstein (2012) apresenta um exemplo: se os fonemas /b/, /p/, /m/ formarem uma PEC no Inglês, então as palavras "*bat*", "*pat*" e "*mat*" estariam no mesmo LEC. Ainda, segundo a autora, a equivalência se refere a matrizes de confusão que representam agrupamentos pelo nível de similaridade perceptual.

Auer e Bernstein (1997) explicaram que "a distribuição das palavras na língua preserva a singularidade lexical em uma ampla gama de distinções fonêmicas potencialmente disponíveis" (Auer e Bernstein, 1997, p. 3704). Por meio da elaboração de um léxico digital para representar classes de fonemas com todas as vogais e consoantes do Inglês, os autores demonstraram que quase todas as palavras em um léxico de 35.000 palavras eram visualmente similares a outras quando eram divididas em 10 classes. No entanto, quando esses fonemas eram agrupados em 12 classes distintas, a maioria das palavras era visualmente diferente das outras. Os autores concluíram que palavras com alta frequência de uso não residem nas mesmas classes de equivalência lexical que outras palavras que também são de alta frequência, por isso, segundo os autores, "pequenos incrementos no número de distinções fonêmicas visualmente perceptíveis podem resultar em mudanças substanciais na singularidade lexical" (Auer e Bernstein, 1997, p. 3704).

As métricas e o modelo descritos por Auer e Bernstein (1997) indicaram que os visemas poderiam ser distinguidos em uma competição lexical que considerasse outras variáveis, como a frequência de uso e palavras existentes na língua. Os autores também propuseram que o treinamento em leitura labial poderia se caracterizar como um paradigma interessante de investigação do reconhecimento visual de palavras, mesmo que este fosse para aumentar a percepção somente de algumas pistas visuais fonéticas.

A pesquisa de Mattys, Bernstein e Auer (2002) foi um dos trabalhos pioneiros referente ao Inglês Americano a investigar os efeitos da distinção lexical no reconhecimento visual de palavras e como estes efeitos poderiam estar associados à redução do empobrecimento fonético nos estímulos visuais. Auer e Bernstein (1997) haviam estabelecido as métricas, porém a competição lexical no processamento de estímulos exclusivamente visuais ainda não havia sido investigada. Mattys, Bernstein e Auer (2002) ressaltaram que os competidores lexicais diferem e podem mudar a depender da modalidade do estímulo (auditivo, visual (pela escrita ou leitura labial) ou audiovisual). Por exemplo, no reconhecimento auditivo de palavras do Inglês, o fonema /p/ na palavra *pack* (que corresponde a um bilabial surdo) é mais semelhante ao /t/ da

palavra *tack* (um alveolar surdo), do que ao /b/ da palavra *back* (um bilabial sonoro). Assim, os autores esclarecem que a concorrência de uma palavra pode divergir conforme a modalidade.

A quantidade de palavras competidoras também pode variar conforme o léxico. Os autores utilizaram-se da métrica da LEC para testar palavras com diferentes taxas de confusabilidade e exemplificaram como essa diferença pode acontecer no Inglês. Mesmo que três fonemas pertencessem a uma mesma classe de equivalência fonêmica (muito parecidos visualmente), como o /b/, /p/ e /m/ do exemplo anterior, algumas palavras que esses fonemas formariam, como *bought*, *pought*, *mought*, não seriam capazes de competir no léxico, pois somente um desses fonemas formaria uma palavra existente no idioma (*bought*), enquanto os outros corresponderiam a pseudopalavras (*pought* e *mought*), ou seja, não seriam competidores lexicais válidos no Inglês.

Os resultados apresentados por Mattys, Bernstein e Auer (2002) indicaram que o desempenho no reconhecimento visual de palavras do Inglês foi consistente com os paradigmas de percepção auditiva da fala, de discriminação lexical e da frequência de uso. Dessa forma, os autores concluíram que a dissimilaridade lexical e a frequência de uso eram dois fatores com forte influência no reconhecimento de palavras, independente da modalidade (auditiva ou visual).

Em paralelo, Auer (2002) também demonstrou a relevância dos processos perceptivos e lexicais no reconhecimento de palavras ao aplicar o Modelo de Ativação por Vizinhança em uma tarefa de reconhecimento visual de palavras. Os resultados para a modalidade visual corroboraram com os obtidos no reconhecimento auditivo. A competição entre os candidatos lexicais no reconhecimento visual da palavra, por meio da leitura labial, favoreceu as palavras existentes no léxico e de uso mais frequente no idioma. Assim, de acordo com Auer (2002), no reconhecimento visual de palavras do Inglês Americano mesmo que haja confusão entre dois fonemas distintos, que pertençam a um visema, mas um desses fonemas formar uma pseudopalavra, a palavra que poderá ser favorecida na competição lexical é a que forma uma palavra existente no idioma, e que tenha maior frequência de uso. No Português Brasileiro, por exemplo, entre as palavras “fato” e "vato”, a segunda não corresponde a uma palavra existente no idioma, logo, a primeira seria acessada previamente no léxico.

Ainda, traçar métricas de competição lexical visual de palavras faladas é complexo e exige especificidades relacionadas a cada idioma, com bases de dados restritas que não englobam toda a realidade da fala. Strand (2014) desenvolveu o *Phi-square Lexical Competition Database* (Phi-Lex), uma base de dados pública que fornece medidas computacionais de competição lexical auditiva e visual (de leitura labial) de aproximadamente

5.000 palavras do Inglês Britânico. A autora explica que, diferente da análise auditiva, as palavras-alvo no domínio visual foram comparadas somente com palavras que tinham a mesma estrutura (por exemplo, estrutura Consoante-Vogal-Consoante (CVC) comparadas com outras CVC, como em *cat* e *bat*). Ainda segundo a autora, na modalidade visual o determinante na identificação dos fonemas não é se algo está presente (detecção), mas sim o que está presente (discriminação). O movimento é facilmente perceptível, mas a identificação de qual o movimento é difícil. Assim, a autora explica que, visualmente, comparar palavras concorrentes de diferentes tamanhos não é aplicável à métrica, e exemplifica que a estrutura CVC é muito mais presente no Inglês Britânico do que a estrutura CCCV (C corresponde a Consoante e V a Vogal), o que distingue as métricas em razão da quantidade maior de palavras em diferentes estruturas.

A tentativas de compreensão de como ocorre o processamento de informações visuais da fala fizeram com que as pesquisas desenvolvidas tivessem diversas abordagens metodológicas, que ainda hoje não são consensuais e não possuem padronização, especialmente pelo caráter interdisciplinar do tema e dificuldade na coleta de informações de maneira manual. Assim, a investigação da leitura labial por meio do treinamento começou a ser amplamente explorada na tentativa de padronizar as tarefas e melhorar o controle metodológico. Bernstein, Jordan, Auer e Eberhardt (2022) explicam que o treinamento analítico, que se refere ao treino isolado de sílabas sem sentido ou palavras, era o paradigma predominante nas pesquisas do século XX, uma vez que as investigações eram mais focadas em compreender a linguagem do que na efetividade do treino. Porém, no início do século XXI, o treinamento sintético, que diz respeito ao treino de fala conectada por meio de sentenças ou frases, ganhou espaço nas tarefas de reconhecimento visual da fala, visto que as pesquisas começaram a ter um enfoque maior na efetividade do treino. As pesquisas demandavam estímulos ecologicamente válidos e que considerassem características do observador, como adivinhação e contexto, propiciados pelo treinamento sintético.

Bernstein, Jordan, Auer e Eberhardt (2022) esclareceram que em ambos os tipos de treinamento, o papel do *feedback* interno e externo foi pouco explorado na aprendizagem perceptiva do reconhecimento visual de palavras e como o treino afetava essa aprendizagem. Os autores explicaram que, até então, quando o *feedback* era fornecido ao observador, ele era relacionado apenas à resposta correta. Isso ocorria por conta das características experimentais pouco controladas. Assim, pesquisas com tarefas experimentais computadorizadas foram desenvolvidas para melhor controle metodológico dos estímulos e do *feedback.*

Bernstein, Auer e Eberhardt (2022) exploraram o papel do *feedback* impresso em três

contextos distintos: em sentenças, por palavras isoladas e por fonemas consonantais em tarefas de resposta aberta em condições de estímulo somente visual, somente auditivo e audiovisual com ruído, considerando teorias sobre aprendizagem perceptiva e a neuroanatomia funcional da leitura labial. Nas três condições de teste, o *feedback* em relação ao fonema consonantal foi o que apresentou melhores índices pós treinamento, em palavras dispostas em sentenças. Os autores concluíram que no treinamento de leitura labial um *feedback* externo precisa direcionar o processamento interno, para que as informações sublexicais sejam discriminadas. Segundo Bernstein, Auer e Eberhardt (2022), o *feedback* pode ocorrer tanto antes quanto depois do estímulo, no entanto, se ocorrer antes, pode afetar o aprendizado da leitura labial. Por isso, posterior ao estímulo é o mais recomendado para a aprendizagem perceptiva. Os autores apontam a necessidade de mais pesquisas sobre aprendizagem perceptiva em leitura labial para entender sobre a generalização do aprendizado e explicam que, na aprendizagem perceptiva visual, o *feedback* externo precisa ser relacionado ao erro perceptivo cometido no treinamento, para que seja eficaz, uma vez que apenas "correto" não fornece orientação para informações sublexicais. O *feedback* para respostas parcialmente corretas pode ser o caminho para impulsionar a aprendizagem perceptiva da fala visual.

No Inglês Britânico, Buchanan-Worster, Hulme, Dennan e MacSweeney (2021) realizaram uma tarefa computadorizada de treinamento em leitura labial com crianças não ouvintes, que indicou a melhora no reconhecimento de palavras. Segundo os autores, os estudos apontam evidências de que a leitura labial não é uma habilidade fixa e pode ser treinada tanto em crianças quanto em adultos surdos e ouvintes.

Em paralelo as investigações realizadas nas pesquisas relacionadas ao treinamento, testes de leitura labial computadorizados também foram desenvolvidos. Um dos testes pioneiros foi apresentado por Mohammed et al. (2006) através do *Test of Adult Speechreading* (TAS) (Teste de Leitura da Fala para Adultos - tradução própria). O TAS é composto por três sub-testes. O primeiro, com palavras isoladas, que tem 19 itens no total – quatro para praticar e 15 que se dividem em palavras monossílabas ou dissílabas. Os participantes veem o estímulo visual falado e tem que selecionar a palavra-alvo entre seis alternativas dispostas em figuras. O segundo, de sentenças, tem 18 itens no total – três para praticar e 15 que são formados por frases com três a seis palavras. Os participantes veem uma sentença falada e selecionam uma figura correspondente ao que perceberam, dentre seis alternativas. Já o terceiro, um sub-teste de fala conectada, em que há seis histórias curtas: a primeira para prática, seguida por cinco outras histórias com duas ou três sentenças. Ao final de cada história, os participantes recebem três

questões a respeito de cada uma e devem responder cada questão selecionando uma figura entre seis alternativas.

Mohammed et al. (2006) usaram o TAS para avaliar as diferenças do desempenho em leitura labial de adultos com surdez profunda pré-lingual, adultos ouvintes disléxicos e adultos ouvintes sem histórico de dislexia. Segundo os autores, o desempenho em leitura labial foi diferente nos três grupos. As pessoas não ouvintes tiveram melhor desempenho do que as pessoas ouvintes sem histórico de dislexia que, por sua vez, foram melhores do que pessoas ouvintes com dislexia. Este último grupo apresentou um déficit residual no processamento da fala quando as informações fornecidas eram apenas visuais (de leitura labial). Os autores concluíram que a habilidade de leitura labial estava correlacionada com a habilidade de leitura.

Kyle et al. (2013) também desenvolveram um teste computadorizado denominado de *Test of Child Speechreading* (ToCS) (Teste de Leitura da Fala Infantil - tradução própria), com linguagem apropriada para crianças ouvintes ou com deficiência auditiva. Segundo os autores, o ToCS mede o desempenho em leitura labial de crianças em três aspectos psicolinguísticos: palavras, frases e histórias curtas.

O desenvolvimento de pesquisas com paradigmas de investigação e estímulos apropriados que influenciam na percepção da leitura labial é atual e tem sido explorado em outros idiomas além do Inglês Americano e Britânico, como o Árabe (Hegazi, Saad e Khodeir, 2021), o Chinês (Chen et al., 2022) e o Francês (Piquard-Kipffer, Cavadini, Sprenger-Charolles e Gentaz, 2021).

No Brasil, poucas investigações a respeito da leitura labial foram realizadas. Oliveira, Soares e Chiari (2014) avaliaram a habilidade de leitura labial com 61 participantes com e sem deficiência auditiva com idades entre 12 e 70 anos, por meio de três testes. O Teste de Leitura da Fala proposto por Tedesco, Chiari e Vieira (1995) é composto por três partes: a primeira parte contemplava questões cotidianas sobre identificação da pessoa e de seu contexto familiar. O participante era orientado a responder as perguntas da locutora para a avaliadora. A segunda parte do teste consistia em 44 frases representadas por meio de figuras em cartões físicos, em agrupamentos que possuíam 11 conjuntos com estrutura gramatical semelhante, divididos em três, quatro ou cinco frases. O participante era orientado a apontar as figuras da prancha que ele achava que correspondiam ao que a locutora pronunciou. Na terceira e última parte do teste, 30 palavras eram representadas por figuras em cartões físicos, que estavam divididos em seis grupos de cinco palavras. Novamente o participante era orientado a apontar para a figura que julgava corresponder ao que a locutora pronunciou. O gabarito do teste era em relação ao "acerto", "erro" e "item sem resposta". O outro teste se referia a 11 listas de sentenças do dia a

dia com dez frases com um total de 50 palavras, proveniente do Centro de Pesquisa Audiológica da Universidade de São Paulo (USP) em Bauru, para reconhecimento de frases em conjunto aberto, ou seja, os participantes não tinham alternativa de resposta. Por fim, o último teste estava relacionado ao reconhecimento de uma história, por meio de um conto denominado "A aposta". O participante ouvia e ao final deveria recontá-la.

Também no Brasil, De Martino (2005) identificou possíveis visemas para uma cabeça artificial falante do Português Brasileiro, a partir de pseudopalavras. Um locutor foi instruído a falar as pseudopalavras mostradas em uma tela de computador e, através do uso de um capacete com pontos de interesse, os autores mapearam os estímulos mais interessantes para compor um conjunto de visemas. O trabalho foi desenvolvido com a proposta de mapear os estímulos para produção da fala em uma face artificial.

Costa (2009) complementou o trabalho de De Martino (2005), ao criar animações faciais em 2D sincronizadas com a fala, a partir de imagens de visemas que dependiam do contexto fonético. Para testar o material produzido, a autora realizou um Teste de Inteligibilidade da Fala, em que o objetivo foi de avaliar a contribuição da informação visual em diferentes situações de ruído, com vídeos de faces reais e artificiais. A autora utilizou 27 pseudopalavras na estrutura CVCV (como "pepe") com fonemas consonantais e vocálicos específicos, apresentados na frase: "Ela fala (pseudopalavra de estímulo)". Os autores concluíram que o vídeo da face real teve um ganho de 38% na inteligibilidade da fala, e 24% no vídeo da animação (face artificial).

Objetivos e Delineamento da Pesquisa

Considerando os descritos e a escassez de estudos e materiais desenvolvidos apropriadamente para investigação da leitura labial do Português Brasileiro, bem como a relevância do estudo desse tema, este trabalho teve como objetivo geral desenvolver uma base de arquivos de vídeo adequada para treinamento de leitura da fala em Português Brasileiro, por meio de estímulos naturais, com palavras que contemplam diferentes fonemas.

E, para se atingir o objetivo geral foram cumpridos os seguintes objetivos específicos: mapear e selecionar palavras que contemplem diferentes fonemas do Português Brasileiro; elaborar vídeos para treinamento em leitura labial a partir das palavras selecionadas; e testar os vídeos elaborados para treinamento de leitura da fala em Português Brasileiro com pessoas ouvintes.

Trata-se de um estudo exploratório de abordagem empírica. A partir do próximo capítulo, a organização do trabalho escrito está dividida em duas partes até a discussão. A primeira parte do capítulo Método foi intitulada "Seleção das Palavras e Produção dos Vídeos

de Estímulo" consistiu na descrição do mapeamento e seleção das palavras e produção dos estímulos, que são gravações da pronúncia dessas palavras por locutores cuidadosamente selecionados e sob critérios de qualidade detalhadamente apresentados, e está relacionada aos objetivos específicos de mapear e selecionar as palavras de estímulo e produzir os vídeos para treinamento. A segunda parte, intitulada "Sessão experimental" se refere ao delineamento da sessão experimental computadorizada aplicada à distância a pessoas ouvintes, voltada para o treinamento analítico (de palavras isoladas), que gerou a coleta de dados com os estímulos e está relacionada ao objetivo específico de testar os vídeos produzidos.

A primeira parte do capítulo Resultados, denominada de "Palavras Selecionadas e Vídeos Produzidos" apresenta as palavras utilizadas para os estímulos, selecionadas a partir do *corpus* C-Oral Brasil, com suas respectivas métricas de frequência de uso e os vídeos produzidos, disponibilizados por meio de um *link* para a banca. E a segunda parte "Análises Descritiva e Exploratória da Sessão Experimental" apresenta a análise descritiva dos dados coletados na sessão experimental e análises exploratórias relacionadas ao tempo de resposta, a classificação dos fonemas consonantais nas palavras de estímulo e agrupamentos dos fonemas baseados na proporção de respostas corretas. Por fim, no capítulo da Discussão, na primeira parte (denominada "Palavras Selecionadas e Vídeos Produzidos") é realizada a discussão sobre as implicações das métricas de fala do Português Brasileiro e do uso de um *corpus* de fala ao invés de escrita para produção dos estímulos de leitura labial; é feito um paralelo entre teste e treinamento, em que são discutidos os testes de leitura labial existentes no Português Brasileiro e tarefas computadorizadas de treinamento de leitura labial em outros idiomas, que tem pontos de convergência com essa pesquisa; são apresentados trabalhos similares do Inglês que produziram estímulos para tarefas experimentais em leitura labial e do Português Brasileiro; são apresentadas as contribuições dessa pesquisa e possíveis direcionamentos futuros que podem ser provenientes dela para ampliar a investigação dessa temática no Brasil. E, na segunda parte ("Sessão Experimental"), os dados analisados são discutidos em paralelo com outras pesquisas já publicadas; são apresentadas implicações provenientes da interdisciplinaridade da pesquisa dentro dessa temática. Na conclusão são apresentados as implicações gerais desse estudo exploratório, a partir de uma retomada do propósito de realização da pesquisa, e direcionamentos futuros.

2. MÉTODO

A pesquisa foi aprovada pelo Comitê de Ética em Pesquisa da Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto (CEP/FFCLRP/USP) - CAAE nº 52025821.5.0000.5407 - conforme consta no parecer do anexo 1.

2. 1. Primeira Parte: Seleção das Palavras e Produção dos Vídeos de Estímulo

2. 1. 1. Mapeamento e Seleção das Palavras

Inicialmente, considerando a interdisciplinaridade do estudo, ressaltam-se os diversos contatos importantes realizados nesta primeira parte, especialmente com pesquisadores que englobam áreas correlatas ao desenvolvimento dessa pesquisa, principalmente relacionadas a Linguística, Psicolinguística e Fonoaudiologia. Em um primeiro momento, via e-mail com uma pesquisadora da leitura labial do Inglês Britânico, da University College London (UCL), foi solicitada autorização de uso de um relatório de informações científicas com indicações de qualidade de imagem para produção de vídeos em leitura labial. Posteriormente, um outro contato, também, com uma pesquisadora da UCL, que estuda essa mesma temática, e para a produção do material adequado e desenho do protocolo experimental de treinamento, recebi indicação do trabalho de uma pesquisadora fonoaudióloga brasileira. Assim, foi realizado contato com a fonoaudióloga brasileira, que pesquisou e trabalhou com oralização e reabilitação de fala de pessoas com deficiência auditiva, nesse contato algumas referências e delineamentos foram indicados. Também, após a busca de corpora do Português Brasileiro para seleção dos estímulos, foi realizado contato com a coordenadora do projeto C-Oral Brasil, *corpus* de base desta pesquisa. Além de contato com fonoaudiólogos durante o balanceamento dos fonemas consonantais nas palavras selecionadas a partir do *corpus*. Por fim, durante a execução do projeto, com linguistas e pesquisadores da Psicolinguística que trabalham com a área de fala para eventuais esclarecimentos de conceitos relacionados às áreas para dirimir dúvidas.

Para a seleção das palavras que compõem o conjunto dos estímulos, foi realizado um mapeamento dos *corpora* existentes do Português Brasileiro. O C-Oral Brasil (https://www.c-oral-brasil.org/) foi escolhido para a seleção das palavras por se tratar de um *corpus* de fala espontânea do Português Brasileiro. Esse *corpus* foi desenvolvido no Laboratório de Estudos Empíricos e Experimentais da Linguagem (LEEL) da Faculdade de Letras da Universidade Federal de Minas Gerais (FALE/UFMG) e é composto por duas listas de palavras oriundas de interações formais – em contextos específicos – e informais – provenientes de interações (monólogos, diálogos e conversações) em locais públicos e privados.

Mello (2012) explica que os *corpora* da língua escrita dominam a produção na área da Linguística, porém esclarece que *corpora* orais e multimodais tem se ampliado e encontram

aplicações não somente em estudos da área, mas também no desenvolvimento de tecnologias relacionadas ao reconhecimento e síntese da fala. Um *corpus* oral eletrônico possibilita acesso a métricas da língua falada para desenvolvimento de "estudos teóricos e aplicados da fala espontânea, com base empírica" (Mello, 2012, p. 33). O C-Oral Brasil foi desenvolvido para fomentar o cenário de *corpora* oral eletrônico no Brasil, que até então era muito restrito.

O C-Oral Brasil possui um total de 36.955 palavras classificadas gramaticalmente como substantivos, dessas, 55 foram selecionadas para utilização neste estudo a partir da lista do Português Brasileiro falado em ambientes informais. A seleção de palavras gramaticalmente classificadas como substantivos foi o primeiro critério estabelecido, em virtude da produção dos estímulos somente visuais e isolados, ou seja, palavras fora de sentenças, o que caracteriza um ambiente sem contexto, e devido a variação diatópica do Português Brasileiro, de sotaque por região geográfica. Os demais critérios para seleção foram relacionados a estrutura da palavra ser Consoante-Vogal (CV) na fala, considerando dígrafos consonantais, que são duas consoantes que representam um fonema, como o ss, ch, nh, lh, rr correspondentes aos fonemas consonantais /s/, /ʃ/, /ɲ/, /ʎ/, /R/, por exemplo em "massa", "chuva", "manhã", "folha" e "carro", e palavras dissílabas.

Paralelamente à seleção, um balanceamento dos fonemas consonantais foi realizado. No Português Brasileiro, 16 fonemas consonantais podem ocupar o início da palavra (sendo /r/, /ɲ/ e /ʎ/, aqueles que não ocupam o início) e 19 fonemas consonantais podem ocupar o meio da palavra. Os fonemas foram balanceados respeitando essa característica, no início da palavra - onset (primeira sílaba) e em posição intervocálica (segunda sílaba). Neste estudo, não foram utilizadas palavras de estímulo com encontros consonantais, palavras com a presença de dois fonemas consonantais juntos (como o BL em "blusa" e PR em "prato"), ou encontros vocálicos (como OE e IA em "poesia").

O balanceamento dos fonemas consonantais foi realizado considerando o inventário fonético apresentado por Issler (1996) (tabela 1). Cabe ressaltar que, de acordo com Silva (2011), a fonética diz respeito aos fenômenos físicos da fala e tem por unidade mínima os traços distintivos de sonoridade, modo e ponto de articulação. Já a fonologia tem por base o fonema, como unidade sem significado, mas com função distintiva, para determinar o significado de uma palavra em detrimento de outra, como os fonemas /s/ e /z/ em "caça" e "casa", respectivamente. O inventário apresentado na tabela 1 demonstra as classificações dos fonemas consonantais do Português Brasileiro baseados em modo de articulação (em azul), ponto de articulação (em cinza) e sonoridade (em verde).

Tabela 1 - Inventário Fonético Consonantal

Papel das Cavidades		CONSOANTES ORAIS							Consoantes Nasais	Semivogais
Modo de Articulação		Plosivas		CONSTRITIVAS					Oclusivas Nasais	
				Fricativas		Laterais	Vibrantes			
							Simples	Múltiplas		
Papel das Cordas Vocais		surdas	sonoras	surdas	sonoras	sonoras	sonoras		sonoras	
Ponto de Articulação	Bilabiais	/p/	/b/						/m/	
	Labiodentais			/f/	/v/					
	Linguodentais	/t/	/d/						/n/	
	Alveolares			/s/	/z/	/l/	/r/			
	Palatais			/ʃ/	/ʒ/	/ʎ/			/ɲ/	/y/
	Velares	/k/	/g/					/R/		/w/

Fonte: Issler, 1996.

A tabela 2 apresenta as 55 palavras selecionadas e balanceadas a depender da posição do fonema alvo, por exemplo, a palavra fogo (destacada em negrito) posicionada referente ao fonema /f/ na primeira sílaba e ao fonema /g/ na segunda sílaba.

2. 1. 2. Produção dos Vídeos de Estímulo

Após a seleção das palavras, os estímulos (vídeo do rosto do locutor falando as palavras selecionadas) foram gravados em ambiente controlado e com fundo neutro, na sala de Percepção Visual do Laboratório de Psicofísica e Percepção, que fica localizado na Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da Universidade de São Paulo (FFCLRP/USP). A gravação atendeu a um conjunto de diretrizes relevantes para produção dos vídeos em leitura labial, a partir de estudos já realizados nessa temática (tabela 3 – elaborado pela autora com base no material disponibilizado e de uso autorizado por Campbell e Mohammed (2010)).

Os vídeos de estímulo tiveram as seguintes especificações técnicas: gravados com a Câmera Semiprofissional Canon Rebel T3i – EOS 600D com Cartão SDXC 128Gb SanDisk Extreme Pro 170Mb/s 4K UHS-I / V30 / U3 Classe 10, com a configuração de 60 Frames Por Segundo (FPS) e qualidade *High Definition* (HD) (1280 x 720), com a câmera em posição vertical, com o rosto e o tronco dos locutores enquadrados, e posicionamento de câmera igual para os dois locutores. Sendo a incidência de luminância no rosto dos locutores de aproximadamente 39,97 cd/m² para o locutor 1 e 43,58 cd/m² para o locutor 2, sem sombra no rosto ou atrás dos locutores.

Tabela 2 - Lista de Palavras Selecionadas e Balanceadas por Fonema Consonantal

Palavras selecionadas organizadas pela posição do fonema consonantal									
Fonema	**Primeira sílaba**					**Segunda sílaba**			
/b/	bife	boca	bicho			lobo	rabo	tubo	
/d/	dado	doce	dica			dado	roda	vida	
/f/	folha	furo	**fogo**	filho		bife	café	sofá	
/g/	galo	gato				**fogo**	lago	vaga	jogo
/k/	café	copo	cura	casa	carro	boca	dica	suco	
/l/	lobo	lixo	lago	luta		pulo	ralo	galo	gelo
/p/	palha	pano	povo	pulo		copo	sopa	tipo	
/R/	rabo	rosa	ralo	roda	ramo	carro			
/r/						muro	furo	cura	
/ʃ/	chuva					lixo	taxa	bicho	
/s/	soma	suco	sopa	sofá	sono	massa	doce	taça	gesso
/t/	taça	tipo	taxa	tubo		moto	gato	luta	
/v/	vaga	vida	vaso	vinho		neve	povo	chuva	
/z/	zona					casa	rosa	vaso	
/m/	muro	moto	massa	manhã		nome	soma	ramo	
/n/	neve	nome	ninho			pano	zona	sono	
/ʒ/	gelo	jogo	gesso						
/ɲ/						manhã	vinho	ninho	
/ʎ/						folha	filho	palha	

Fonte: elaborado pela autora, 2024.

Ainda conforme as diretrizes apresentadas por Campbell e Mohammed (2010), os seguintes critérios foram estabelecidos para inclusão dos locutores voluntários: possuir como língua nativa o Português Brasileiro, ser alfabetizado, ter entre 18 e 55 anos de idade, não apresentar deficiência auditiva em nenhum grau, não ter obstruções na região da boca, não fazer uso de aparelhos ou *piercings* na região externa ou interna da cavidade oral, estar em bom estado de saúde na época de gravação dos vídeos (em virtude da pandemia da Covid-19) e ter boa articulação (sem prejuízos ou obstrução) na região da cavidade oral. Dois fonoaudiólogos que cumprem os critérios de inclusão foram convidados para participar do estudo e aceitaram contribuir com a pesquisa como locutores voluntários. Os locutores voluntários assinaram o Termo de Autorização do Uso de Imagem e receberam as orientações para participação via e-mail institucional.

O locutor voluntário 1 é um homem "cisgênero", que na época das gravações estava com 29 anos de idade, fonoaudiólogo, natural de São Joaquim da Barra (São Paulo), com sotaque paulistano. Para as gravações, o locutor voluntário estava com blusa preta, sem óculos ou adereços na cabeça e tronco, sem barba, tendo cabelo curto castanho escuro, pele clara e

altura de 1,77 m. Já a locutora voluntária 2 é uma mulher "cisgênero", que na época das gravações estava com 25 anos de idade, fonoaudióloga, natural de São José do Rio Preto (São Paulo), com sotaque paulistano. Para as gravações a locutora voluntária estava com blusa preta, cabelo solto posicionado atrás dos ombros, sem óculos ou adereços na cabeça e tronco, tendo cabelo longo castanho escuro, pele clara e altura de 1,68 m.

Tabela 3 - Diretrizes para Produção de Vídeos de Leitura Labial

Iluminação	Iluminação superior e frontal, mais parecida com a iluminação natural (luz do dia) Mudanças de luminância na faixa de 1 a 120 cd/m² Sem sombras na região da boca
Visualização	Visualização completa do rosto Não há estudos que indiquem diferenças significativas na percepção entre imagens 3D e 2D
Cor	Não há estudos que indiquem que imagens com escala de cor diferem na leitura labial em comparação com imagens em escala de cinza
Distância	É sugerido que a cabeça e o tronco do locutor sejam capturados na imagem, ao invés de apenas rosto ou boca, a fim de auxiliar a interpretação do observador com informações não orais do locutor
Percepção	Leitores de fala pousam o olhar na região dos olhos do locutor Boa parte das informações usadas na leitura da fala não só depende da visão central, mas também da visão periférica
Características de imagem	Dimensão / taxa de quadros (FPS) de pelo menos 352 x 288 pixels e 25 FPS Taxa mínima: é possível distinguir o estímulo com uma taxa reduzida de 12 FPS, mas não é recomendado Uma taxa maior, com cerca de 120 FPS (mais qualidade de imagem), pode indicar características distintas da fala que podem ser observadas na leitura da fala por observadores não treinados
Escolha do locutor	A maioria dos testes de leitura de fala possui locutores barbeados, falando em voz normal e em tom neutro Não há evidências para diferenças de locutor com base na idade, sexo, cor da pele ou etnia O melhor indicador da legibilidade da fala de um locutor é a familiaridade tanto com o locutor quanto com seu estilo de fala e sotaque
Aspectos gerais	Não são indicadas edições/alterações na amplitude visual da imagem da boca, como aumento ou redução da taxa de quadros, a fim de capturar padrões. Essas alterações podem atrapalhar a percepção do observador

Fonte: Campbell e Mohammed (2010) adaptado pela autora, 2024.

Ao total, cada locutor gravou 55 vídeos referentes a cada uma das palavras selecionadas. Os vídeos produzidos foram editados pelo programa Filmora para selecionar o *take* completo (boca fechada, movimento/pronúncia completa da palavra e boca fechada novamente), rotacionar os vídeos, remover o som e exportar o vídeo mantendo a qualidade indicada.

2. 2. Segunda Parte: Sessão Experimental

2. 2. 1. Elaboração da Sessão Experimental e Coleta de Dados

A última parte do trabalho visava testar os vídeos produzidos com pessoas ouvintes. O enfoque era incorporar os vídeos em uma tarefa experimental de treinamento analítico, que pudesse estar relacionada com a identificação dos fonemas alvo dentro das palavras e estes pudessem ser incorporados em uma base de dados criada para estudos da percepção visual da fala do Português Brasileiro.

2. 2. 2. Participantes

Ao todo, 330 participantes responderam a sessão experimental, sendo que 281 preencheram os critérios de inclusão: possuíam como língua nativa e mais usada o Português Brasileiro, eram alfabetizados, tinham entre 18 e 55 anos de idade, não apresentavam deficiência auditiva diagnosticada e, se tinham problemas na visão, faziam uso de lentes corretivas, e tinham acesso a um dispositivo digital com internet (como celular, computador ou tablet).

Foram excluídos do estudo os dados de 49 participantes que não possuíam como língua nativa o Português Brasileiro, que não usavam esse idioma como mais frequente de fala/escrita/escuta/leitura para se comunicar (n=5), não alfabetizados, menores de 18 anos ou maiores de 55 anos (n=22), que apresentassem deficiência auditiva diagnosticada em algum grau (n=2), que não fizessem uso de lentes corretivas, se tivessem problemas na visão (n=18), que não tinham acesso a um dispositivo digital ou não concordaram em participar do estudo (n=2).

2. 2. 3. Caracterização da Amostra

A idade mediana dos participantes foi de 26 anos, visto que 50% da amostra tinha idade entre 24 e 34 anos. A maioria dos participantes informou ser do sexo feminino – 60,9% (171 participantes) – e 38,1% (107 participantes) do sexo masculino, 2 participantes não binários e um que preferiu não informar. A análise dos dados não foi norteada pela idade ou gênero dos participantes. A amostra foi composta por 69% de participantes que, na época da sessão experimental, possuíam ensino superior completo, estavam cursando ou já haviam finalizado uma pós-graduação (tabela 4).

A maioria dos participantes reside na região Sudeste (57,6%), seguida pelo Centro-Oeste (33,8%), com a maior parte proveniente dos estados de São Paulo (40,2%) e Mato Grosso do Sul (30,6%) (tabela 5). Ainda, 64% da amostra informou ter miopia ou astigmatismo, mas que fazem uso de lentes corretivas.

Tabela 4 - Escolaridade dos Participantes da Pesquisa

Grau de escolaridade	Número de participantes (%)
Ensino Fundamental Completo	1 (0,4%)
Ensino Médio Completo	12 (4,3%)
Ensino Superior (cursando)	74 (26,3%)
Ensino Superior Completo	58 (20,6%)
Pós-Graduação (cursando)	57 (20,3%)
Pós-Graduação Completa	79 (28,1%)

Fonte: elaborado pela autora, 2024.

Tabela 5 - Região e Estado de Residência dos Participantes da Pesquisa

Região	Estado de Residência	N (%)
Centro-Oeste	Distrito Federal	7 (2,5%)
	Goiás	2 (0,7%)
	Mato Grosso do Sul	86 (30,6%)
Nordeste	Bahia	1 (0,4%)
	Ceará	2 (0,7%)
	Paraíba	1 (0,4%)
Norte	Pará	1 (0,4%)
Sudeste	Minas Gerais	36 (12,8%)
	Rio de Janeiro	13 (4,6%)
	São Paulo	113 (40,2%)
Sul	Paraná	4 (1,4%)
	Rio Grande do Sul	6 (2,1%)
	Santa Catarina	9 (3,2%)

Fonte: elaborado pela autora, 2024.

2. 2. 4. *Design* da Sessão Experimental

A coleta de dados foi realizada de forma *online* e assíncrona, ou seja, os participantes podiam responder a sessão experimental sem contato direto com a pesquisadora. Por meio de um *link* do Formulários Google, os participantes acessavam o questionário sociodemográfico e o Termo de Consentimento Livre e Esclarecido (TCLE), via dispositivo móvel ou computador. Ao completar o questionário e responder o TCLE, o *link* para participar da sessão experimental era disponibilizado na tela, com a indicação da necessidade de um dispositivo com *mouse* e teclado para responder a segunda parte. Com o dispositivo adequado, ao acessar o *link*, o participante era redirecionado para iniciar a sessão experimental.

A sessão experimental consistiu em um protocolo experimental específico para testar os vídeos produzidos visando a identificação do fonema em um contexto de competição lexical com possíveis competidores reais (palavras próximas com troca de fonema), fomentada por uma tarefa de treinamento que propiciava *feedback* quando a palavra-alvo não era indicada corretamente.

Para estimular a participação e obter uma amostra apropriada, 30 dos 110 vídeos produzidos foram selecionados para compor a sessão experimental ainda respeitando o balanceamento dos fonemas consonantais nas palavras selecionadas. Dessa forma, a sessão tinha duração aproximada de 15 minutos, para que a coleta de dados e a atenção do participante à tarefa não fossem prejudicadas. Como o intuito não era comparar o desempenho na tarefa com dois locutores diferentes, somente os vídeos com a locutora voluntária 2 foram usados para compor a sessão experimental. A tabela 6 apresenta as palavras de estímulo que foram usadas na sessão experimental, novamente balanceadas a depender da posição do fonema alvo, por exemplo, a palavra fogo (destacada em negrito) posicionada referente ao fonema /f/ na primeira sílaba e ao fonema /g/ na segunda sílaba. Na sessão experimental, 16 fonemas foram contemplados na primeira sílaba e 18 fonemas consonantais na segunda sílaba.

A sessão experimental foi construída por meio do PsychoPy 3 e hospedada no Pavlovia. O PsychoPy 3 é uma plataforma aberta com interface para construir e rodar experimentos em ciência do comportamento – https://www.psychopy.org/ – vinculada ao Pavlovia, um repositório *online* que hospeda os experimentos em ciência do comportamento que podem ser acessados ao redor do mundo – https://pavlovia.org/. Ambos são iniciativas da Universidade de Nottingham.

Ao iniciar a sessão experimental, uma tela de boas-vindas aparecia e o participante deveria indicar que estava pronto para começar ao clicar na barra de espaço do teclado do dispositivo que estava usando. Em seguida, um vídeo de um minuto com as instruções da tarefa iniciava automaticamente. O participante tinha a opção de repetir o vídeo de instrução se quisesse, caso contrário, iniciava a tarefa. Em cada tentativa, o vídeo de estímulo era repetido por duas vezes seguidas com a locutora voluntária falando a palavra-alvo (por exemplo: “fato”). Em seguida, aparecia uma tela com três alternativas clicáveis, a palavra-alvo por escrito, por exemplo, “fato”, e duas alternativas distratoras, também, por escrito com troca de fonema consonantal na primeira sílaba, por exemplo, “gato” e segunda sílaba, “faro”. Os participantes indicavam uma resposta e recebiam *feedback* instantâneo para cada tentativa, novamente por escrito. Caso acertassem a palavra-alvo, o *feedback* era “A resposta está certa”, caso contrário, a resposta alvo era indicada para o participante “A resposta certa é fato”. Para iniciar o vídeo

de estímulo seguinte era necessário pressionar a barra de espaço para indicar que estava pronto para uma nova tentativa.

Tabela 6 - Palavras de Estímulo Utilizadas na Sessão Experimental

Fonema	**Primeira Sílaba**			**Segunda Sílaba**		
/b/	bife	boca		rabo		
/d/	dado	doce		dado	roda	
/f/	folha	**fogo**		bife	café	
/g/	gato			**fogo**	jogo	
/k/	café	copo	carro	boca	suco	
/l/	lixo	luta		pulo	gelo	
/p/	pano	pulo		copo	tipo	
/R/	rabo	roda		carro		
/r/				muro	furo	
/ʃ/	chuva			lixo		
/s/	soma	suco		taça	doce	
/t/	taça	tipo		moto	gato	luta
/v/	vaso	vinho		neve	chuva	
/z/	zona			casa	vaso	
/m/	muro	moto		nome	soma	
/n/	neve	nome		pano	zona	
/ʒ/	gelo	jogo				
/ɲ/				vinho		
/ʎ/				folha		

Fonte: elaborado pela autora, 2024.

Os vídeos que compuseram a sessão experimental foram apresentados como estímulos exclusivamente visuais (imagem sem o som) e isolados, ou seja, nenhuma dica foi adicionada durante a execução do vídeo. A sequência dos vídeos de estímulo e a posição das alternativas eram randomizadas, de forma que não apareciam na mesma ordem para os participantes, em virtude de evitar que o processo atencional, o cansaço ou a interrupção por fatores externos durante a sessão fossem atrelados a estímulos específicos, como por exemplo, o primeiro e o último. O tempo de resposta era coletado a partir do momento em que as alternativas apareciam na tela até o momento em que o participante clicava em uma resposta.

2. 2. 5. Estabelecimento das Alternativas Distratoras

A sessão experimental foi uma tarefa de escolha forçada, ou seja, os participantes indicavam a resposta por meio de alternativas impressas na tela. Para definir as alternativas distratoras na tarefa foram estabelecidas trocas dos fonemas consonantais, baseados nas classificações apresentadas no Inventário Fonético de Issler (1996) e que configuravam possíveis competidores fonológicos e lexicais, obedecendo aos critérios de não corresponderem

a pseudopalavras/logatomas e possuírem a mesma estrutura dissílaba e de Consoante-Vogal na fala. A tabela 7 apresenta a transcrição fonológica da palavra-alvo (resposta certa) e as alternativas distratoras (1 e 2).

Tabela 7 - Transcrição Fonológica das Palavras de Estímulo e Distratoras

Palavra de Estímulo	Distrator 1	Distrator 2
/tipo/	/sipo/	/tiko/
/nome/	/fome/	/nove/
/kaRo/	/ʒaRo/	/karo/
/ʒogo/	/fogo/	/ʒoRo/
/dado/	/lado	/dano/
/luta/	/ʒuta/	/lupa/
/kafe/	/Rafe/	/kale/
/boka/	/foka/	/boʎa/
/foʎa/	/RoʎaI	/fora/
/ʃuva/	/luva/	/ʃuta/
/fogo/	/ʒogo/	/foto/
/dose/	/pose/	/doze/
/moto/	/foto/	/moRo/
/zona/	/lona/	zoRa/
/Roda/	/moda/	/RoʃaI
/kopo/	/topo/	/kolo/
/pano/	/kano/	/pato/
/muro/	/furo/	/mudo/
/viɲo/	/liɲo/	/vivo/
/bife/	/blefe/	/bile/
/gato/	/fato/	/gaʎo/
/ʒelo/	/zelo/	/ʒeso/
/liʃo/	/biʃo/	/lizo/
/suko/	/muko/	/suʒo/
/vazo/	/kazo/	/vago/
/neve/	/leve/	/nele/
/soma/	/goma/	/sopa/
/pulo	/nulo/	/puɲo/
/tasa/	/masa/	/tala/
/Rabo/	/kabo/	/Ralo/

Fonte: elaborado pela autora, 2024.

3. RESULTADOS

3. 1. Primeira Parte: Palavras Selecionadas e Vídeos produzidos

A tabela 8 apresenta as 55 palavras selecionadas, com suas respectivas frequências de uso, disponibilizadas no site do C-ORAL Brasil (http://www.c-oral-brasil.org/livro.php). As frequências são baseadas nos *tokens*, que corresponde à frequência de ocorrência da palavra no *corpus* como um todo.

Tabela 8 - Frequência de Uso das Palavras de Estímulo

Frequência	Palavra	Frequência	Palavra	Frequência	Palavra	Frequência	Palavra	Frequência	Palavra
25	bicho	10	doce	10	lixo	13	pano	13	sono
9	bife	130	filho	10	lobo	23	povo	9	suco
32	boca	25	fogo	2	luta	1	pulo	8	taça
36	café	23	folha	34	massa	2	rabo	1	taxa
146	carro	3	furo	2	manhã	2	ralo	299	tipo
322	casa	4	galo	30	moto	6	ramo	2	tubo
23	chuva	10	gato	10	muro	13	roda	8	vaga
22	copo	1	gesso	2	neve	5	rosa	5	vaso
2	cura	3	gelo	3	ninho	17	sofá	107	vida
5	dado	54	jogo	100	nome	4	soma	13	vinho
7	dica	3	lago	13	palha	2	sopa	8	zona

Fonte: elaborado pela autora, 2024.

Um total de 110 vídeos foram produzidos em qualidade *high definition* (HD) com 60 *frames per second* (FPS) e estão sem som. Esses vídeos compõem a base de dados resultante deste trabalho e estão disponíveis pelo link (com acesso permitido a partir de solicitação à pesquisadora): https://github.com/VidalFernanda/VideosProduzidosBasedeArquivos.git

3. 2. Segunda Parte: Análises Descritiva e Exploratória da Sessão Experimental

Além dos dados sociodemográficos, a resposta do participante e o tempo de resposta foram coletados. Os dados da sessão experimental foram analisados por meio do *software* R na versão 4.3.1, pela interface do RStudio. As análises se dividiram em quatro momentos: (1) de estatística descritiva, que apresenta as proporções de respostas e o tempo médio de resposta por tipo de resposta; que motivou análises exploratórias (2) da correlação entre o tempo de resposta pelo tipo de resposta por meio de Kendall, um teste não paramétrico para observar se dois postos estão correlacionados, (3) a análise da proporção de respostas corretas em relação as classificações dos fonemas consonantais entre a palavra-alvo e distratoras, comparadas por meio do método de Newcombe-Wilson; (4) e a análise de agrupamento (*cluster*) dos fonemas pela proporção de acerto, baseado no algoritmo *k-means*.

A proporção de respostas por tipo de resposta (correta ou incorreta) foram calculados por palavra (tabela 9). A resposta "correta" se refere à quando os participantes indicaram corretamente a palavra-alvo (estímulo), seguida pela proporção de respostas incorretas da palavra distratora em que a troca do fonema consonantal estava na primeira sílaba e a proporção de respostas incorretas da palavra distratora em que a troca de fonema consonantal estava na segunda sílaba. Os dados estão organizados na tabela por ordem alfabética das palavras de estímulo.

Quando a resposta indicada foi incorreta, as palavras de estímulo "neve", "rabo", "dado", "tipo", "café" , "moto" e "bife" apresentaram maior confusão em relação a palavra distratora com troca do fonema consonantal na primeira sílaba. As demais palavras tiveram uma proporção de respostas maior em palavras com troca do fonema consonantal na segunda sílaba, sendo que o item "soma" foi o que apresentou maior confusão, com 14,9% de respostas corretas e 83,3% de respostas em relação a alternativa distratora. As palavras com as menores proporções de respostas corretas foram "muro" (62,6%), "suco" (54,8%), "zona" (51,2%), "neve" (39,1%) e "soma" (14,9%), enquanto "chuva" (96,8%), "pulo" (91,8%), "roda" (89,7%) e "pano" (89,3%) tiveram as maiores proporções de respostas corretas.

As médias e o intervalo de confiança (IC) de 95% para os tempos de resposta foram calculados por palavra e por tipo de resposta (tabela 10). Os ICs sinalizados por "(-)" se referem a palavras que naquela alternativa só tiveram uma resposta. A média total do tempo de resposta foi de 3,33 segundos (95% IC: 3,00; 3,67) para as respostas corretas; de 6,03 segundos (95% IC: 5,19; 6,88) para respostas incorretas com troca de fonema na primeira sílaba; e de 5,19 segundos (95% IC: 4,56; 5,82) para respostas incorretas com troca de fonema na segunda sílaba.

Devido a grande quantidade de dados, um gráfico Box-plot (gráfico 1) foi gerado por palavra de estímulo para visualização geral desses dados descritivos. Cabe ressaltar que a apresentação do gráfico se refere somente a descrição dos dados e o Box-plot utiliza como base a mediana ao invés da média. Foi realizada a transformação logarítmica do tempo de resposta (eixo Y) para melhor observação desses dados. Em cada palavra é apresentado o tipo de resposta (da esquerda para direita do leitor: correta (em vermelho), incorreta 1 – referente a troca de fonema na 1ª sílaba (em verde) e incorreta 2 referente a troca de fonema na 2ª sílaba (em azul)) e, também, os *outliers* – participantes que divergiram no tempo de resposta em relação aos demais – indicados pelos pontos em preto. Os *outliers* podem indicar possíveis interrupções externas ou travamento durante a execução do vídeo, uma vez que essas são consequências pertinentes a uma tarefa remota computadorizada. Por isso, o número amostral alto e o detalhamento dos *outliers* auxilia na observação de limitações passíveis desse tipo de

coleta. Também, cabe ressaltar que há um *delay* no registro do tempo de resposta, em virtude do uso de uma ferramenta virtual.

Tabela 9 - Proporção de Resposta por Tipo de Resposta

Palavra de Estímulo	Proporção de respostas corretas (em %)	Proporção de respostas com troca na 1ª sílaba (em %)	Proporção de respostas com troca na 2ª sílaba (em %)
bife	87,2%	11,4%	1,4%
boca	67,6%	2,1%	30,2%
café	86,8%	11,0%	2,1%
carro	86,8%	4,3%	8,9%
chuva	96,8%	1,4%	1,8%
copo	89,0%	6,8%	4,3%
dado	82,2%	13,2%	4,6%
doce	77,9%	5,3%	16,7%
fogo	80,1%	1,8%	18,1%
folha	86,8%	1,8%	11,4%
gato	80,1%	1,8%	18,1%
gelo	72,6%	3,2%	24,2%
jogo	73,3%	2,5%	24,2%
lixo	81,9%	3,6%	14,6%
luta	82,9%	3,9%	13,2%
moto	73,3%	14,2%	12,5%
muro	62,6%	10,0%	27,4%
neve	39,1%	59,1%	1,8%
nome	77,2%	7,8%	14,9%
pano	89,3%	1,4%	9,3%
pulo	91,8%	2,8%	5,3%
rabo	73,3%	23,5%	3,2%
roda	89,7%	3,9%	6,4%
soma	14,9%	1,8%	83,3%
suco	54,8%	1,8%	43,4%
taça	89,3%	4,6%	6,0%
tipo	85,4%	10,3%	4,3%
vaso	82,2%	0,4%	17,4%
vinho	87,5%	0,4%	12,1%
zona	51,2%	5,7%	43,1%
Total	**76,45%**	**7,39%**	**16,14%**

Fonte: elaborado pela autora, 2024

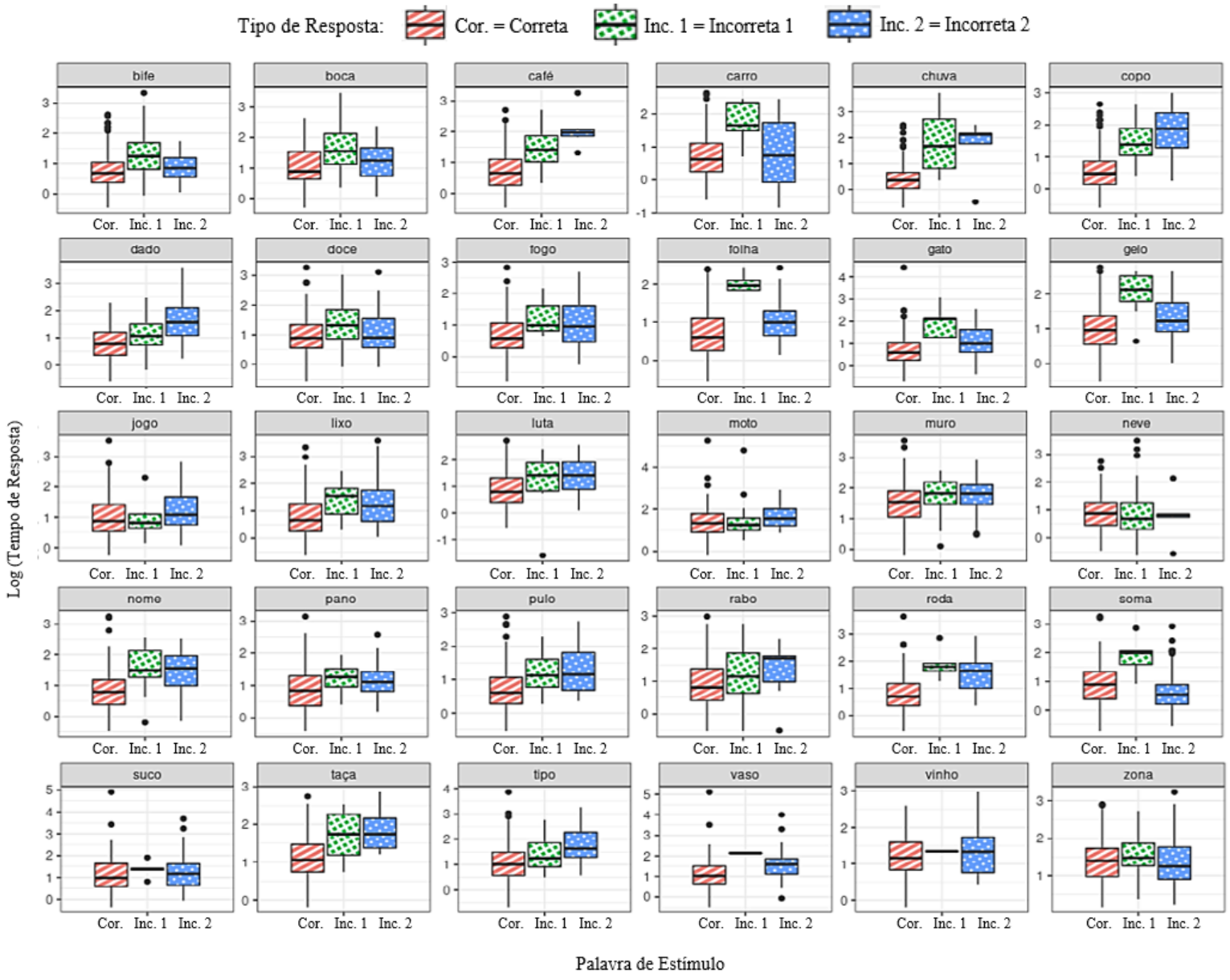
Tabela 10 - Média (95% IC) do Tempo de Resposta por Palavra e Tipo de Resposta

Palavra de Estímulo	Média do tempo de resposta para respostas corretas (95% IC) (segundos)	Média do tempo de resposta para respostas com troca de fonema na 1ª sílaba (95% IC) (segundos)	Média do tempo de resposta para respostas com troca de fonema na 2ª sílaba (95% IC) (segundos)
bife	2,62 (2,36;2,87)	5,06 (3,21;6,91)	2,89 (0,91;4,86)
boca	3,60 (3,22;3,97)	9,28 (0,01;18,55)	4,00 (3,52;4,48)
café	2,61 (2,35;2,87)	4,93 (3,89;5,97)	9,72 (3,22;16,23)
carro	2,62 (2,34;2,90)	6,65 (4,62;8,67)	3,40 (2,16;4,64)
chuva	1,77 (1,59;1,94)	14,17 (-4,30;32,63)	7,15 (3,39;10,91)
copo	2,16 (1,94;2,38)	5,16 (3,65;6,67)	7,65 (4,59;10,71)
dado	2,65 (2,43;2,87)	3,73 (2,85;4,61)	7,76 (2,78;12,74)
doce	3,26 (2,89;3,63)	6,09 (3,09;9,09)	3,94 (2,87;5,02)
fogo	2,42 (2,17;2,67)	4,14 (1,63;6,65)	3,87 (2,95;4,78)
folha	2,48 (2,25;2,71)	7,83 (5,94;9,71)	3,29 (2,52;4,05)
gato	2,82 (2,06;3,57)	9,09 (2,57;15,60)	3,69 (2,95;4,42)
gelo	3,28 (2,96;3,61)	8,85 (5,99;11,71)	4,53 (3,82;5,24)
jogo	3,42 (2,97;3,87)	3,34 (1,07;5,60)	4,21 (3,47;4,94)
lixo	3,14 (2,70;3,57)	4,93 (2,92;6,93)	5,74 (3,46;8,02)
luta	3,15 (2,81;3,49)	4,79 (2,75;6,82)	4,93 (3,89;5,97)
moto	5,51 (3,65;7,36)	6,98 (1,10;12,86)	6,14 (4,87;7,40)
muro	5,64 (5,00;6,27)	6,46 (5,30;7,62)	6,91 (6,12;7,71)
neve	3,17 (2,69;3,65)	3,07 (2,50;3,63)	3,16 (0,44;5,88)
nome	2,97 (2,58;3,36)	5,84 (4,37;7,30)	4,99 (4,14;5,85)
pano	3,17 (2,83;3,50)	3,89 (1,65;6,14)	3,84 (2,89;4,79)
pulo	2,55 (2,28;2,82)	4,14 (2,11;6,17)	5,10 (2,79;7,41)
rabo	3,22 (2,84;3,60)	4,32 (3,52;5,13)	4,99 (3,06;6,92)
roda	2,82 (2,44;3,19)	6,76 (4,53;8,98)	5,83 (3,90;7,75)
soma	3,98 (2,35;5,62)	8,03 (3,01;13,05)	2,24 (1,99;2,48)
suco	4,76 (3,01;6,51)	4,18 (2,75;5,61)	4,49 (3,64;5,33)
taça	3,58 (3,30;3,87)	6,60 (4,57;8,64)	7,17 (5,21;9,12)
tipo	3,89 (3,36;4,42)	4,93 (3,64;6,22)	7,78 (3,89;11,67)
vaso	4,32 (2,90;5,74)	8,48 (-)	6,46 (4,18;8,75)
vinho	3,81 (3,56;4,07)	3,82 (-)	5,28 (3,89;6,66)
zona	4,75 (4,24;5,26)	5,54 (3,85;7,23)	4,64 (4,02;5,26)

Fonte: elaborado pela autora, 2024

Gráfico 1 - Box Plot do Tempo de Resposta por Tipo de Resposta com os *Outliers*

Legenda do gráfico: considerando os limites extremos superiores (linha superior no gráfico) 3º Quartil (Q3) - (Q3 + 1,5 [Q3 - Q1]) - e inferiores (linha inferior no gráfico) 1º Quartil (Q1) - (Q1 – 1,5 [Q3 – Q1]) dos gráficos, Q3 se refere ao valor observado que divide os dados entre 75% menores valores observados e 25% dos maiores valores observados e Q1 se refere ao valor observado que divide os dados entre os 25% menores valores observados e 75% dos maiores valores observados, dispostos por palavra e na sequência da esquerda para a direita de correta, incorreta 1 e incorreta 2.



Fonte: elaborado pela autora, 2024.

As observações da análise descritiva motivaram novas análises exploratórias sobre a correlação entre o tempo de resposta e o tipo de resposta. O nível de significância de 5% foi considerado para todas as análises realizadas. Os coeficientes de correlação posto-ordem de Kendall foram calculados para a média do tempo de resposta e a frequência da resposta por tipo de resposta e a significância estatística dos coeficientes de correlação estimados foi verificada (tabela 11). Observa-se uma correlação negativa em que quanto maior a quantidade de

respostas, menor o tempo de resposta e uma correlação positiva quando quanto maior a quantidade de respostas, maior o tempo de resposta.

Foram verificadas correlações negativas significativas do tempo de resposta e respostas corretas em 23 das 30 palavras (“bife”, “boca”, “café”, “carro”, “chuva”, “copo”, “dado”, “fogo”, “folha”, “gato”, “gelo”, “jogo”, “lixo”, “luta”, “muro”, “nome”, “pano”, “pulo”, “rabo”, “roda”, “taça”, “tipo” e “vaso”), ou seja, quanto maior a frequência de respostas corretas, menor foi o tempo de resposta. Apenas no item “soma” foi observada correlação positiva significativa, o que indica que para esse item quanto maior foi o tempo de resposta, maior foi a frequência de resposta correta. Também foram observadas correlações positivas significativas do tempo de resposta e respostas incorretas com troca de fonema na primeira sílaba (indicado na tabela como Incorretas 1) em 17 das 30 palavras (“bife”, “café”, “carro”, “chuva”, “copo”, “dado”, “folha”, “gato”, “gelo”, “lixo”, “nome”, “pulo”, “rabo”, “roda”, “soma”, “taça” e “tipo”), ou seja, quanto maior a frequência de resposta, maior o tempo de resposta. Não houve palavras em que o coeficiente de correlação estimado foi negativo e significativo em relação ao tempo de resposta e repostas incorretas com troca na primeira sílaba. O mesmo foi observado para o tempo de resposta e as respostas incorretas com troca de fonema na segunda sílaba (indicado na tabela como Incorretas 2), em 21 das 30 palavras houve correlação significativa positiva (“café”, “chuva”, “copo”, “dado”, “fogo”, “folha”, “gato”, “gelo”, “jogo”, “lixo”, “luta”, “moto”, “muro”, “nome”, “pano”, “pulo”, “roda”, “soma”, “taça”, “tipo” e “vaso”).

Não houve palavras em que o coeficiente de correlação foi negativo e significativo para a troca de fonema na segunda sílaba. No entanto, cabe ressaltar que essas correlações, apesar de significativas, foram fracas, ou seja, próximas de zero, isso pode ter acontecido, pois na correlação de Kendall as propriedades estatísticas são mais robustas e os valores observados geralmente são menores do que outros tipos de correlação. Kendall lida com pares concordantes e discordantes e não baseados em desvio, como é o caso de Spearman, por exemplo.

As palavras de estímulo da sessão experimental foram separadas em quatro categorias (tabela 12) que se referiam à relação dos fonemas consonantais a nível silábico entre a palavra-alvo e as palavras distratoras, com base no inventário fonético de Issler (1996): “mesmo modo de articulação”, “mesmo ponto de articulação”, “mesmo ponto ou modo de articulação” e “ponto e modo de articulação diferentes”.

Tabela 11 - Correlação entre Tempo de Frequência de Resposta por Tipo de Resposta

Palavra-alvo	Corretas	p-valor	Incorretas 1	p-valor	Incorretas 2	p-valor
bife	-0,200	**0,001**	0,205	**0,001**	0,015	0,766
boca	-0,116	**0,018**	0,071	0,145	0,095	0,051
café	-0,307	**0,001**	0,252	**0,001**	0,172	**0,001**
carro	-0,112	**0,021**	0,209	**0,001**	-0,014	0,767
chuva	-0,159	**0,001**	0,115	**0,019**	0,109	**0,026**
copo	-0,324	**0,001**	0,238	**0,001**	0,207	**0,001**
dado	-0,180	**0,001**	0,106	**0,031**	0,158	**0,001**
doce	-0,082	0,093	0,094	0,054	0,034	0,481
fogo	-0,173	**0,001**	0,088	0,070	0,149	**0,002**
folha	-0,200	**0,001**	0,173	**0,001**	0,141	**0,004**
gato	-0,209	**0,001**	0,149	**0,002**	0,166	**0,001**
gelo	-0,221	**0,001**	0,175	**0,001**	0,158	**0,001**
jogo	-0,134	**0,006**	-0,016	0,749	0,144	**0,003**
lixo	-0,200	**0,001**	0,107	**0,029**	0,162	**0,001**
luta	-0,201	**0,001**	0,083	0,090	0,176	**0,001**
moto	-0,075	0,126	-0,043	0,381	0,145	**0,003**
muro	-0,185	**0,001**	0,061	0,215	0,160	**0,001**
neve	0,078	0,111	-0,079	0,106	0,006	0,907
nome	-0,317	**0,001**	0,191	**0,001**	0,229	**0,001**
pano	-0,137	**0,005**	0,049	0,315	0,126	**0,010**
pulo	-0,173	**0,001**	0,097	**0,048**	0,139	**0,004**
rabo	-0,161	**0,001**	0,131	**0,008**	0,089	0,068
roda	-0,301	**0,001**	0,216	**0,001**	0,203	**0,001**
soma	0,123	**0,012**	0,155	**0,002**	-0,173	**0,001**
suco	-0,066	0,175	0,047	0,334	0,054	0,269
taça	-0,261	**0,001**	0,138	**0,005**	0,217	**0,001**
tipo	-0,168	**0,001**	0,106	**0,030**	0,134	**0,006**
vaso	-0,209	**0,001**	0,075	0,126	0,199	**0,001**
vinho	-0,076	0,119	0,016	0,739	0,074	0,130
zona	0,025	0,608	0,065	0,184	-0,056	0,255

Fonte: elaborado pela autora, 2024.

A partir dessa separação por grupos, as proporções de respostas corretas foram calculadas utilizando a razão entre a quantidade de respostas corretas e a quantidade total de respostas nas palavras de cada categoria (tabela 13). O método de Newcombe-Wilson foi utilizado para calcular os intervalos de confiança de 95% para a diferença entre as proporções de respostas corretas entre as categorias (Wilson, 1927; Newcombe, 1998).

Tabela 12 - Categorias das Palavras de Estímulo pela Relação com as Palavras Distratoras

Mesmo Modo de Articulação em um dos Distratores
tipo
carro
luta
fogo
copo
bife
lixo
Mesmo Modo ou Ponto de Articulação em Distratores Diferentes
jogo
doce
pano
gelo
Mesmo Ponto de Articulação em um dos Distratores
dado
café
zona
soma
taça
rabo
Ponto e Modo de Articulação Diferentes nos Distratores
nome
boca
folha
chuva
moto
roda
muro
vinho
gato
suco
vaso
neve
pulo

Fonte: elaborado pela autora, 2024

A proporção de respostas corretas para palavras em que um dos distratores a nível silábico tinha o "mesmo modo de articulação" foi significativamente maior do que as outras categorias consideradas. Palavras em que uma das alternativas distratoras tinha mesmo ponto de articulação tiveram uma proporção de respostas corretas menor do que as outras categorias consideradas. A única diferença não significativa encontrada foi entre a proporção de respostas

corretas em palavras que um dos distratores tinham ou mesmo ponto ou mesmo modo de articulação e palavras com ponto e modo de articulação diferentes (tabela 14).

Tabela 13 - Proporção de Respostas Corretas por Categoria das Palavras de Estímulo

Categoria da palavra-alvo e distratores	Proporção (em %) de respostas corretas (95% IC)
Mesmo modo de articulação	84,7% (83,1; 86,3)
Ponto e modo de articulação diferente	76,1% (74,7; 77,5)
Mesmo ponto ou modo de articulação	78,2% (75,7; 80,6)
Mesmo ponto de articulação	66,3% (64,0; 68,6)

Fonte: elaborado pela autora, 2024

Tabela 14 - Diferença das Proporções de Respostas Corretas entre Categorias das Palavras de Estímulo

Proporções comparadas		Diferença entre as proporções de resposta (IC)	p-valor
Mesmo modo	Mesmo ponto	0,184 (0,156; 0,213)	**<0,001**
Mesmo modo	Mesmo ponto ou modo	0,064 (0,035; 0,094)	**<0,001**
Mesmo modo	Ponto e modos diferentes	0,086 (0,065; 0,108)	**<0,001**
Mesmo ponto	Mesmo ponto ou modo	-0,119 (-0,154; -0,086)	**<0,001**
Mesmo ponto	Ponto e modos diferentes	-0,098 (-0,125; -0,071)	**<0,001**
Mesmo ponto ou modo	Ponto e modos diferentes	0,0216 (-0,007; 0,050)	0,14

Fonte: elaborado pela autora, 2024.

Por fim, foi realizada uma análise de *cluster*, exploratória, para identificar possíveis agrupamentos entre os fonemas na primeira sílaba, segunda sílaba e considerando a primeira e segunda sílaba. O método utilizado foi de classificação dos fonemas em grupos não estabelecidos a priori, *k-means*, na versão de Hartigan-Wong, que atribuí inicialmente os valores dos dados a centroides aleatórios. Nesse método, o agrupamento dos dados é considerado ao redor de um ponto comum de modo que a variabilidade entre os grupos seja maior do que a variabilidade dos dados dentro de cada grupo. Os centroides são pontos centrais utilizados para selecionar esses grupos, e seu valor se refere ao valor ao qual os elementos se agrupam no entorno.

Para a seleção da quantidade de centroides considerados na análise, foi realizada uma simulação considerando de 1 a 15 centroides, comparando a soma de quadrados entre grupos e foi selecionado a quantidade de centroides equivalente à quando a variação no percentual de

variabilidade explicado pelo modelo deixou de ser significativa. Esse método é informalmente conhecido como regra do cotovelo. Seguindo a regra do cotovelo, foi selecionada a quantidade de grupos quando 90% ou mais da variabilidade estava sendo explicada pelos grupos e quando a diferença entre a variabilidade explicada pela quantidade maior de centroides era menor que 4%. As análises foram realizadas considerando a proporção de respostas corretas e a identificação do fonema alvo a depender da posição dele na palavra. A proporção total de respostas corretas foi de aproximadamente 76,5%. Devido a limitação computacional de símbolos, alguns fonemas foram representados por letras, “x” para /ʃ/, “j” para /ʒ/, “lh” para /ʎ/, “nh” para /ɲ/ e “R” para /R/ (se refere ao “r forte”, que seria o “rr”, como em “rabo”).

Análise de Cluster dos Fonemas na Primeira Sílaba

Para a definição dos agrupamentos dos fonemas na primeira sílaba, a variabilidade explicada pelo modelo era de 96,6% quando foram considerados 5 grupos e 98% quando foram considerados 6 grupos. Pela diferença entre a variabilidade ser de 1,4% entre 5 e 6 grupos, foram definidos 5 grupos (centroides), apresentados na tabela 15.

Tabela 15 - Cluster e Centroides dos Fonemas Consonantais da Primeira Sílaba

Representação no Gráfico	Fonema Primeira Sílaba	Cluster	Centroide
k	/k/	5	0,9057
t	/t/		
x	/ʃ/		
p	/p/	4	0,8139
b	/b/		
d	/d/		
f	/f/		
g	/g/		
l	/l/		
v	/v/		
R	/R/		
m	/m/	3	0,7046
j	/ʒ/		
n	/n/	2	0,5472
z	/z/		
s	/s/	1	0,3488

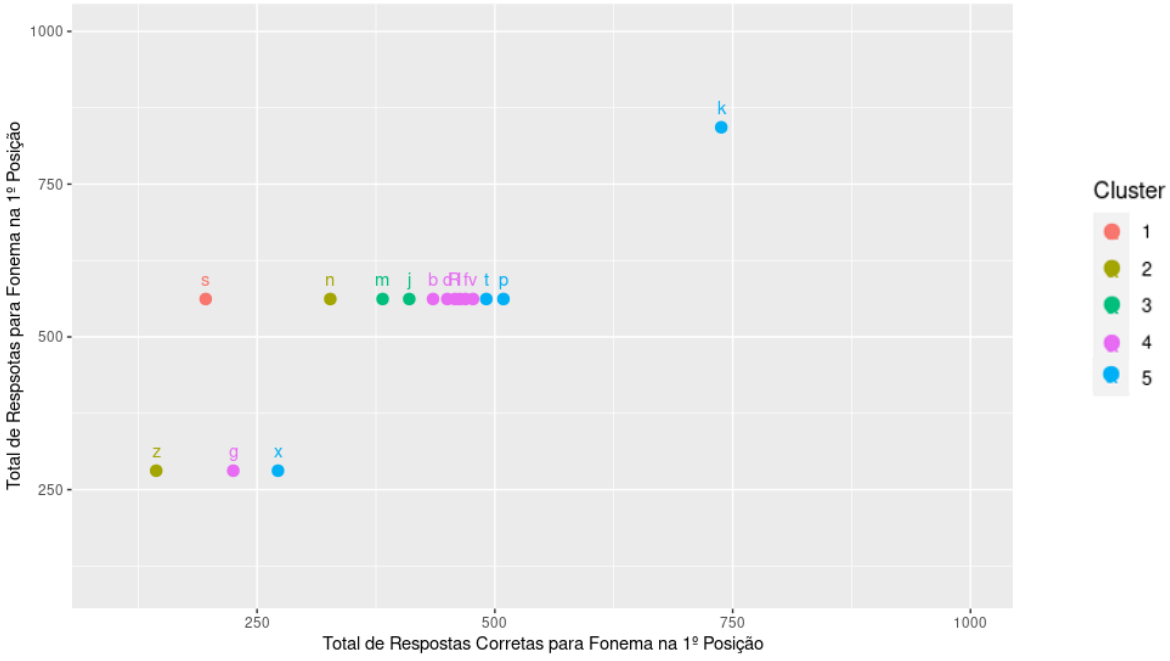
Fonte: elaborado pela autora, 2024.

O gráfico 2 apresenta os agrupamentos para os 16 fonemas consonantais que ocupavam a primeira sílaba. A quantidade de palavras em que o fonema aparecia na primeira sílaba não era a mesma para todos os fonemas. Assim, para fins de visualização, o gráfico 2 está disposto em total de respostas corretas para o fonema na primeira sílaba (no eixo X) pelo total de

respostas para o fonema (no eixo Y), por exemplo, três palavras iniciavam com o fonema /k/, enquanto apenas uma iniciava com o fonema /ʃ/. As palavras em que o fonema /s/ estava na primeira sílaba tiveram a menor proporção de respostas corretas e formaram um grupo separado. Considerando os valores dos centroides, os grupos ficaram divididos em:

- Grupo 1: composto apenas pelo fonema /s/, com a menor proporção de respostas corretas;
- Grupo 2: fonemas /z/ e /n/, com a proporção de respostas corretas de aproximadamente 51% e 58% dentro do grupo;
- Grupo 3: fonemas /m/ e /ʒ/, com a proporção de respostas corretas de aproximadamente 68% e 73% dentro do grupo;
- Grupo 4: fonemas /b/, /d/, /f/, /g/, /l/, /R/ e /v/, com a proporção de respostas corretas entre aproximadamente 77% e 84% dentro do grupo;
- Grupo 5: fonemas /k/, /t/, /p/ e /ʃ/, que apresentaram as maiores proporções de respostas corretas entre aproximadamente 87% e 96% dentro do grupo.

Gráfico 2 - Clusters dos Fonemas Consonantais na Primeira Sílaba



Fonte: elaborado pela autora, 2024.

Análise de Cluster dos Fonemas na Segunda Sílaba

Para a definição dos agrupamentos dos fonemas na segunda sílaba, a variabilidade explicada pelo modelo era de 94,8% quando foram considerados 4 grupos e 97,8% quando

foram considerados 5 grupos. Nesse caso, a diferença entre a variabilidade foi de 3 %, assim foram definidos 4 grupos, apresentados na tabela 16.

Tabela 16 - Cluster e Centroides dos Fonemas Consonantais da Segunda Sílaba

Representação no Gráfico	Fonema Segunda Sílaba	Cluster	Centroide
nh	/ɲ/	4	0,8512
p	/p/		
x	/ʃ/		
z	/z/		
l	/l/		
lh	/ʎ/		
d	/d/		
R	/R/		
s	/s/		
f	/f/		
n	/n/	3	0,7476
b	/b/		
t	/t/		
g	/g/		
k	/k/	2	0,6394
v	/v/		
r	/r/		
m	/m/	1	0,4609

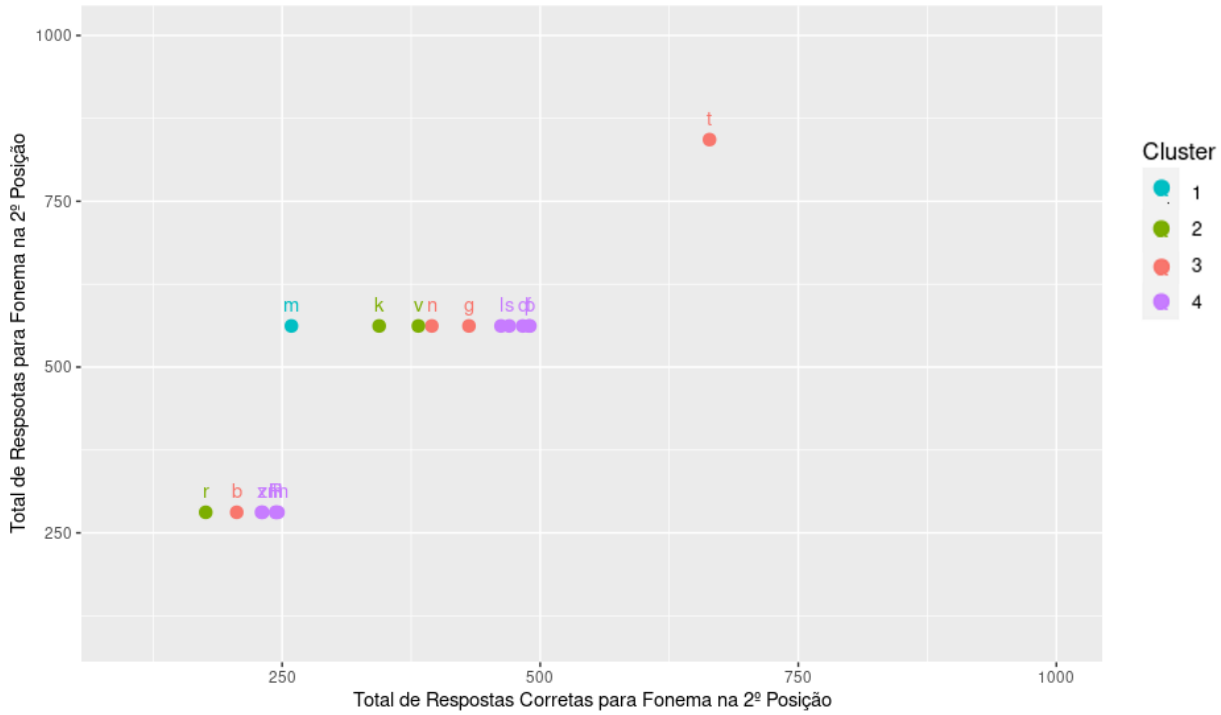
Fonte: elaborado pela autora, 2024.

O gráfico 3 apresenta os agrupamentos para os 18 fonemas consonantais que ocupavam a segunda sílaba. Também, a quantidade de palavras em que o fonema aparecia na segunda sílaba não era a mesma para todos os fonemas. Por isso, novamente para fins de visualização, o gráfico 3 está disposto em total de respostas corretas para o fonema na segunda sílaba (no eixo X) pelo total de respostas por fonema (no eixo Y), por exemplo, três palavras tinham o fonema /t/ no meio da palavra, enquanto apenas uma tinha o fonema /b/. Considerando os valores dos centroides, os grupos ficaram divididos em:

- Grupo 1: composto apenas pelo fonema /m/, com a menor proporção de respostas corretas;
- Grupo 2: fonemas /k/, /r/ e /v/, com proporções de respostas corretas entre aproximadamente 61% e 68% dentro do grupo;
- Grupo 3: fonemas /b/, /g/, /n/ e /t/, com proporções de respostas corretas entre aproximadamente 70% e 78% dentro do grupo;

- Grupo 4: fonemas /d/, /f/, /l/, /ʎ/, /ɲ/, /p/, /R/, /s/, /ʃ/ e /z/, que apresentaram as maiores proporções de respostas corretas entre aproximadamente 82% e 88% dentro do grupo.

Gráfico 3 - Clusters dos Fonemas Consonantais na Segunda Sílaba



Fonte: elaborado pela autora, 2024.

Análise de Cluster dos Fonemas na Primeira e Segunda Sílaba

Para a definição dos agrupamentos dos fonemas independentemente da posição na palavra, a variabilidade explicada pelo modelo era de 88,9% quando foram considerados 4 grupos e 93,3% quando foram considerados 5 grupos. Pela diferença entre a variabilidade ser de 4,4% e seguindo os critérios estabelecidos previamente, foram definidos 5 grupos, apresentados na tabela 17.

O gráfico 4 apresenta a proporção de respostas corretas para quando o fonema estava na primeira sílaba (eixo x) e quando o mesmo fonema aparecia na segunda sílaba (eixo y). Os fonemas /ʎ/, /ɲ/ e /r/, não ocupam a posição inicial da palavra e o fonema /ʒ/ não apareceu na segunda sílaba, por isso esses fonemas estão nas margens do gráfico, visto que para um dos eixos, seu valor é zero. Considerando os valores dos centroides, os grupos ficaram divididos em:

- Grupo 1: fonema /ʒ/, que não foi apresentado na segunda sílaba, com a média da proporção de respostas corretas de aproximadamente 73% quando apareceu na primeira sílaba;

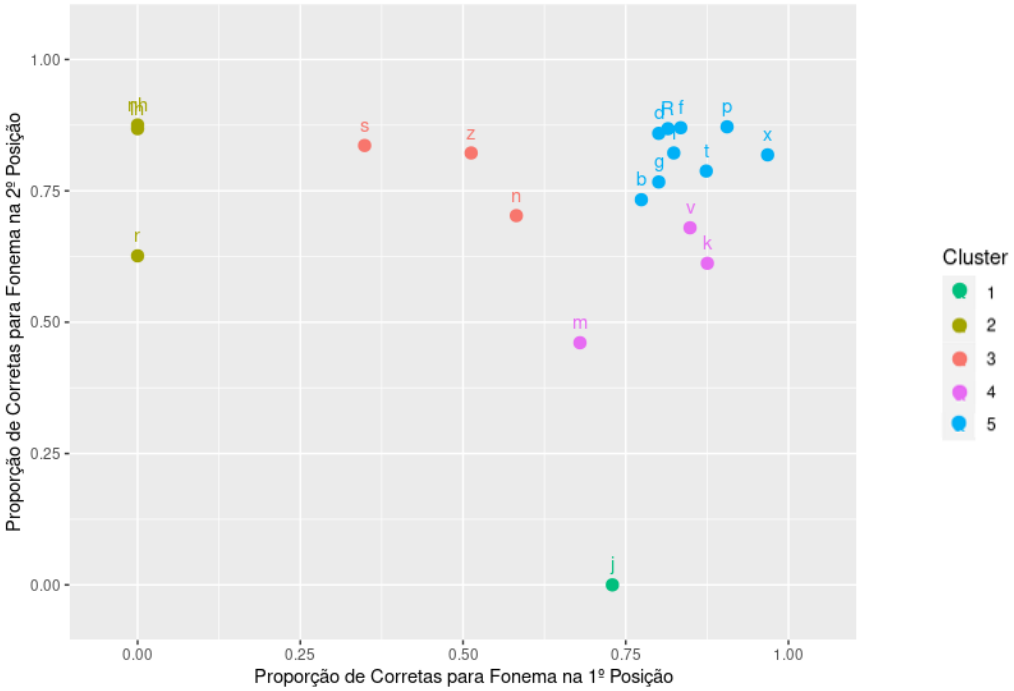
- Grupo 2: fonemas /ʎ/, /ɲ/ e /r/, que não foram apresentados na primeira sílaba, com a média da proporção de respostas corretas de aproximadamente 79% quando apareceram na segunda sílaba;
- Grupo 3: fonemas /n/, /s/ e /z/, com a média da proporção de respostas corretas de aproximadamente 48% quando apareceram na primeira sílaba e de aproximadamente 79%, quando apareceram na segunda sílaba;
- Grupo 4: fonemas /k/, /m/ e /v/, com a média da proporção de respostas corretas de aproximadamente 79% quando apareceram na primeira sílaba e de aproximadamente 58% quando apareceram na segunda sílaba;
- Grupo 5: fonemas /b/, /d/, /f/, /g/, /l/, /p/, /R/, /t/ e /ʃ/, que apresentaram as maiores proporções de respostas corretas independentemente da posição na palavra, com a média da proporção de respostas corretas de aproximadamente 84% quando apareceram na primeira sílaba e de aproximadamente 82% quando apareceram na segunda sílaba.

Tabela 17 - Cluster e Centroides dos Fonemas Consonantais da Primeira e Segunda Sílaba

Representação no Gráfico	Fonema	Cluster	Centroide
l	/l/		
b	/b/		
d	/d/		
f	/f/		
g	/g/	5	0,8311
p	/p/		
t	/t/		
x	/ʃ/		
R	/R/		
k	/k/		
m	/m/	4	0,6925
v	/v/		
n	/n/		
s	/s/	3	0,6340
z	/z/		
lh	/ʎ/		
nh	/ɲ/	2	0,7953
r	/r/		
j	/ʒ/	1	0,7295

Fonte: elaborado pela autora, 2024.

Gráfico 4 - Clusters dos Fonemas Consonantais na Primeira e Segunda Sílaba



Fonte: elaborado pela autora, 2024.

4. DISCUSSÃO

O presente estudo foi conduzido para desenvolver uma base de arquivos de vídeo que possa ser utilizado para treinamento em leitura labial do Português Brasileiro, em virtude da escassez de materiais e trabalhos relacionados a leitura labial nesse idioma. E, devido a esta escassez, foi identificada uma oportunidade, o que fomentou o caráter exploratório dessa pesquisa. Esse trabalho foi pioneiro na proposição do primeiro conjunto de materiais adequados para treinamento em leitura labial do Português Brasileiro.

Assim, considerando o contexto interdisciplinar e as diferentes etapas realizadas na pesquisa, a discussão também está dividida em duas partes: a primeira relacionada à seleção das palavras de estímulos e ao material produzido; e a segunda ao estabelecimento do protocolo e resultados da sessão experimental.

4. 1. Primeira Parte: Palavras Selecionadas e Vídeos Produzidos

As inferências em leitura labial podem ser extraídas de estudos publicados e relatos que surgiram em contextos diferentes (como por exemplo, em outros idiomas), por isso é necessário cautela em abordagens empíricas ao considerar como utilizar a leitura labial para fins de coleta de informações, conforme explicam Campbell e Mohammed (2010). Não há consenso metodológico relacionado a investigação da percepção visual da fala por meio da leitura labial, especialmente por ser uma área interdisciplinar relacionada a linguagem e pouco explorada em diferentes idiomas. As investigações nessa temática têm sido isoladas ou com pouco aprofundamento interdisciplinar no desenvolvimento e estabelecimento de critérios metodológicos. A escolha das palavras de estímulo e os vídeos produzidos nesse trabalho tiveram embasamento teórico e metodológico provenientes de pesquisas do Inglês, com adaptações relacionadas ao idioma, que ocorreram em interface com a Linguística e a Fonoaudiologia. Essa interface entre as áreas, configura também discussões metodológicas em outros aspectos. Alguns pontos passíveis dessa adaptação são pertinentes para a discussão que se segue.

Na Fonoaudiologia, as diferentes abordagens teóricas que fornecem ou não interface com a Linguística foram amplamente discutidas por Cristófaro-Silva (2015), ao abordar o paralelo entre a fonética e a fonologia. Segundo a autora, em uma dessas abordagens, a Fonologia de Uso, a fonética e a fonologia não são dissociadas, ou seja, são analisadas em conjunto e as representações fonológicas dizem respeito a generalizações baseadas na experiência. A autora explica que “o nível fonológico trata das generalizações observadas na estrutura sonora e expressa formalmente o conhecimento abstrato dos falantes (representação fonêmica, forma subjacente, representação lexical). O nível fonético é visto como a saída do

componente gramatical onde o detalhe fonético é observado (representação fonética)" (Cristófaro-Silva, 2015, p. 224).

Cristófaro-Silva e De Bona (2017) ressalta que em teorias "formais", as representações são simples, ou seja, excluem o detalhe fonético e parâmetros extralinguísticos, já em representações complexas, as teorias compreendem o detalhe fonético e os parâmetros extralinguísticos e não dissociam a fonética e a fonologia. A Fonologia de Uso se situa em uma representação complexa da fala. A autora alerta que a modelagem multimodal da fala gera grandes desafios metodológicos, mas que é também o que propiciará explicações consistentes sobre a natureza da linguagem. Cristófaro-Silva (2015) apresenta os pressupostos teóricos da Fonologia de Uso, com base em Bybee (2001) como:

> "Experiência afeta representações; representações mentais de objetos linguísticos têm as mesmas propriedades de representações mentais de outros objetos; categorização é baseada em identidade e em similaridade; generalizações em relação a formas não são separadas de representações, e sim emergem a partir das formas; a organização lexical oferece generalizações e segmentações em vários níveis de abstração e generalização; o conhecimento gramatical tem caráter de procedimento" (Cristófaro-Silva, 2015, p. 225).

O estudo em percepção visual da fala tem se baseado em paradigmas de investigação da fala relacionados a modalidade auditiva, por isso atualmente trabalha com pressupostos de frequência de uso e na tentativa de traçar competidores adequados para compreender a modalidade visual. Trabalhar com esses dois pressupostos exigem desafios metodológicos que geram alguns entraves e limitações da investigação em leitura labial. São esses desafios que são importantes de serem ressaltados e discutidos brevemente a seguir.

Existem implicações relacionadas à utilização de um *corpus* de fala ao invés de um *corpus* de escrita para a seleção dos estímulos. Cresti (2005) explica que a organização no nível lexical da língua falada é diferente da organização da língua escrita. Dessa forma, as métricas de fala também divergem da escrita. Na organização do C-Oral Brasil, Mello (2012) discute a respeito das interfaces entre um *corpus* de fala e um de escrita e exemplifica que na fala há uma maior proporção de uso de verbos do que substantivos. Isso implica, por exemplo, em métricas diferentes de frequência de uso das palavras. Além disso, o acesso a essas métricas ainda é muito mais robusto em relação a escrita do que a fala, pois as bases de dados de textos escritos disponíveis para extrair métricas da língua escrita é maior e mais acessível do que bases de dados para extrair métricas da língua falada. O mesmo é válido para densidade de vizinhança das palavras.

A utilização de um *corpus* de fala para seleção dos estímulos de leitura labial pode permitir a produção de um material mais completo e adequado, por ser baseado na realidade da fala. Em um contexto real, o observador tem que extrair informações a partir de interações diárias, por isso é importante trabalhar com informações linguísticas que estejam mais próximas dessas interações. Assim, usar um *corpus* de fala com métricas adequadas sobre as palavras no idioma pode auxiliar no delineamento de tarefas experimentais em blocos mais completos de treinamento de leitura labial, bem como propiciar o estabelecimento de contexto baseados na realidade. No caso do *corpora* C-Oral Brasil, além de ter informações extraídas de interações informais em contextos reais, o material conta também com outra parte proveniente de interações formais, ou seja, vinculadas a ambientes específicos de interação, como salas de aula, audiências etc. Essa separação de acordo com o ambiente pode contribuir para mapear e produzir estímulos específicos para tarefas experimentais e possibilitar novas investigações sobre como o contexto influencia, por exemplo, na competição lexical de palavras isoladas. Apesar de ser um *corpus* baseado na diatopia mineira, de variação pela região geográfica, a percepção da fala se difere da produção, ou seja, mesmo que uma palavra tenha uma pronúncia diferente a depender do sotaque, isso não impede o reconhecimento da palavras, tanto na modalidade auditiva como visual. No entanto, novas investigações a respeito dos diferentes sotaques do Brasil pode ser interessante para estudos futuros.

A frequência de uso era uma variável considerada a princípio nessa pesquisa, no entanto, devido as implicações vinculadas ao uso dessa variável e da densidade de vizinhança, apresentadas anteriormente, se faz necessário um estudo interdisciplinar mais aprofundado a respeito dessas métricas relacionadas as palavras selecionadas.

Nesse trabalho, substantivos foram selecionados a partir do C-Oral Brasil, a fim de padronizar os estímulos produzidos. As palavras selecionadas foram balanceadas em relação aos fonemas consonantais do Português Brasileiro em duas posições diferentes em que esses fonemas podem aparecer na palavra (onset (começo da palavra) e intervocálica (entre vogais)). Esse balanceamento consistiu no aparecimento dos fonemas nessas duas posições nas palavras selecionadas, por exemplo, palavras que o fonema /g/ estivesse em posição de onset, como gota, e em posição intervocálica, como fogo. Nas palavras de estímulo dessa pesquisa, 16 fonemas consonantais aparecem em posição inicial (onset) e 19 em posição intervocálica.

É interessante trabalhar com a identificação de fonemas dentro de palavras, pois a competição lexical pode impactar na percepção e discriminação do estímulo físico (fonema-alvo). Assim, mesmo que uma tarefa experimental considere sílabas sem sentido para investigar a discriminação dos fonemas de um idioma, não trabalhará com a realidade, uma vez que as

palavras que competem no léxico divergem conforme a estrutura e existência dessas palavras no idioma. Por exemplo, como já foi explicado anteriormente, a discriminação de um estímulo físico não é única e exclusivamente influenciada por outro estímulo físico parecido, como nas palavras “fato” e “vato”, que apesar de terem o mesmo ponto de articulação, não competem no léxico, pois “vato” não é uma palavra com significado no idioma.

Cabe ressaltar também que o material produzido nessa pesquisa tem uma proposta diferente das ferramentas relacionadas a leitura labial já existentes em Português Brasileiro, uma vez que visava ao treinamento e a investigação dessa habilidade. Por isso, é necessário diferenciar teste e treinamento dentro dessa temática. O teste de leitura labial elaborado em Português Brasileiro por Tedesco, Chiari e Vieira (1995), contempla palavras e sentenças do idioma e visa a avaliar o desempenho em leitura labial, no entanto não está disponível, mas foi descrito e utilizado por Oliveira, Soares e Chiari (2014) para avaliar a habilidade da leitura labial de 61 participantes com e sem deficiência auditiva, com idades entre 12 e 70 anos. As autoras produziram os estímulos do teste em vídeos, gravados com uma locutora feminina, para utilizarem em uma coleta presencial. Os vídeos produzidos não foram disponibilizados e não têm qualidade de imagem/características de produção indicadas. Esse teste não é uma tarefa relacionada ao treinamento e investigação da leitura labial, não é computadorizado e não possui estímulos padronizados, o que impacta em alguns pontos de investigação, como por exemplo, o locutor é o próprio aplicador/examinador em coletas presenciais.

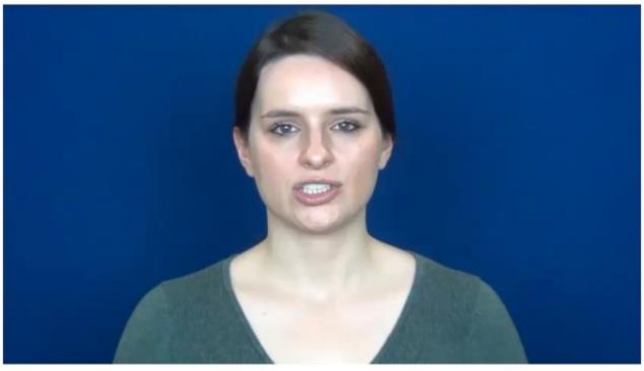
O material desenvolvido para tarefas experimentais de treinamento nessa pesquisa é passível de se tornar um teste computadorizado robusto e completo que tenha o intuito de avaliar o desempenho na habilidade de leitura labial do Português Brasileiro, a partir de validação psicométrica. No entanto, para isso ainda é necessário ampliar a base de arquivos com a produção de mais vídeos de estímulo. Para que essa ampliação ocorra é interessante contemplar palavras que tenham diferentes frequência de uso e densidades de vizinhança, encontros consonantais, contextos vocálicos e de coarticulação diversos, além de usar somente palavras e não pseudopalavras, ter mais de um tipo locutor e inserir as palavras em sentenças. Essas colocações são sugeridas com base nos estudos apresentados anteriormente, do que já se conhece sobre a investigação em leitura labial em diferentes idiomas, mas principalmente provenientes do Inglês.

Já em relação ao treinamento, diferentes tarefas experimentais podem ser elaboradas. Tarefas computadorizadas têm sido cada vez mais comuns devido a manipulação e padronização dos estímulos, o que possibilita maior controle metodológico e melhor adesão à participação. Outro trabalho que também selecionou, produziu estímulos e inseriu em uma

tarefa experimental foi Pimperton, Ralph-Lewis e MacSweeney (2017). As autoras selecionaram palavras do Inglês Britânico, produziram e testaram vídeos de estímulo em uma sessão experimental para verificar a compensação perceptual de leitura labial em adultos sem e com deficiência auditiva (que faziam uso de Implante Coclear (IC)). Foram selecionadas 123 palavras do Inglês que eram gramaticalmente classificadas como substantivos ou correspondentes a cores. O principal critério para a seleção das palavras foram as métricas de competição lexical na modalidade visual. Segundo as autoras, as métricas utilizadas refletiram "a competição geral no léxico de referência para a palavra-alvo com base na semelhança das distribuições de resposta de seus fonemas constituintes (de uma tarefa de identificação visual apenas de fonemas de escolha forçada) com as dos fonemas em todas as outras palavras do mesmo tipo de padrão no léxico" (Pimperton, Ralph-Lewis e MacSweeney, 2017, p. 5, tradução própria). Essas métricas foram extraídas do Phi-Lex Database (https://osf.io/ynqxr/), proposto por Strand (2014).

A partir das palavras selecionadas, as autoras produziram vídeos com uma locutora feminina pronunciando todas as palavras. No entanto, as autoras não forneceram informações a respeito das especificações técnicas dos vídeos produzidos. A imagem 1 apresenta uma captura de um dos vídeos produzidos por Pimperton, Ralph-Lewis e MacSweeney (2017) com o fonema alvo dentro da palavra do Inglês Britânico "*shoe*", visando o fonema /ʃ/. Em paralelo, a imagem 2 mostra a captura de um dos vídeos produzidos nesse trabalho, da palavra do Português Brasileiro "chuva", visando demonstrar o mesmo fonema /ʃ/.

Imagem 1 - Captura de Tela do Vídeo da Locutora Pronunciando a Palavra "shoe" no Inglês Britânico



Fonte: Pimperton, Ralph-Lewis e MacSweeney (2017), adaptado pela autora, 2024.

Imagem 2 - Captura de Tela do Vídeo da Locutora Pronunciando a Palavra “chuva” no Português Brasileiro

Fonte: elaborado pela autora, 2024.

Os vídeos produzidos por Pimperton, Ralph-Lewis e MacSweeney (2017) tem semelhanças com o material produzido nesse trabalho. As autoras descreveram que os vídeos foram produzidos com a locutora pronunciando palavras em um tom de voz de conversa normal e foram editados posteriormente para remoção do som, com o intuito de que os participantes vissem a palavra a partir de uma pronúncia natural, porém sem som. Os vídeos de Pimperton, Ralph-Lewis e MacSweeney (2017) foram utilizados em outros trabalhos do Inglês (ver Pimperton et al., 2019). No entanto, o protocolo experimental estabelecido foi diferente dessa pesquisa, que será discutido posteriormente, em paralelo ao protocolo realizado aqui.

Costa (2009) também produziu vídeos de uma locutora feminina pronunciando sílabas sem sentido (pseudopalavras) para testar em paralelo a uma face artificial em um Teste de Inteligibilidade da Fala, em que ambos os estímulos foram apresentados com diferentes níveis de ruído. Segundo a autora, as gravações foram realizadas em um laboratório de multimídia, com uma câmera SONY DSR-PD170, posicionada a aproximadamente 1 metro a frente da locutora. A taxa de quadros por segundo utilizada foi de 29,97 FPS e a resolução foi de 720 x 486 pixels. O áudio também foi capturado. A locutora estava posicionada na frente de um fundo de cor azul com iluminação direcionada para a face. As imagens 3 e 4 fazem um paralelo com um exemplo dos vídeos produzidos por Costa (2009) (imagem 3) e nesse trabalho (imagem 4). Os vídeos produzidos por Costa (2009) não foram disponibilizados após a pesquisa, pois foram produzidos para comparar os resultados do teste com os vídeos utilizando a face artificial que a autora produziu. Todos os estímulos produzidos pela autora se referiam a sílabas sem sentido.

Imagem 3 - Imagem Capturada no Processo de Gravação dos Estímulos de Costa (2009)

Fonte: Costa, 2009, p. 29

Imagem 4 - Imagem Capturada no Processo de Gravação dos Estímulos dessa Pesquisa

Fonte: elaborado pela autora, 2024

As palavras selecionadas e os vídeos produzidos preenchem uma parte da lacuna da necessidade de material adequado para investigação e treinamento da leitura labial do Português Brasileiro. A partir de métricas mais robustas a respeito das palavras de estímulo e da ampliação do material produzido, novas pesquisas poderão se desenvolver a partir dessa, com diferentes protocolos experimentais. Alguns exemplos de novas pesquisas que podem ser desenvolvidas com outros protocolos experimentais são:

- Observar qual o efeito da frequência de uso e densidade de vizinhança no reconhecimento visual das palavras de estímulo selecionadas nesse trabalho, partindo da hipótese de que as palavras com maior frequência de uso e vizinhança esparsa serão mais bem reconhecidas do que as palavras com menor frequência de uso e vizinhança densa;
- Adaptar a tarefa experimental dessa pesquisa para pessoas com deficiência auditiva e comparar o desempenho na identificação do fonema em palavras do Português Brasileiro

por pessoas sem e com deficiência auditiva (oralizadas ou não oralizadas, que fazem uso ou não de Implante Coclear);

- Observar o desempenho na identificação das palavras a depender da idade, sexo e escolaridade;
- Observar a diferença de desempenho a depender do tipo de locutor (masculino ou feminina);
- Utilizar figuras nas alternativas de resposta, para observar as diferenças na identificação de fonemas por crianças ouvintes e não ouvintes;
- Investigar a respeito da relação entre dislexia, consciência fonológica, desempenho em leitura e leitura labial;
- Investigar a respeito das bases neurobiológicas da leitura labial (plasticidade em relação a compensação no processamento auditivo para pessoas com deficiência auditiva).

Além da base de arquivos de vídeo permanecer disponível para ser usada em novas pesquisas, os dados gerados pela coleta de dados proveniente da sessão experimental desse estudo também podem ser utilizados para novas análises em algumas dessas temáticas sugeridas, por exemplo, em relação a frequência de uso e densidade de vizinhança e observação do desempenho pelo sexo e idade. No Inglês, algumas linhas de investigação tem sido traçadas relacionadas a um melhor desempenho de mulheres em leitura labial e pesquisas associadas a outras habilidades cognitivas em paralelo com a leitura labial (como a memória de trabalho, dislexia e consciência fonológica). É importante padronizar estímulos, para investigações computadorizadas e tarefas de treinamento apropriadas, para que as investigações do Inglês sejam acompanhadas pelo Português Brasileiro, desde que o estudo dessa temática seja fomentado nesse idioma.

4. 2. Segunda Parte: Sessão Experimental

O caráter inicial dessa investigação no Português Brasileiro permite que o protocolo experimental de treinamento apresentado possa ser mais bem explorado em estudos posteriores, por exemplo com uma tarefa de resposta aberta. O tipo de resposta fornecida pelo participante pode ser aberta, fechada (de escolha forçada), indicadas por meio de imagens ou por escrito, ou seja, há elementos que podem ampliar a abrangência do caráter exploratório. Nesse trabalho, a tarefa de treinamento foi com alternativas de resposta fechada para cada estímulo, que apareciam por escrito para o participante. Pimperton, Ralph-Lewis e MacSweeney (2017) realizaram uma tarefa de resposta aberta para treinamento em leitura labial do Inglês Britânico. A tarefa estabelecida pelas autoras tem similaridades com a tarefa dessa pesquisa. Para delinear a sessão experimental, as autoras dividiram os 123 vídeos produzidos em quatro grupos

aleatórios. Participantes ouvintes (61) e não ouvintes (15) foram designados, também de maneira aleatória, para completar uma das quatro tarefas (baseadas na divisão dos grupos de vídeos). A sessão experimental foi hospedada na plataforma *online* Opinio, semelhante a Pavlovia, usada nessa pesquisa.

Na tarefa de Pimperton, Ralph-Lewis e MacSweeney (2017), assim como nesse trabalho, o participante acessava e realizava a sessão experimental a partir de seu próprio dispositivo e, antes de iniciar, recebia as instruções. O participante precisava clicar no vídeo de estímulo para iniciar a reprodução. Alguns pontos divergiram no protocolo experimental, por exemplo, os vídeos foram mostrados somente uma vez, enquanto nesse trabalho foram mostrados duas vezes. A resposta era indicada por escrito, por meio de uma caixa de resposta em que o participante deveria escrever o que julgava ter visto. As respostas eram pontuadas quando houvesse acerto ou indicação de palavras homófonas (palavra com escrita e significado diferente, mas pronúncia igual, como as palavras do Inglês "*I*" e "*eye*").

Costa (2009) também produziu estímulos e testou em uma tarefa de discriminação de escolha forçada, que utilizava sílabas sem sentido como estímulo e alternativas de resposta, apresentado na imagem 5. O trabalho da autora complementou a pesquisa de De Martino (2005), ambos estudos pioneiros no Brasil, que mapearam os visemas do Português Brasileiro a depender de contexto fonético em pseudopalavras. Apesar de ser voltado para a produção da fala de uma animação/face artificial, Costa (2009) realizou uma tarefa de percepção dos estímulos em situações de ruído para comparar o reconhecimento das pseudopalavras através uma face natural e uma artificial. A face natural teve melhor reconhecimento do que a face artificial.

Apesar de haver pontos de convergência com essa pesquisa na investigação de leitura labial do Português Brasileiro, tanto Costa (2009) quanto De Martino (2005) trabalharam com o mapeamento de visemas em segmentos consonantais e vocálicos para explorar a similaridade articulatória na produção desses segmentos. As pseudopalavras de estímulo, que eram dissílabas com a mesma estrutura Consoante Vogal (CV), propiciavam dois contextos de coarticulação dos fonemas (CVC e VCV), por exemplo no segmento "papa" (CVCV), "pap" (CVC) e "apa" (VCV), para configurar a articulação de uma *talking head* (cabeça falante), pois juntando diferentes contextos, a face artificial consegue pronunciar diversas palavras. Esse mapeamento se distingue do estabelecimento de estímulos baseados em contextos reais (dentro de palavras) para investigar a percepção da leitura labial e configurar uma tarefa de treinamento baseada no reconhecimento de palavras, que é o enfoque dessa pesquisa.

Imagem 5 - Recorte da Tarefa Experimental com Pseudopalavras de Costa (2009)

Fonte: Costa, 2009, p. 78.

As limitações do presente trabalho estão mais relacionadas à tarefa experimental do que ao desenvolvimento dos estímulos. A seleção dos estímulos e delineamento da tarefa experimental precisou ser padronizada e restrita devido a quantidade de variáveis envolvidas. Ainda, cabe ressaltar que, as investigações sobre leitura labial podem esbarrar em limitações em que os achados não são passíveis de generalização, independentemente do tamanho da amostra, justamente pelo caráter interdisciplinar e das barreiras de idioma envolvidos nas pesquisas dessa temática.

A amostra desse estudo foi somente de pessoas ouvintes pelo caráter pioneiro, exploratório e interdisciplinar, além da realização da tarefa experimental de maneira computadorizada e remota e pelo momento de execução da pesquisa, que foi durante a pandemia da COVID-19. Estudos de treinamento em leitura labial do Inglês muitas vezes realizam comparações entre o desempenho na tarefa por pessoas com e sem deficiência auditiva, como pode ser observado no trabalho de Pimperton, Ralph-Lewis e MacSweeney (2017) e Buchanan-Worster, Hulme, Dennan e MacSweeney (2021). A comparação de grupos propicia uma maior robustez e consistência na investigação da leitura labial, especialmente quando o grupo com maior potencial de ser beneficiado nessa investigação é de pessoas com deficiência auditiva. Bernstein (2012) ressalta que a compreensão do processamento visual da fala por meio da leitura labial pode ser acelerada se for investigada com pessoas que tiveram pouco acesso a informações auditivas, ou seja, pessoas com deficiência auditiva congênita.

Segundo Bernstein, Jordan, Auer e Eberhardt (2022), na leitura labial, os observadores

podem perceber a sonoridade do fonema, o modo de articulação e informações prosódicas (como entonação e ênfase). A ideia de que as informações visuais se referem somente a leitura labial do local de articulação se mostrou incorreta. Fonemas com o mesmo ponto de articulação são mais passíveis de serem confundidos, mas isso não impede o reconhecimento a depender do contexto, por exemplo, dentro da palavra, como discutido anteriormente.

Na análise exploratória dessa pesquisa, a menor proporção de respostas corretas foi de 14,9% em "soma" e a maior de 96,8% em "chuva". A palavra "soma" foi confundida com o distrator "sopa" (que teve 83,3% de proporção de resposta). A alta confusabilidade entre soma e sopa pode ser explicada por /m/ e /p/ pertencerem ao mesmo visema, ou seja, possuem o mesmo ponto de articulação. No entanto, outras palavras que também tinham um dos distratores com mesmo ponto de articulação tiveram uma alta proporção de respostas corretas, como a palavra "dado" que teve 82,2% em relação aos distratores "lado" (13,2%) e "dano" (4,6%). Os fonemas /d/ e /n/ possuem o mesmo ponto de articulação, já os fonemas /d/ e /l/ possuem pontos de articulação próximos, podendo ser classificados, a depender do inventário fonético, também com o mesmo ponto de articulação (linguodentais ou alveolares). No entanto, não foram tão confundidos como "soma" e "sopa". Isso pode ser explicado, como já foi ressaltado anteriormente e que é atualmente a principal linha de investigação em outros idiomas, pela frequência de uso e densidade de vizinhança das palavras. Para retomar, a densidade de vizinhança se refere a quantos vizinhos lexicais uma palavra pode ter, esses vizinhos se diferenciam a partir da modalidade de estímulo (auditivo ou visual, por exemplo). A palavras "sopa" e "soma" podem ter frequências de uso e densidade de vizinhança similares, enquanto "dado", "lado" e "dano" podem ter frequências distintas e densidade de vizinhança esparsa.

Como estudos futuros, parece interessante investigar a hipótese de que também na modalidade visual, palavras mais frequentes e com vizinhança esparsa sejam mais facilmente reconhecidas do que palavras com frequência baixa e vizinhança densa. Para essa investigação sugere-se padronizar os distratores quanto aos pontos e modos de articulação e observar a frequência de uso e densidade de vizinhança deles. Essa mesma hipótese pode explicar a alta proporção de respostas corretas em "chuva", visto que as variáveis de frequência de uso e densidade de vizinhança combinadas, devem estabelecer uma frequência alta da palavra e vizinhança mais esparsa do que as demais.

Certamente um desafio nas tarefas de escolha forçada é estabelecer os distratores lexicais com a mesma estrutura de palavra e com diferentes variações de fonemas. As palavras podem ter diferentes significados e diversas classificações gramaticais, o que pode impactar na competição da palavra-alvo com os possíveis distratores. Ainda é necessário, no Português

Brasileiro, estabelecer de maneira mais assertiva possíveis competidores visuais das palavras de estímulo, baseados na modalidade visual.

Ainda, a investigação entre a relação da leitura labial com habilidades cognitivas tem sugerido que as diferenças individuais pertinentes ao desempenho em leitura labial podem estar vinculadas à memória de trabalho e à velocidade de processamento, como bem apresentado por Feld e Sommers (2009). Os autores sugerem que o declínio das habilidades cognitivas com a idade pode explicar também o declínio do desempenho em leitura labial, uma vez que verificaram que adultos jovens apresentaram melhor desempenho em leitura labial e em habilidades perceptivas do que idosos, bem como maior tempo de memória de trabalho e velocidade de processamento mais rápida.

Feld e Sommers (2009) citam em seu trabalho que, na modalidade auditiva, Pichora-Fuller, Schneider e Daneman (1996) mostraram que quanto maior a dificuldade do processamento perceptivo, maiores são as demandas de habilidades cognitivas e sugerem que o mesmo ocorre na modalidade visual. Isso poderia explicar a correlação significativa entre o tempo de resposta e o tipo de resposta encontrada nessa pesquisa. No entanto, quando houve correlação significativa (que não foi observada em todas as palavras), essa correlação foi baixa, por isso, é necessário um estudo mais aprofundado a respeito da relação entre o tempo de resposta (velocidade de processamento), habilidades cognitivas e a leitura labial. Cabe lembrar, também, que por ser uma tarefa experimental *online*, a plataforma utilizada apresenta *delay* em relação ao tempo de resposta coletado, ou seja, o tempo de resposta é maior do que o tempo de resposta coletado manualmente em tarefas experimentais presenciais, porém é menos passível de erros relacionados ao aplicador da tarefa. Sugere-se, para investigações relacionadas a essa natureza, a aplicação da tarefa em coleta computadorizada, porém presencial, para maior controle metodológico do tempo de resposta.

Feld e Sommers (2009) apresentam ainda a hipótese de que se a leitura labial está diretamente relacionada a habilidades cognitivas, isso poderia explicar a baixa melhoria no desempenho e pouca efetividade do treinamento em leitura labial. No entanto, os autores sugerem que a associação entre o treino de habilidades cognitivas ao de leitura labial possa produzir benefícios mais significativos, por exemplo, em contexto clínico para pessoas com deficiência auditiva.

Em relação a classificação dos fonemas consonantais em palavras, Bernstein e Liebenthal (2014) explicam que “a visibilidade das características da fala ou fonemas não pode ser inferida com precisão a partir de um simples mapeamento um-para-um entre a visibilidade da anatomia da produção da fala (por exemplo, lábios, boca, língua, glote) e as características

da fala (por exemplo, vozeamento, lugar, maneira, nasalidade)" (Bernstein e Liebenthal, 2014, p. 4, tradução própria). Porém, as autoras esclarecem que o conceito de visema foi concebido para descrever padrões de confusões de fonemas na leitura labial e que grande parte das pesquisas envolve visemas e contabilização de erros nas tarefas de driscriminação. Assim, informações provenientes de visemas podem ser significativas, informativas e funcionar como recurso. Isso é elucidado no trabalho de Bernstein, Iverson e Auer Jr (1997) em que palavras impressas isoladas foram mostradas a adultos jovens com e sem deficiência auditiva, que deveriam indicar qual das duas palavras faladas, que tinham o mesmo visema, correspondiam a palavra de estímulo. Segundo os autores, mesmo nas palavras com visemas, as pontuações ficaram na faixa de 65% a 80%. Assim, Bernstein, Iverson e Auer Jr (1997) concluíram que para reconhecer as palavras, os participantes perceberam detalhes fonéticos visuais para além do visema.

Nesse trabalho, a média das proporções de respostas corretas em que uma das alternativas distratoras tinha o mesmo ponto de articulação, que configura um visema, foi de 66,3%, considerando que uma dessas era a palavra "soma", com uma proporção de acerto de 14,9%, e as demais palavras nesse grupo tiveram proporções de respostas corretas de 89,3% (taça), 86,8% (café), 82,2% (dado), 73,3% (rabo) e 51,2% (zona).

No entanto, nas palavras que um dos distratores tinha o mesmo modo de articulação, a média das proporções de respostas corretas foi de 84,7%, sendo que todas as proporções de respostas corretas das palavras nesse grupo foram acima de 80%: 80,1% (fogo), 81,9% (lixo), 82,9% (luta), 85,4% (tipo), 86,8% (carro), 87,2% (bife) e 89,0% (copo). Assim, o modo de articulação pareceu desempenhar um papel relevante na identificação das palavras isoladas, uma vez que a média das proporções de repostas corretas para palavras em que um dos distratores a nível silábico tinha o "mesmo modo de articulação" foi significativamente maior do que as outras categorias consideradas. Ainda, esta colocação corrobora com a observação do visema associado ao ponto de articulação, mesmo assim, isso não impediu a identificação dos fonemas com mesmo ponto de articulação em palavras, ainda que essa identificação tenha sido significativamente menor do que em relação a outras classificações dos fonemas. Sugere-se para estudos futuros que ainda seja observado se o modo e ponto de articulação dos fonemas consonantais também tem papel decisivo na identificação, mesmo considerando outras variáveis.

Como os fonemas consonantais foram apresentados em dois contextos na estrutura da palavra (CV e VCV), foram realizadas três análises de agrupamento a depender da posição do fonema consonantal: em onset (CV), intervocálica (VCV) e considerando a proporção de

respostas corretas para o fonema independentemente da posição em que foi apresentado na palavra. Todas as análises foram feitas considerando a proporção de respostas corretas, ou seja, quando o participante indicou corretamente a palavra-alvo.

A análise de *cluster* demonstrou que na primeira sílaba (em posição de onset), o fonema com a pior identificação foi o fonema /s/ (fricativa surda alveolar), esse mesmo fonema apareceu no grupo com melhor identificação em posição intervocálica, no entanto em relação ao agrupamento do fonema independentemente da posição, esse fonema permaneceu no grupo com pior identificação. Nesse caso, excepcionalmente, cabe observar que isso aconteceu possivelmente em virtude da alta confusabilidade ocorrida em "soma", o que ocasionou com que esse fonema permanecesse no grupo com pior identificação independentemente da posição na palavra. O mesmo ocorreu para o fonema /m/ (oclusiva nasal sonora bilabial) sendo esse fonema formador de um grupo único de pior identificação em posição intervocálica, entretanto, em posição de onset, esse fonema apareceu em um grupo de identificação moderada.

Os grupos com piores identificações em posição de onset tinham mesmo modo de articulação (fricativas surda /s/, fricativa sonora /ʒ/ e /z/, oclusivas nasais sonoras /m/ e /n/) e apareceram em grupos distintos na análise da identificação em posição intervocálica. O fonema /ʒ/ não apareceu em posição intervocálica em nenhum estímulo, por isso não foi considerado na segunda análise.

O grupo com melhor identificação em posição de onset foi formado por plosivas surdas, com diferentes pontos de articulação (bilabial /p/, linguodental /t/, velar /k/) e a fricativa surda palatal (/ʃ/), cabe ressaltar que o fonema /ʃ/ foi apresentado na palavra que teve a maior proporção de respostas corretas. O segundo grupo com melhor identificação em posição de onset foi formado por plosivas sonoras (bilabial /b/, linguodental /d/, velar /g/), fricativas sonora e surda labiodental (/v/ e /f/), lateral sonora alveolar (/l/) e vibrante múltipla sonora velar (/R/).

O grupo com pior identificação em posição intervocálica foi formado pela plosiva surda velar (/k/), vibrante simples sonora alveolar (/r/) e fricativa sonora labiodental (/v/), seguido pelo grupo formado pelas plosivas sonoras bilabial e velar (/b/ e /g/), plosiva surda linguodental (/t/) e oclusiva nasal sonora linguodental (/n/).

É interessante ressaltar a diferença de agrupamento nas análises, todas as plosivas aparecem nos grupos com melhor identificação em posição de onset. Já em posição intervocálica, somente as plosivas /p/ e /d/ aparecem no grupo com melhor identificação. Esse mesmo grupo foi formado também pelas fricativas surdas alveolar, palatal e labiodental (/s/, /ʃ/, /f/), fricativa sonora alveolar (/z/), lateral sonora alveolar e palatal (/l/ e /ʎ/), oclusiva nasal palatal (/ɲ/), plosiva sonora linguodental (/d/) e plosiva surda bilabial (/p/).

Os fonemas com pior identificação independentemente da posição foram a oclusiva nasal sonora linguodental (/n/), a fricativa surda alveolar (/s/) e a fricativa sonora alveolar (/z/), seguido pelo grupo formado pela plosiva surda velar (/k/), a oclusiva nasal sonora bilabial (/m/) e a fricativa sonora labiodental (/v/). Seguidos pelos grupos formados por aqueles fonemas que apareceram apenas em um dos contextos, da lateral sonora palatal (/ʎ/), oclusiva nasal sonora palatal (/ɲ/) e vibrante simples sonora alveolar (/r/). Os fonemas com melhor identificação geral (independentemente da posição) foram as plosivas sonoras (bilabial /b/, linguodental /d/, velar /g/), plosivas surdas (bilabial /p/, linguodental /t/), fricativas surdas (palatal /ʃ/, labiodental /f/), lateral sonora palatal (/l/) e vibrante múltipla sonora velar (/R/). As tabelas apresentam os agrupamentos dos fonemas consonantais em relação a classificação conforme o inventário fonético, em três análises distintas: em posição de onset (tabela 18), intervocálica (tabela 19) e independentemente da posição do fonema na palavra (tabela 20).

Tabela 18 - Agrupamentos dos Fonemas Consonantais em Posição de Onset Especificados pela Classificação Conforme Inventário Fonético

Fonema Primeira Sílaba	Classificação Conforme Inventário Fonético	Cluster
/ʃ/	Fricativa Surda Palatal	5
/t/	Plosiva Surda Linguodental	
/k/	Plosiva Surda Velar	
/p/	Plosiva Surda Bilabial	4
/b/	Plosiva Sonora Bilabial	
/d/	Plosiva Sonora Linguodental	
/g/	Plosiva Sonora Velar	
/R/	Vibrante Múltipla Sonora Velar	
/l/	Lateral Sonora Alveolar	
/f/	Fricativa Surda Labiodental	
/v/	Fricativa Sonora Labiodental	
/ʒ/	Fricativa Sonora Palatal	3
/m/	Oclusiva Nasal Sonora Bilabial	
/n/	Oclusiva Nasal Sonora Linguodental	2
/z/	Fricativa Sonora Alveolar	
/s/	Fricativa Surda Alveolar	1

Fonte: elaborado pela autora, 2024.

Não foi observado nenhum padrão de agrupamento pela sonoridade, nasalidade ou ponto de articulação, no entanto, o modo de articulação de plosivas no geral teve melhor identificação. Esse descrito corrobora com a discussão de De Martino (2005), a partir de trabalhos do Inglês, de que o vozeamento e a nasalidade são marcantes em estímulos auditivos, porém não permitem contraste visual. No entanto, cabe ressaltar que, como apresentado anteriormente, isso não significa que o observador não perceba a sonoridade e a nasalidade, por

exemplo, nos estímulos completos, mas sim que eles não parecem ser determinantes para a identificação dos fonemas.

Tabela 19 - Agrupamentos dos Fonemas Consonantais em Posição Intervocálica Especificados pela Classificação Conforme Inventário Fonético

Fonema Segunda Sílaba	Classificação Conforme Inventário Fonético	Cluster
/ɲ/	Oclusiva Nasal Sonora Palatal	4
/p/	Plosiva Surda Bilabial	
/d/	Plosiva Sonora Linguodental	
/ʃ/	Fricativa Surda Palatal	
/z/	Fricativa Sonora Alveolar	
/s/	Fricativa Surda Alveolar	
/f/	Fricativa Surda Labiodental	
/l/	Lateral Sonora Alveolar	
/ʎ/	Lateral Sonora Palatal	
/R/	Vibrante Múltipla Sonora Velar	
/n/	Oclusiva Nasal Sonora Linguodental	3
/b/	Plosiva Sonora Bilabial	
/t/	Plosiva Surda Linguodental	
/g/	Plosiva Sonora Velar	
/k/	Plosiva Surda Velar	2
/v/	Fricativa Sonora Labiodental	
/r/	Vibrante Simples Sonora Alveolar	
/m/	Oclusiva Nasal Sonora Bilabial	1

Fonte: elaborado pela autora, 2024.

A identificação de fonemas consonantais a depender da posição desse fonema na palavra tem relação com os efeitos de coarticulação. Em posição de onset, a identificação será diferente do que em posição intervocálica. De Martino (2005) também discute trabalhos que a visibilidade dos fonemas consonantais em posição intervocálica reduziu. Uma análise mais aprofundada dos efeitos da coarticulação se faz necessária. Discutir a redução dessa visibilidade relacionada a efeitos da coarticulação com trabalhos do Inglês não corresponde aos mesmos contextos que podem ser encontrados no Português Brasileiro, especialmente na identificação em palavras, no entanto, pode indicar caminhos relacionados às classificações dos fonemas consonantais em relação aos efeitos de coarticulação.

De Martino (2005) descreveu homofemas consonantais do Português Brasileiro. Segundo o autor, homofemas se referem aos segmentos sonoros que não são possíveis de serem identificados visualmente (nome atribuído a junção de homo e morfema). Esse conceito, na verdade, não diz respeito a identificação dos fonemas em palavras, mesmo com a proximidade na definição de visema e homofema. O aspecto trazido por De Martino (2005) sobre a

classificação dos fonemas consonantais ocasiona implicações interessantes de serem discutidas nesse trabalho, uma vez que os homofemas definidos pelo autor (tabela 21) estão diretamente relacionados aos pontos e modos de articulação estabelecidos no inventário fonético. De Martino (2005) descreve que os fonemas do Português Brasileiro, que aparecem no Inglês, /f, v/, /p, b, m/ e /ʃ, ʒ/ formam os grupos de homofemas na maioria dos estudos e que os demais fonemas /t/, /d/, /n/, /l/, /s/, /z/, /k/ e /g/ são menos visíveis.

Tabela 20 - Agrupamentos dos Fonemas Consonantais Independente da Posição Especificados pela Classificação Conforme Inventário Fonético

Fonema	Classificação Conforme Inventário Fonético	Cluster
/b/	Plosiva Sonora Bilabial	5
/d/	Plosiva Sonora Linguodental	
/g/	Plosiva Sonora Velar	
/p/	Plosiva Surda Bilabial	
/t/	Plosiva Surda Linguodental	
/l/	Lateral Sonora Alveolar	
/f/	Fricativa Surda Labiodental	
/ʃ/	Fricativa Surda Palatal	
/R/	Vibrante Múltipla Sonora Velar	
/k/	Plosiva Surda Velar	4
/m/	Oclusiva Nasal Sonora Bilabial	
/v/	Fricativa Sonora Labiodental	
/n/	Oclusiva Nasal Sonora Linguodental	3
/s/	Fricativa Surda Alveolar	
/z/	Fricativa Sonora Alveolar	
/ʎ/	Lateral Sonora Palatal	2
/ɲ/	Oclusiva Nasal Sonora Palatal	
/r/	Vibrante Simples Sonora Alveolar	
/ʒ/	Fricativa Sonora Palatal	1

Fonte: elaborado pela autora, 2024.

Segundo o autor, os fonemas /p/, /b/, /m/, /f/ e /v/ são articulados com uso dos lábios e em lugares externos na cavidade vocal e os fonemas /ʃ/, /ʒ/ envolvem protusão labial, o que facilita a percepção visual. Já os fonemas /t/, /d/, /n/, /s/, /z/, /k/, /g/ e /l/ não têm características visuais muito explícitas, o que dificulta a identificação. Ainda conforme De Martino (2005), há uma grande quantidade de fonemas classificados como alveolares (o inventário fonético utilizado pelo autor considera os fonemas linguodentais como alveolares), o que torna a discriminação desses fonemas mais difícil. Além dos velares, que possuem articulação no interior da cavidade oral e por isso também são difíceis de serem reconhecidos.

Tabela 21 - Homofemas do Português Brasileiro Definidos por De Martino (2005)

Homofema	Designação
[p, b, m]	Bilabial
[f, v]	Labiodental
[t, d, n]	Alveolar plosivo/nasal
[s, z]	Alveolar fricativo
[ɾ]	Alveolar tepe
[l]	Alveolar lateral
[ʃ, ʒ]	Pós-alveolar
[ʎ, ɲ]	Palatal
[k, g]	Velar plosivo
[ɣ]	Velar fricativo

Fonte: De Martino, 2005, p. 68.

Nesse trabalho, não foi observada nenhuma tendência quanto ao ponto de articulação, no entanto, no geral, alveolares e palatais tiveram pior identificação em palavras, além de que velares e linguodentais estiveram nos grupos com melhor identificação.

5. CONCLUSÃO

Poucas pesquisas foram desenvolvidas a respeito da leitura labial do Português Brasileiro, o que ocasiona pouca compreensão da percepção visual nesse idioma. Grande parte dos estudos da fala no Português são relacionados a escrita, a percepção auditiva, audiovisual ou a produção da fala. Após anos de investigação da leitura labial do Inglês, ainda há pouco consenso metodológico. Bernstein (2012) ressalta que "agora sabemos que o sistema visual pode ser um jogador completo no jogo de percepção de fala, mas muito ainda precisa ser aprendido sobre como o jogo é jogado" (Bernstein, 2012, p. 39, tradução própria), que enfatiza a relevância de estudo da leitura labial, mesmo com poucas evidências de como essa percepção ocorre e pode ser melhor investigada.

Traçar e elaborar estímulos e tarefas experimentais que sejam exclusivos e adequados para investigação da leitura labial parece ser uma via para inserir o Brasil no jogo e jogar apropriadamente, como Bernstein bem colocou. Materiais adaptados, que utilizam como base outras modalidades de percepção (como auditiva ou escrita), podem não fornecer base adequada para avanço científico dessa temática. No entanto, profissionais de diferentes áreas precisam trabalhar em conjunto para estudar a leitura labial de maneira a englobar diferentes cenários que são de extrema relevância para esses avanços. De forma pioneira, esse trabalho objetivou desenvolver uma base de arquivos de vídeo para treinamento de leitura da fala em Português Brasileiro, por meio de estímulos naturais, com palavras que contemplem diferentes fonemas. Novos trabalhos vão poder usar a base de arquivos produzidos nessa pesquisa (110 vídeos de estímulo exclusivamente visual produzidos em qualidade HD com 60 FPS de 55 palavras do Português Brasileiro, contemplando todos os fonemas da língua em posição de onset e intervocálica, com estrutura CVCV, de dois locutores diferentes (um homem cis e uma mulher cis).

Os estímulos desenvolvidos nessa pesquisa são passíveis de serem utilizados para desenvolver um teste computadorizado robusto de leitura labial, desde que ampliados com mais variações de estruturas e tamanhos de palavras e diferentes informações linguísticas (como frequências de uso e densidades de vizinhança), e validação psicométrica.

A leitura labial envolve não só a identificação do estímulo físico, mas também o processamento linguístico de nível superior. Por isso, para o treinamento dessa habilidade e compreensão da percepção é necessário considerar particularidades da língua relacionadas ao léxico. No protocolo experimental dessa pesquisa, os estímulos tinham características lexicais presentes que indicaram um processamento *top-down* envolvido na leitura labial, o tempo de resposta apresentou correlação negativa significativa com respostas indicadas corretamente. No

entanto, essa correlação precisa ser melhor investigada, visto que pode estar associada a outros processos cognitivos, como a memória de trabalho. Dessa forma, o material desenvolvido propicia não só o fomento de tarefas de treinamento, mas também outras investigações associadas às bases biológicas e processos cognitivos relacionados a leitura labial. Além de estudos, por exemplo, envolvendo a neuroplasticidade, uma vez que a leitura labial auxilia na oralização e aquisição do Português escrito como segunda língua para pessoas com deficiência auditiva.

A identificação dos fonemas em palavras não demonstrou nenhuma tendência específica de reconhecimento baseada na classificação estabelecida no inventário fonético, de sonoridade, modo e ponto de articulação. Porém, no geral, fonemas classificados quanto ao modo de articulação, como plosivas, apresentaram melhor identificação, independentemente da posição na palavra, enquanto a sonoridade e nasalidade não parecem ter desempenhado um papel relevante na identificação.

Apesar da tarefa experimental ter enfoque na identificação dos fonemas em palavras, novas hipóteses surgem a partir dessa pesquisa que podem fomentar alguns avanços na investigação dessa temática. Por exemplo, se o reconhecimento visual da fala se der realmente por palavras armazenadas no léxico, que competem em função da sua similaridade e baseadas na frequência de uso e quantidade de vizinhos, então palavras seriam reconhecidas mais facilmente do que pseudopalavras com fonemas semelhantes entre elas. Além de que, palavras com alta frequência de uso e com densidade de vizinhança esparsa seriam reconhecidas mais facilmente do que palavras de baixa frequência e densidade de vizinhança densa. É necessário traçar de maneira mais apropriada essas métricas para a modalidade visual.

Já quanto ao treinamento, é interessante fazer um paralelo entre a leitura labial e processos de aprendizagem de leitura e escrita, pois algumas pesquisas do Inglês já demonstraram a associação dessa habilidade à melhora no desempenho de pessoas com dislexia. Ainda, uma nova linha de investigação relacionada a aprendizagem reflexiva, baseada na *Reverse Hierarchy Theory* (RHT) (Teoria da Hierarquia Reversa – tradução própria), tem sido fomentada no Inglês e pode ser próspera também para pesquisas em Português Brasileiro.

Não é possível dissociar o estudo desse tema à investigação interdisciplinar, apesar de ser um trabalho que estava inicialmente vinculado à Percepção, é fundamental para novos avanços o envolvimento das áreas de Fonoaudiologia, Linguística (ao que parece principalmente na perspectiva da Fonologia de Uso), Linguística Computacional (para traçar métricas mais robustas e competidores adequados, baseados na realidade da fala e na modalidade visual), Psicologia, Psicolinguística (para investigar outras questões cognitivas,

como a memória) e Neurociência (para entender as bases biológicas da leitura labial).

A evolução dessas discussões e áreas se relacionam diretamente com as formas de estudo da leitura labial. A percepção de faces com informações linguísticas é complexa e não se baseia somente no estímulo externo/físico. As áreas correlatas ao estudo da percepção visual da fala por meio da leitura labial tem cada vez mais aprimorado as discussões para entender de maneira interdisciplinar o processamento da fala, seja na modalidade auditiva, audiovisual ou visual. No entanto, ainda parece haver um descompasso entre discussões teóricas e pesquisas empíricas realizadas dentro dessa temática, portanto, vislumbra-se que o potencial de desenvolvimento é vasto e pode beneficiar diferentes pessoas e áreas. Espera-se ter-se contribuído para que este seja apenas o início dessa investigação em Português Brasileiro.

REFERÊNCIAS

AUER, E. T. The influence of the lexicon on speech read word recognition: Contrasting segmental and lexical distinctiveness. **Psychonomic Bulletin and Review**, v. 9, n. 2, 2002.

AUER, E. T.; BERNSTEIN, L. E. Speechreading and the structure of the lexicon: Computationally modeling the effects of reduced phonetic distinctiveness on lexical uniqueness. **The Journal of the Acoustical Society of America**, v. 102, n. 6, 1997.

BERNSTEIN, L. E.; AUER, E. T.; EBERHARDT, S. P. During Lipreading Training with Sentence Stimuli, Feedback Controls Learning and Generalization to Audiovisual Speech in Noise. **American Journal of Audiology**, v. 31, n. 1, 2022.

BERNSTEIN, L. E.; JORDAN, N.; AUER, E. T.; EBERHARDT, S. P. Lipreading: A Review of Its Continuing Importance for Speech Recognition with an Acquired Hearing Loss and Possibilities for Effective Training. **American Journal of Audiology**, 2022.

BERNSTEIN, L. E.; LIEBENTHAL, E. Neural pathways for visual speech perception. **Frontiers in Neuroscience**, 2014.

BERNSTEIN, L. E. Visual speech perception. In: **Audiovisual Speech Processing**. Cambridge: Cambridge University Press, p. 21-39, 2012.

BERNSTEIN, L. E.; IVERSON, P.; AUER Jr., E. T. Elucidating the complex relationships between phonetic perception and word recognition in audiovisual speech perception. Workshop on Auditory-Visual Speech Processing. Rhodes, Greece, p. 89-92, 1997.

BUCHANAN-WORSTER, E.; HULME, C.; DENNAN, R.; MACSWEENEY, M. Speechreading in hearing children can be improved by training. **Developmental Science**, v. 24, n. 6, 2021.

BYBEE, J. Phonology and Language Use. Cambridge Studies in Linguistics 94. Cambridge: CUP, 2001.

CAMPBELL, R; MOHAMMED, T. J. Speechreading for information gathering: a survey of scientific sources. (DCAL Associated Projects). Deafness Cognition and Language (DCAL) Research Centre, Division of Psychology and Language Sciences, University College London: London, UK, 2010.

COSTA, P. D. P. Animação facial 2D sincronizada com a fala baseada em imagens de visemas dependentes do contexto fonético. Dissertação (mestrado). Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação, Campinas, SP, 2009. Disponível em: https://hdl.handle.net/20.500.12733/1610148. Acesso em: 20 out. 2023.

CRESTI, E. Notes on lexical strategy, structural strategies and surface clause indexes in the C-ORAL-ROM spoken *corpora*. In: CRESTI, E.; MONEGLIA, M. (Ed.). C-ORAL-ROM: Integrated reference *corpora* for spoken Romance Languages. Amsterdam/Philadelphia: John Benjamins, p. 209-256, 2005.

CRISTÓFARO-SILVA, T. Fonética e Fonologia do Português: Roteiro de Estudos e Guia de Exercícios 10 ed., 6ª reimpressão. São Paulo: Contexto, 2015.

CRISTÓFARO-SILVA, T.; DE BONA, C. O papel do léxico na variação fonológica: uma entrevista com Thais Cristófaro-Silva. ReVEL, edição especial n. 14, 2017.

DE MARTINO, J. M. Animação facial sincronizada com a fala: visemas dependentes do contexto fonético para o português do Brasil. Tese (doutorado). Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação, Campinas, SP, 2005. Disponível em: https://hdl.handle.net/20.500.12733/1600896. Acesso em: 20 out. 2023.

FELD, J. E.; SOMMERS, M. S. Lipreading, processing speed, and working memory in younger and older adults. **Journal of Speech, Language, and Hearing Research**, v. 52, n. 6, 2009.

FILES, B. T.; TJAN, B. S.; JIANG, J.; BERNSTEIN, L. E. Visual speech discrimination and identification of natural and synthetic consonant stimuli. **Frontiers in Psychology**, v. 6, 2015.

FISHER, C. G. Confusions among visually perceived consonants. **J Speech Hear Res**., p. 796-804, 1968.

HEGAZI, M. A. F.; SAAD, A. M.; KHODEIR, M. S. Development of a test for assessment of the lipreading ability for children in the Arabic-speaking countries. **Egyptian Journal of Otolaryngology**, v. 37, n. 1, 2021.

ISSLER, S. Articulação e Linguagem. 1. ed. Editora Revinter, 1996.

KATHLEEN PICHORA-FULLER, M.; SCHNEIDER, B. A.; DANEMAN, M. How young and old adults listen to and remember speech in noise. **Journal of the Acoustical Society of America**, v. 97, n. 1, 1995.

KYLE, F. E.; CAMPBELL, R.; MOHAMMED, T.; COLEMAN, M.; MACSWEENEY, M. Speechreading development in deaf and hearing children: Introducing the test of child speechreading. **Journal of Speech, Language, and Hearing Research**, v. 56, n. 2, 2013.

LUCE, P. A.; PISONI, D. B. Recognizing spoken words: The neighborhood activation model. **Ear and Hearing**, v. 19, n. 1, 1998.

MACDONALD, J.; MCGURK, H. Visual influences on speech perception processes. **Perception and Psychophysics**, v. 24, n. 3, 1978.

MASSARO, D. W. Speech Perception. In: **International Encyclopedia of the Social and Behavioral Sciences**, 2. ed, v.23, p. 235-242, 2015.

MATTYS, S. L.; BERNSTEIN, L. E.; AUER, E. T. Stimulus-based lexical distinctiveness as a general word-recognition mechanism. **Perception and Psychophysics**, v. 64, n. 4, 2002.

MELLO, H. Os *corpora* orais e o C-ORAL-BRASIL. In: RASO, T.; MELLO, H. C-ORAL-BRASIL: *corpus* de referência do português brasileiro falado informal. Editora UFMG, p. 31-54, 2012.

MCGURK, H.; MACDONALD, J. Hearing lips and seeing voices. **Nature**, v. 264, n. 5588, 1976.

MOHAMMED, T.; CAMPBELL, R.; MACSWEENEY, M.; BARRY, F.; COLEMAN, M. Speechreading and its association with reading among deaf, hearing and dyslexic individuals. **Clinical Linguistics and Phonetics**. Anais. v. 20, 2006.

NEWCOMBE R.G. Interval Estimation for the Difference Between Independent Proportions: Comparison of Eleven Methods. **Statistics in Medicine**, 17, 873–890, 1998.

OLIVEIRA, L.; SOARES, A. D.; CHIARI, B. M. Leitura da fala como mediadora da comunicação. **CoDAS**, v. 26, n. 1, p. 53-60, 2014.

PIMPERTON, H.; KYLE, F.; HULME, C.; HARRIS, M.; BEEDIE, I.; RALPH-LEWIS, A.; WORSTER, E.; REES, R.; DONLAN, C.; MACSWEENEY, M. Computerized speechreading training for deaf children: A randomized controlled trial. **Journal of Speech, Language, and Hearing Research**, v. 62, n. 8, 2019.

PIMPERTON, H.; RALPH-LEWIS, A.; MACSWEENEY, M. Speechreading in deaf adults with cochlear implants: Evidence for perceptual compensation. **Frontiers in Psychology**, v. 8, 2017.

PIQUARD-KIPFFER, A.; CAVADINI, T.; SPRENGER-CHAROLLES, L.; GENTAZ, É. Impact of lip-reading on speech perception in French-speaking children at risk for reading failure assessed from age 5 to 7. **Annee Psychologique**, v. 121, n. 2, 2021.

SEARA, I. C., NUNES, V. G., VOLCÃO, C. L. Fonética e fonologia do português brasileiro: 2º período. Florianópolis: LLV/CCE/UFSC, 2011.

SILVA, F. M. da. Processos Fonológicos Segmentais na Língua Portuguesa. **Littera: Revista de Estudos Linguísticos e Literários**, v. 2, n. 4, 7 Mar., 2011.

STRAND, J. F. Phi-square Lexical Competition Database (Phi-Lex): An online tool for quantifying auditory and visual lexical competition. **Behavior Research Methods**, v. 46, n. 1, 2014.

TEDESCO, M. R. M.; CHIARI, B. M.; VIEIRA, R. M. Influências do método oral e da comunicação total no desenvolvimento da habilidade de leitura da fala de deficientes auditivos. **Revista Brasileira de Medicina - Otorrinolaringologia**. 2(5): 348, 350-1, 354, 1995.

TOFFOLO, A. C. R.; BERNARDINO, E. L. A.; VILHENA, D. A.; PINHEIRO, Â. M. V. Os benefícios da oralização e da leitura labial no desempenho de leitura de surdos profundos usuários da Libras. **Revista Brasileira de Educação**, v. 22, n. 71, 2017.

WILSON, E.B. Probable inference, the law of succession, and statistical inference. **Journal of the American Statistical Association**, 22, 209–212, 1927.

ZHANG, F.; LEI, J.; GONG, H.; WU, H.; CHEN, L. The development of speechreading skills in Chinese students with hearing impairment. **Frontiers in Psychology**, v. 13, 2022.

ANEXO

Anexo 1 – Aprovação do Comitê de Ética em Pesquisa

Universidade de São Paulo
Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto
Comitê de Ética em Pesquisa

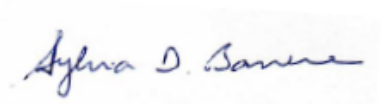
OF.143/CEP/FFCLRP/USP/22.11.2021

Prezada Pesquisadora,

Comunicamos a V. Sa. que o projeto de pesquisa intitulado "**Leitura da Fala do Português-Brasileiro: Elaboração de Vídeos para Treinamento**" foi analisado pelo Comitê de Ética em Pesquisa da FFCLRP-USP em sua 221ª Reunião Ordinária, realizada em 18.11.2021, e enquadrado na categoria: **APROVADO** (CAAE nº 52025821.5.0000.5407).

Solicitamos que eventuais modificações ou emendas ao projeto de pesquisa sejam apresentadas ao CEP, de forma sucinta, identificando a parte do projeto a ser modificada e suas justificativas. De acordo com a Resolução nº 466 de 12.12.2012, devem ser entregues relatórios semestrais e, ao término do estudo, um relatório final, sempre via Plataforma Brasil.

Atenciosamente,



Profa. Dra. Sylvia Domingos Barrera
Coordenadora

Ilma. Sra.
Fernanda De Barros Vidal
Programa de Pós-Graduação em Psicobiologia da FFCLRP-USP

CEP - Comitê de Ética em Pesquisa da FFCLRP USP
Av. Bandeirantes, 3900 - Bloco 01 da Administração - Sala 07.
14040-901 - Ribeirão Preto - SP – Brasil. Fone: (16) 3315-4811.
Homepage: http://www.ffclrp.usp.br. E-mail: coetp@listas.ffclrp.usp.br.